

**SOLUTION OF
EQUATIONS IN
EUCLIDEAN AND
BANACH SPACES**

Pure and Applied Mathematics

A Series of Monographs and Textbooks

Editors **Paul A. Smith and Samuel Eilenberg**

Columbia University, New York

RECENT TITLES

ERNST AUGUST BEHRENS. Ring Theory

MORRIS NEWMAN. Integral Matrices

GLEN E. BREDON. Introduction to Compact Transformation Groups

WERNER GREUB, STEPHEN HALPERIN, AND RAY VANSTONE. Connections, Curvature, and Cohomology : Volume I, De Rham Cohomology of Manifolds and Vector Bundles. Volume II, Lie Groups, Principal Bundles, and Characteristic Classes

XIA DAO-XING. Measure and Integration Theory of Infinite-Dimensional Spaces : Abstract Harmonic Analysis

RONALD G. DOUGLAS. Banach Algebra Techniques in Operator Theory

WILLARD MILLER, JR. Symmetry Groups and Their Applications

ARTHUR A. SAGLE AND RALPH E. WALDE. Introduction to Lie Groups and Lie Algebras

T. BENNY RUSHING. Topological Embeddings

JAMES W. VICK. Homology Theory : An Introduction to Algebraic Topology

E. R. KOLCHIN. Differential Algebra and Algebraic Groups

GERALD J. JANUSZ. Algebraic Number Fields

A. S. B. HOLLAND. Introduction to the Theory of Entire Functions

WAYNE ROBERTS AND DALE VARBERG. Convex Functions

A. M. OSTROWSKI. Solution of Equations in Euclidean and Banach Spaces, Third Edition of Solution of Equations and Systems of Equations

In preparation

SAMUEL EILENBERG. Automata, Languages, and Machines : Volume A

H. M. EDWARDS. Riemann's Zeta Function

WILHELM MAGNUS. Non-Euclidean Tesselations and Their Groups

J. DIEUDONNÉ. Treatise on Analysis, Volume IV

MORRIS HIRSCH AND STEPHEN SMALE. Differential Equations, Dynamical Systems, and Linear Algebra

SOLUTION OF EQUATIONS IN EUCLIDEAN AND BANACH SPACES

Third Edition of
**Solution of Equations
and Systems of Equations**

A. M. OSTROWSKI

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BASEL
BASEL, SWITZERLAND

1973



ACADEMIC PRESS New York and London
A Subsidiary of Harcourt Brace Jovanovich, Publishers

**COPYRIGHT © 1973, BY ACADEMIC PRESS, INC.
ALL RIGHTS RESERVED.**

**NO PART OF THIS PUBLICATION MAY BE REPRODUCED OR
TRANSMITTED IN ANY FORM OR BY ANY MEANS, ELECTRONIC
OR MECHANICAL, INCLUDING PHOTOCOPY, RECORDING, OR ANY
INFORMATION STORAGE AND RETRIEVAL SYSTEM, WITHOUT
PERMISSION IN WRITING FROM THE PUBLISHER.**

**ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003**

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 73-8310

**AMS (MOS) 1970 Classifications: 65H05;
65H10; 65J05; 65B99; 12D10; 15A60**

PRINTED IN THE UNITED STATES OF AMERICA

Contents

PREFACE TO THE THIRD EDITION	xiii
From the Preface to the Second Edition	xiv
From the Preface to the First Edition	xiv
LIST OF NOTATIONS AND ABBREVIATIONS	xvii

1A. DIVIDED DIFFERENCES

Divided Differences for Distinct Arguments	1
Symmetry	2
Hermite's Integral Representation	3
Mean Value Formulas	5

1B. CONFLUENT CASE. INTERPOLATION

Confluent Divided Differences	8
Continuity of Confluent Divided Differences	9
Various Formulas for Divided Differences	10
Newton's Interpolation Formula	12
General Interpolation Problem	14
Polynomial Interpolation	15
The Remainder for a General Interpolating Function	15
Triangular Schemes for Computing Divided Differences	16

2. INVERSE INTERPOLATION. DERIVATIVES OF THE INVERSE FUNCTION. ONE INTERPOLATION POINT

The Concept of Inverse Interpolation	18
Darboux's Theorem on Values of $f'(x)$	19
Derivatives of the Inverse Function	20
One Interpolation Point	22
A Development of a Zero of $f(x)$	25

3. METHOD OF FALSE POSITION (REGULA FALSI)

Definition of the Regula Falsi	27
Use of Inverse Interpolation	28
Geometric Interpretation (Fourier's Conditions)	30
Iteration with Successive Adjacent Points	31
Horner Units and Efficiency Index	32
The Rounding-Off Rule	33
Locating the Zero with the Regula Falsi	34
Examples of Computation by the Regula Falsi	35

4. ITERATION

A Convergence Criterion for an Iteration	38
Points of Attraction and Repulsion	38
Improving the Convergence	40

5. FURTHER DISCUSSION OF ITERATIONS. MULTIPLE ZEROS

Iterations by Monotonic Iterating Functions	47
Multiple Zeros	48
Connection of the Regula Falsi with the Theory of Iteration	51

6. THE NEWTON-RAPHSON METHOD

The Idea of the Newton-Raphson Method	53
The Use of Inverse Interpolation	53
Comparison of Regula Falsi and Newton-Raphson Method	55

7. FUNDAMENTAL EXISTENCE THEOREMS IN THE NEWTON-RAPHSON ITERATION

Error Estimates A Priori and A Posteriori	56
Fundamental Existence Theorems	56

8. AN ANALOG OF THE NEWTON-RAPHSON METHOD FOR MULTIPLE ROOTS

Convergence of Schröder's Iteration for Multiple Roots	61
Error Estimates A Priori	63
Recurrent Error Estimates	65
Evaluation of Exact Multiplicity	67

9. FOURIER BOUNDS FOR THE NEWTON-RAPHSON ITERATION	69
10. DANDELIN BOUNDS FOR THE NEWTON-RAPHSON ITERATION	73
11. THREE INTERPOLATION POINTS	
Interpolation by Linear Fractions	79
Two Coincident Interpolation Points.	80
Error Estimates	81
Use in Iteration Procedure	82
12. LINEAR DIFFERENCE EQUATIONS	
Inhomogeneous and Homogeneous Difference Equations	84
General Solution of the Homogeneous Equation.	85
Lemma on Division of Power Series	86
Asymptotic Behavior of Solutions of (12.1)	87
Asymptotic Behavior of Errors in the Regula Falsi Iteration.	90
A Theorem on Roots of Certain Equations	91
13. n DISTINCT POINTS OF INTERPOLATION	
Error Estimates	94
Iteration with n Distinct Interpolation Points	95
Discussion of the Roots of Some Special Equations	97
14. $n+1$ COINCIDENT INTERPOLATION POINTS AND TAYLOR DEVELOPMENT OF THE ROOT	
Statement of the Problem	103
A Theorem on Inverse Functions and Conformal Mapping	103
Theorem on the Error of the Taylor Approximation to the Root	106
Discussion of the Conditions of the Theorem	107
15. THE SQUARE ROOT ITERATION	
Polynomials with Simple Real Zeros Only	110
Modification for Multiple Zeros	113
Differentiable Functions and Complex Zeros	115

16. FURTHER DISCUSSION OF SQUARE ROOT ITERATION	
Local Formulation of the Existence and Convergence Theorem	119
Extension to Entire Functions	124
17. A GENERAL THEOREM ON ZEROS OF INTERPOLATING POLYNOMIALS	127
18. APPROXIMATION OF EQUATIONS BY ALGEBRAIC EQUATIONS OF A GIVEN DEGREE. ASYMPTOTIC ERRORS FOR SIMPLE ROOTS	
Convergence of Zeros of Interpolating Polynomials	131
Asymptotic Errors for Simple Zeros	132
19. NORMS OF VECTORS AND MATRICES	
Vector Norms	135
Matrix Norms $ A _1$ and $ A _\infty$	137
Eigenvalues of A	140
20. TWO THEOREMS ON CONVERGENCE OF PRODUCTS OF MATRICES	143
21. A THEOREM ON DIVERGENCE OF PRODUCTS OF MATRICES	146
22. CHARACTERIZATION OF POINTS OF ATTRACTION AND REPULSION FOR ITERATIONS WITH SEVERAL VARIABLES	
Points of Attraction and Repulsion	150
An Example	153
23. EUCLIDEAN NORMS	
Euclidean Length and Frobenius Norm	155
Hermitian Matrices	156
Euclidean Norm of a Matrix	157

24. MINKOWSKI NORMS, $\Delta_p(A)$, $\Delta_{p,p'}(A)$	
Minkowski Norms	159
$ A _p$ and $ A _{p,p'}$	159
$\Delta_{p,p'}(A)$ and $\Delta_p(A)$	161
Inequalities for $\Delta_{p,p'}(A)$	163
Variation of the Inverse Matrix	164
25. METHOD OF STEEPEST DESCENT. CONVERGENCE OF THE PROCEDURE	
Idea of the Method	166
Convergence of the Procedure	168
Application to $ f(x+iy) ^2$	170
26. METHOD OF STEEPEST DESCENT. WEAKLY LINEAR CONVERGENCE OF THE ζ_μ	
The Derived Set of the ζ_μ	173
Weakly Linear Convergence	174
Condition for the Regular Minimum of the Function (25.3)	176
Algebraic Equations with One Unknown	177
27. METHOD OF STEEPEST DESCENT. LINEAR CONVERGENCE OF THE ζ_μ	
Conditions for Strictly Linear Convergence	178
An Example	180
Connection with the Newton-Raphson Procedure	183
28. CONVERGENT PROCEDURES FOR POLYNOMIAL EQUATIONS	
The First Step of the Procedure	186
Convergence of the Iteration Procedure	188
Switching over to the Newton-Raphson Procedure	190
The Ω -Test	190
29. J-TEST AND J-ROUTINE	
Basic Theorem	194
The J -Test	196
The J_m -Routine	197

30. <i>q</i>-ACCELERATION. THE PRACTICE OF THE PROCEDURE	
The Definition of <i>q</i> -Acceleration	200
The Basic Lemma	201
The Convergence Discussion	203
Speed of Convergence	203
Flow Charts	205
31. NORMED LINEAR SPACES	
Linear Spaces	207
Norms	208
Convergence	209
Completeness and Compactness	210
Examples	210
Spaces $C^k(J)$	211
Spaces $L_a(G)$	212
32. METRIC SPACES	
Definition of Metric Spaces	214
Principle of Contracting Operators	215
33. OPERATORS IN NORMED LINEAR SPACES	
Mappings and Operators	219
Bounded Operators	219
Linear Operators	220
Strong and Weak Convergence	221
34. INVERSE OPERATORS	
Definition of the Inverse Operator	224
Existence of the Inverse Operator	225
Another Existence Theorem	226
A Banach Theorem	227
35. OPERATORS MAPPING A LINEAR INTERVAL	
A Refinement of Borel's Covering Theorem	229
Lipschitz Condition for $H(t)$	231
Taylor Development	233

36. THE DIRECTIONAL DERIVATIVES AND GRADIENTS OF OPERATORS	
Directional Derivatives	236
Gateau Gradient	237
F-Differentials and F-Gradients	239
37. CENTRAL EXISTENCE THEOREM	
Formulation of the Central Existence Theorem	241
A Local Existence Theorem	242
Proof of Theorem 37.1	245
38. NEWTON-RAPHSON ITERATION IN BANACH SPACES. STATEMENT OF THE THEOREMS	
Definition of the α_v	247
Formulation of Theorems 38.1–38.3	248
A Lemma	250
39. PROOF OF THEOREMS 38.1–38.3	
A Further Lemma	253
Specialization for Quadratic Polynomials	254
Proofs of Theorems 38.1–38.3	256
40. COMPLEMENTS TO THE NEWTON-RAPHSON METHOD	
Equalities in Estimates for Quadratic Polynomials	260
Multiple Solutions	260
Unicity Theorem	261
41. CENTRAL EXISTENCE THEOREM FOR FINITE SYSTEMS OF EQUATIONS	
Formulation of the Central Existence Theorem	265
The Choice of Norms	266
A Uniqueness Theorem	266
Example	267

**42. NEWTON-RAPHSON ITERATION FOR FINITE SYSTEMS
OF EQUATIONS**

Formulation of the Theorem	270
The Choice of the Norms	271
Application to Complex Functions of a Complex Variable	274

APPENDICES

A. Continuity of the Roots of Algebraic Equations	276
B. Relative Continuity of the Roots of Algebraic Equations	281
C. An Explicit Formula for the n th Derivative of the Inverse Function	290
D. Analog of the Regula Falsi for Two Equations with Two Unknowns	294
E. Steffensen's Improved Iteration Rule	296
F. The Newton-Raphson Algorithm for Quadratic Polynomials	301
G. Some Modifications and Improvements of the Newton-Raphson Method	306
H. Rounding Off in Inverse Interpolation	310
I. Accelerating Iterations with Supralinear Convergence	320
J. Roots of $f(z) = 0$ in Terms of the Coefficients of the Development of $1/f(z)$	324
K. Continuity of the Fundamental Roots as Functions of the Elements of the Matrix	334
L. The Determinantal Formulas for Divided Differences	336
M. Remainder Terms in Interpolation Formulas	339
N. Generalization of Schröder's Series to the Case of Multiple Roots	343
O. Laguerre Iterations	353
P. Approximation of Equations by Algebraic Equations of a Given Degree. Asymptotic Errors for Multiple Zeros	363
Q. Feedback Techniques for Error Estimates	372
A Numerical Example	377
R. Reduced Polynomial Equations	378
S. Discussion of the q -Acceleration	383
T. Remainder in the Taylor Formula for Analytic Functions	389
U. Equality Conditions for the Newton-Raphson Iteration	391
A Lemma	391
Equality Conditions for Normed Spaces	392
Equality Conditions for Strictly Normed Spaces	394
Bibliographical Notes	399
INDEX	409

The most practical thing in the world is a good theory.

Attributed to H. von Helmholtz

Preface to the Third Edition

The treatment of finite systems of equations, as reflected in the 1952 course from which the first edition of this book developed, has been completely superseded by the general theory of equations in Banach spaces. Thus it was felt that the old title was no longer adequate. My intent with this new edition has been to bring the reader up to date with the developments in this theory, correct what I felt were some deficiencies in both previous editions, and to retain my original hope, as expressed in the preface to the first edition, that the book would "help to a certain degree to bridge the gap that still exists between 'pure' and 'practical' mathematics."

For this third edition, some of the chapters of the second edition have been revised and amplified, others dropped and the contents covered in new contexts. The chapter on divided differences, added in a not completely satisfactory manner to the second edition, has been rewritten and enlarged, and divided into two chapters, 1A and 1B.

Chapter 8, concerning Schröder's modification of the Newton-Raphson procedure for multiple zeros, has been rewritten, partly in order to prepare for utilization of the feedback techniques of Appendix Q and partly to account for the case in which the multiplicity index is unknown. Further, the contents of Chapters 23 and 24 on the norms of vectors and matrices have been considerably developed to prepare for the discussions in Chapters 41 and 42 of finite systems of equations. The new Chapters 28, 29, and 30 discuss an automatic and always convergent procedure for solution of polynomial equations. Although we do not give an explicit program for this procedure, the flow charts at the end of Chapter 30 ought to make the preparation of programs sufficiently easy.

The last part of the book, Chapters 31-42, discusses the solution of equations in a Banach space. In order to prepare for this discussion a self-contained exposition of the theory of normed linear and metric spaces is first given in Chapters 31-36. Although there are excellent treatises on this subject, it was felt that such an exposition would be useful, since a unified terminology in this domain is apparently not as yet fully established. In Chapter 37, we

introduce a central existence theorem which supersedes the existence theorem of the second edition's Chapter 24, covering finite systems of equations in Euclidean space.

Finally, we give in Chapters 38–40 the complete theory of the Newton–Raphson iteration in the case of Banach spaces with best possible constants and discuss, in Chapters 41 and 42, application of the general theory to finite systems of equations.

Different points of a more special character, which would make the discussion in the main text a little cumbersome, are dealt with in the new Appendices Q–U. Appendices A–P of the second edition have undergone only small changes save Appendix A, where a partly incorrect statement of the text in the second edition had to be corrected.

In preparation of this edition I was considerably helped by Dr. Rita Jeltsch (Mrs. François Fricker). In the preparation of the indexes I was assisted by Mrs. Doris Wiedemann. I feel very grateful for this help.

FROM THE PREFACE TO THE SECOND EDITION

I should like to extend my appreciation to the reviewers of the first edition for their constructive criticism. In particular I am indebted to Professor D. Greenspan for a considerable number of corrections. Further, I owe thanks to L. S. and B. L. Rumshiski, who prepared the Russian translation of my book with unusual and penetrating care, resulting in several corrections and improvements in details. I also thank my assistants in Basel, K. Goetschi, P. Gschwind, H. Rigganbach, and R. Rüedi, for their help in the preparation of the manuscript.

Finally I have to thank Dr. Pierre Banderet and Professor D. Greenspan for their valuable help in correcting the proofs.

FROM THE PREFACE TO THE FIRST EDITION

I wish to take this opportunity to extend my thanks to Dr. John H. Curtiss, then chief of the Mathematics Division of the NBS, on whose invitation these lectures were originally given and who was certainly mainly responsible for the auspicious atmosphere which proved so stimulating to the research associates in the Mathematics Division. Conversations with Dr. Olga Taussky-Todd and Dr. John Todd were always particularly inspiring. I also thank Mr. M. Ticson, who made the first draft of the notes of these lectures, and Mr. W. F. Cahill, who was my assistant at that time. During the preparation of the lectures I had much help from the computing staff of the NBS, both in Washington and

Los Angeles, help that was extremely valuable in trying out different methods in actual computing practice. Finally, I have to thank my assistants in Basel, B. Marzetta, S. Christeller, T. Witzemann, and R. Bürki, and also Drs. E. V. Haynsworth and Wa. Gautschi of the NBS for the help they gave me in the preparation of the manuscript, and Dr. Pierre Banderet and Mr. Howard Bell for their valuable help in correcting the proofs.

This Page Intentionally Left Blank

List of Notations and Abbreviations

If the definition of the notation is given in the text we give here only the corresponding page number.

\coloneqq	read: is the notation for
$=:$	is the definition of, will be denoted by
\lessgtr	signifies \neq between real numbers
\subset	is contained in
\in	is a member of
\supset	contains
\ni	has as a member
\wedge	as well as, \vee or. These two logical signs have the absolute precedence on operation, equality and inequality signs. For instance $0 < x < 1$, $0 < y < 1$, $0 < z < 1$ can be written as $0 < x \wedge y \wedge z < 1$; again “either $0 < x < 1$ or $0 < y < 1$ ” can be written as $0 < x \vee y < 1$.
θ	signifies a convenient real number from the interval $\langle 0, 1 \rangle$.
θ^*	signifies an element with the norm ≤ 1 . These symbols need not have the same value from one formula to another, while if they occur in the same formula they are equal; for instance

$$\frac{1+\theta}{1-\theta}, \quad \frac{1+\theta^*}{1-\theta^*}.$$

$J_x, (J_x)$	1
$[x]f$	1
$[x_1, \dots, x_n]f$	2, 8
$\langle x_1, \dots, x_n \rangle$	4
$J_{m_1}(t)$	6
D_t	We denote the differentiation operator with respect to t with the symbol D_t . Then the meaning

of a product of powers of such operators is obvious:

$$D_{x_1}^{\alpha_1} \cdots D_{x_n}^{\alpha_n}.$$

X_t	6
$U(x)$	6
$[x^k]f$	8
$[x^{m_1+1}, \dots, z^{m_k+1}]f$	8
\otimes^μ	12
$A \cap B$	denotes the “intersection” of the sets A and B , that is the set of all elements common to A and B .
$A \cup B$	is the set theoretical sum of A and B , that is the set of all elements that occur either in A or in B .
$\bigcup_v A_v (v \in N)$	is the set theoretical sum of all sets A_v , where v runs through all different elements of N .
\square	18
$\uparrow\downarrow$	24
O, o	26
(a_1, \dots, a_k)	is, as a set, the set of all inner points of $\langle a_1, \dots, a_k \rangle$ relatively to the “smallest” linear manifold containing all a_κ .
$\operatorname{sgn} \alpha$	31
$\operatorname{Sup} A$	for a real number set A , is the same as $\operatorname{lub} A$, the least upper bound of A .
$\operatorname{Inf} A$	is the same as $\operatorname{glb} A$, the greatest lower bound of A .
$0.50_n 3$	36
\leftrightarrow	signifies simply: does not converge to the limit indicated; a different limit may exist or not at all.
$21.2(\pm 6)$	42
$[A]$	the “Gaussian parenthesis” of A is the unique integer n such that $n \leq A < n+1$.
$f \sim g$	for some limiting process, signifies that $f/g \rightarrow 1$ for this limiting process.
\ll	87
(x_1, \dots, x_n)	135. The symbol (a_1, \dots, a_k) introduced above is apparently the same but appears in a different context, as a <i>set of points</i> .
$ \xi _p, \xi _\infty$	135
$p \wedge q$	as conjugate numbers, p. 135.
(ξ, η)	136
$ A _\infty$	137
A'	137

$ A _1$	138
$\det A$	denotes the determinant of A .
λ_A	140
$\text{tr}(A)$	142
(A)	denotes, if A is a determinant, the corresponding matrix.
$J(\xi)$	151
$ \xi _2$	155
$ A _F$	155
$H_A(\xi)$	156
\bar{A}, A^*	156
$\lambda(A), \Lambda(A)$	157
$ A _p$	159
$ A _{p,p'}$	160
$\Delta_p(\xi), \Delta_{p,p'}(\xi)$	161
$\text{grad } f(\xi)$	167, 237
$f''(\xi), f''(\xi_1, \dots, \xi_n)$	168
$ \xi, \Omega^* $	168
$\Re f$	is the real part of f
$\Im f$	is the imaginary part of f
$\langle \xi_1, \xi_2 \rangle, (\xi_1, \xi_2)$	208
$\ \xi\ $	208
NLS	208
$C^k(J)$	210
$\ f\ _0^*$	211
$L_a(G)$	212
$\ f\ _\alpha, \ f\ _\infty$	213
$ \xi, \eta $	214
$(U_r(\xi_0))$	215
$f(U \rightarrow V), f\xi$	219
$\ f\ $	219
$S_{U,Y}$	220
$f(U \leftrightarrow V)$	224
f^{-1}	224
$H'(t)$	233
$H^{(v)}(t)$	233
$df(\xi; \Delta)$	236
$d^v f(\xi; \Delta)$	237
$\text{Sin } \varphi, \text{Cos } \varphi, \text{Tg } \varphi, \text{Ctg } \varphi$	247
P_0, α_v, S_0	247
σ_0, ρ_0	248
K_0	248

$P(\xi)$	249
$P_v, Q_v, f_v, h_v, \sigma_v$	249
$\hat{\Delta}_{p,p'}(\xi), \Delta_{p,p'}^*(\xi)$	271
$D_{1,v}$	325
$D_{2,v}$	326
Δ_y	327
$\varphi_n(x), \rho_n^*$	378
$\chi_n(x)$	380
$A_v, \dots, E_v, \bar{A}_v, \dots, \bar{E}_v$	392

1A

Divided Differences

DIVIDED DIFFERENCES FOR DISTINCT ARGUMENTS

1. In this book J_x will denote an interval on the x -axis. J_x can be open, closed, or open at one end and closed at the other. (J) will denote the *interior* of the interval J , i.e., the interval J excluding the end points. (J) is an open interval.

The functions and the variables used need not be real. If the variables are complex and the derivatives of functions are used, these functions are assumed to be *analytic* on the corresponding sets of points. In the real case the derivatives are supposed to be continuous as far as they occur in the discussion, unless the opposite is explicitly stated (cf. Appendix L).

2. For any function $f(x)$ defined for $x = x_1$ and $x = x_2$, we define the symbol $[x_1, x_2]_x f$ for $x_1 \neq x_2$ by

$$[x_1, x_2]_x f := \frac{f(x_2) - f(x_1)}{x_2 - x_1} \quad (x_1 \neq x_2). \quad (1A.1)$$

For $x_1 = x_2$ this symbol is, of course, defined by

$$[x_1, x_1]_x f := f'(x)_{x=x_1}, \quad (1A.2)$$

provided $f'(x_1)$ exists.

The variable x in the subscript will be denoted as the *target available*. The subscript with the target variable can be dropped if $f(x)$ is defined as a function of *one* variable x only. The letter denoting the target variable can be replaced by any other letter, provided this letter does not already occur in the function f as a different variable:

$$[x_1, x_2] f(x) = [x_1, x_2]_u f(u). \quad (1A.3)$$

If $f(x)$ does not depend on x_1, x_2 as parameters, expression (1A.1) is a *symmetric function* of x_1 and x_2 .

3. If x, x_1, x_2 are all distinct, the two operators $[x_1, x]$ and $[x_2, x]$ *commute*, that is to say, we have

$$[x_1, x][x_2, x] f = [x_2, x][x_1, x] f, \quad (1A.4)$$

provided $f(x)$ does not depend on x_1 and x_2 . Indeed, the common value of both sides of (1A.4) is seen at once to be

$$\frac{(x-x_1)f(x_2)+(x_1-x_2)f(x)+(x_2-x)f(x_1)}{(x-x_1)(x_1-x_2)(x-x_2)},$$

and this is obviously a symmetric function of x, x_1, x_2 .

Put, if x, x_1, \dots, x_m are all distinct,

$$[x_1, x][x_2, x] \cdots [x_m, x], f =: [x, x_1, \dots, x_m], f. \quad (1A.5)$$

This is called the *divided difference (of the order m) of the function f(t)*.

When we form (1A.5), the term containing $f(x)$ is at the application of $[x_v, x]$ divided by $(x-x_v)$. It follows that the term containing $f(x)$ in (1A.5) can be written as

$$f(x) \left/ \prod_{\mu=1}^m (x-x_\mu) \right..$$

In what follows the symbol $[x]_u f(u)$ signifies $f(x)$.

SYMMETRY

4. Expression (1A.5) is, by (1A.4), a symmetric function in the m arguments x_1, \dots, x_m , if $f(t)$ does not contain any of the x, x_1, \dots, x_m as parameters. We are going to prove that (1A.5) is *symmetric in all $m+1$ arguments*. For this it is sufficient to prove that (1A.5) does not change if x is interchanged with x_1 . But, if we put

$$[x_2, x] \cdots [x_m, x], f = g(x, x_2, \dots, x_m),$$

expression (1A.5) becomes

$$[x_1, x], g(t, x_2, \dots, x_m)$$

and this, by Section 2, is obviously symmetric in x, x_1 .

It follows now from the symmetry, by what has been said about (1A.5), that we have the general formula

$$[x_1, \dots, x_m], f = \sum_{\mu=1}^m f(x_\mu) \left/ \prod_{v \neq \mu} (x_\mu - x_v) \right.. \quad (1A.6)$$

In (1A.5) $[x, x_1, \dots, x_m]$ is a *linear operator* in the sense that for two arbitrary functions f, g and two arbitrary constants a, b we have

$$[x, x_1, \dots, x_m], (af + bg) = a[x, x_1, \dots, x_m], f + b[x, x_1, \dots, x_m], g,$$

provided the right-hand expression exists.

In the symbol (1A.5) the subscript t can be dropped if f is introduced as a function of one variable only. This expression is also sometimes written as $f(x, x_1, \dots, x_m)$.

From the symmetry of the right-hand expression in (1A.5) follows (if we interchange x with x_k , replace m by $m-1$, and correspondingly shift the indices) the formula

$$[x_1, \dots, x_m] f = [x_1, \dots, x_k]_{x_k} [x_k, \dots, x_m] f; \quad (1A.7)$$

more generally we can write

$$[x_1, \dots, x_m] f = [x_1, \dots, x_k]_u [u, x_{k+1}, \dots, x_m] f, \quad (1A.8)$$

where the target variable u is unconnected with f and the x_μ .

Iterating (1A.7), we can write, if we have k groups of variables,

$$x_1, \dots, x_{m_1+1}, \quad y_1, \dots, y_{m_2+1}, \dots, z_1, \dots, z_{m_k+1}, \quad (1A.9)$$

$$\begin{aligned} & [x_1, \dots, z_{m_k+1}] f \\ &= [x_1, \dots, x_{m_1+1}]_x [y_1, \dots, y_{m_2+1}]_y \cdots [z_1, \dots, z_{m_k+1}]_z [x, y, \dots, z] f, \end{aligned} \quad (1A.10)$$

where we must assume that all variables x_1, \dots, z_{m_k+1} remain distinct.

5. As an example, consider $f(x) = x^m$. We obtain recurrently, denoting by p, q_1, \dots, q_m nonnegative integers,

$$\begin{aligned} [x_1, x] x^m &= \frac{x_1^m - x^m}{x_1 - x} = \sum x^p x_1^{q_1} \quad (p + q_1 = m-1), \\ [x_2, x] [x_1, x] x^m &= \sum_{p+q_1=m-1} x_1^{q_1} [x_2, x] x^p = \sum x_1^{q_1} x^p x_2^{q_2} \\ &\quad (p + q_1 + q_2 = m-2), \\ [x_k, x_{k-1}, \dots, x_1, x] x^m &= \sum x^p x_1^{q_1} \cdots x_k^{q_k} \\ &\quad (p + q_1 + \cdots + q_k = m-k), \end{aligned} \quad (1A.11)$$

$$\begin{aligned} [x_{m-1}, \dots, x_1, x] x^m &= x_{m-1} + x_{m-2} + \cdots + x_1 + x, \\ [x_m, x_{m-1}, \dots, x_1, x] x^m &= 1, \quad [x_{m+1}, x_m, \dots, x_1, x] x^m = 0. \end{aligned} \quad (1A.12)$$

The divided difference of the order $>m$, applied to a polynomial of degree $\leq m$, gives 0.

HERMITE'S INTEGRAL REPRESENTATION

6. We will denote the smallest convex (closed) polygonal domain containing all x, x_v ($v = 1, \dots, m$) by $\langle x, x_1, \dots, x_m \rangle$ and in particular, in the case of

real x and x_v , the corresponding interval by $\langle x, x_1, \dots, x_m \rangle$. Both reduce to a point if all x, x_1, \dots, x_m coincide. Assume that $f^{(m)}(t)$ is continuous in $\langle x, x_1, \dots, x_m \rangle$ or, if not all x, x_v are real, analytic in $\langle x, x_1, \dots, x_m \rangle$. We have obviously

$$[x_1, x_2] f = \frac{f(x_1) - f(x_2)}{x_1 - x_2} = \frac{1}{x_2 - x_1} \int_{x_1}^{x_2} f'(t) dt.$$

Introducing here the new variable of integration t_1 by $t = (1-t_1)x_1 + t_1 x_2$, $dt = (x_2 - x_1) dt_1$, we obtain, replacing x_2 by x ,

$$[x_1, x] f = \int_0^1 f'[(1-t_1)x_1 + t_1 x] dt_1.$$

Applying the operator $[x_2, x]$ to both sides and differentiating under the sign of integration, we get similarly

$$[x_2, x_1, x] f = \int_0^1 dt_1 \int_0^1 t_1 f''[(1-t_1)x_1 + t_1(1-t_2)x_2 + t_1 t_2 x] dt_2;$$

here we introduce in the inner integral the new variable of integration $t_2 = t_1 t_2$, $dt_2 = t_1 dt_2$, and obtain

$$[x_2, x_1, x] f = \int_0^1 \int_0^{t_1} f''[(1-t_1)x_1 + (t_1 - t_2)x_2 + t_2 x] dt_2 dt_1.$$

7. Proceeding in the same way, we get

$$\begin{aligned} [x_m, \dots, x_1, x] f &= \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{m-1}} f^{(m)}[(1-t_1)x_1 + (t_1 - t_2)x_2 + \cdots \\ &\quad + (t_{m-1} - t_m)x_m + t_m x] dt_m \cdots dt_1, \end{aligned} \quad (1A.13)$$

and this formula is immediately verified by induction.

For $f = x^m/m!$ the left-hand expression in (1A.13) becomes, by (1A.12), $1/m!$ and $f^{(m)} \equiv 1$.

We see that we have

$$\int_0^1 \int_0^{t_1} \cdots \int_0^{t_{m-1}} dt_m \cdots dt_1 = \frac{1}{m!}. \quad (1A.14)$$

Observe that all coefficients $1-t_1, t_1-t_2, \dots, t_{m-1}-t_m, t_m$ in the argument of $f^{(m)}$ in (1A.13) are *nonnegative*. One consequence of this is that the argument

$$(1-t_1)x_1 + (t_1 - t_2)x_2 + \cdots + (t_{m-1} - t_m)x_m + t_m x$$

lies both in the real and in the complex case in $\langle x, x_1, \dots, x_m \rangle$.

MEAN VALUE FORMULAS

8. Suppose now first that x and all x_i are *real* and denote by Ω and ω the maximum and the minimum of $f^{(m)}(t)$ in $\langle x, x_1, \dots, x_m \rangle$. Then we obtain an upper and a lower bound of expression (1A.13), replacing there $f^{(m)}$ by the constants Ω and ω , for instance, replacing f by $\Omega x^m/m!$ and $\omega x^m/m!$. We see that (1A.13) is contained between $\Omega/m!$ and $\omega/m!$. We have therefore, as $f^{(m)}$ is assumed to be continuous,

$$[x_m, \dots, x_1, x] f = \frac{f^{(m)}(\xi)}{m!}, \quad \xi \in \langle x, x_1, \dots, x_m \rangle. \quad (1A.15)$$

If x, x_i are not all real, the modulus of $f^{(m)}$ assumes its maximum in a point ξ of $\langle x, x_1, \dots, x_m \rangle$ and the modulus of (1A.13) is then

$$\leq \frac{|f^{(m)}(\xi)|}{m!},$$

so that we have

$$[x_m, \dots, x_1, x] f = \theta \frac{f^{(m)}(\xi)}{m!}, \quad |\theta| \leq 1, \quad \xi \in \langle x, x_1, \dots, x_m \rangle. \quad (1A.16)$$

9. In the complex case, and also in the real case if $f^{(m+1)}(x)$ is continuous in $\langle x, x_1, \dots, x_m \rangle$, we can obtain a better result. Put

$$M_{m+1} = \max |f^{(m+1)}(t)| \quad (t \in \langle x, x_1, \dots, x_m \rangle).$$

Then we have, denoting by ζ an arbitrarily chosen point of $\langle x, x_1, \dots, x_m \rangle$, subtracting $f^{(m)}(\zeta)/m!$ from both sides of (1A.13), and using (1A.14),

$$\begin{aligned} [x_m, \dots, x_1, x] f - \frac{f^{(m)}(\zeta)}{m!} &= \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{m-1}} F(t_1, \dots, t_m) dt_m \cdots dt_1, \\ F(t_1, \dots, t_m) &= f^{(m)}[(1-t_1)x_1 + \cdots \\ &\quad + (t_{m-1}-t_m)x_m + t_m x] - f^{(m)}(\zeta). \end{aligned}$$

On the other hand, if D is the greatest distance between two points of $\langle x, x_1, \dots, x_m \rangle$ (the *diameter* of $\langle x, x_1, \dots, x_m \rangle$), we have $|F(t_1, \dots, t_m)| \leq DM_{m+1}$, and therefore

$$\begin{aligned} \left| \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{m-1}} F(t_1, \dots, t_m) dt_m \cdots dt_1 \right| &\leq \frac{DM_{m+1}}{m!}, \\ [x_m, \dots, x_1, x] f &= \frac{f^{(m)}(\zeta)}{m!} + \theta D \frac{M_{m+1}}{m!}, \quad |\theta| \leq 1. \quad (1A.17) \end{aligned}$$

10. In order to generalize (1A.13), we must write it in a more compact way, introducing convenient notations. In formula (1A.13), we have an operator which can be written as

$$J_{m_1}(t) := \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{m_1-1}} dt_{m_1} \quad (1A.18)$$

and which has to be applied to a function of t_1, \dots, t_{m_1} , continuous in the corresponding m_1 -dimensional domain. It is then clear what has to be understood by the symbols

$$J_{m_2}(u), \dots, J_{m_k}(v).$$

Further, we write, if x_1, \dots, x_{m_1+1} are given,

$$X_t := (1-t_1)x_1 + (t_1-t_2)x_2 + \cdots + (t_{m_1-1}-t_{m_1})x_{m_1} + t_{m_1}x_{m_1+1}. \quad (1A.19)$$

It is then clear what has to be understood by the symbols

$$Y_u, Z_v, \dots$$

11. Formula (1A.13) obviously can be written as

$$[x_1, \dots, x_{m_1+1}]_x f(x) = J_{m_1}(t) f^{(m_1)}(X_t). \quad (1A.20)$$

Consider now a function of k variables $x, y, \dots, z, F(x, y, \dots, z)$, and assume that the derivative $D_x^{m_1} D_y^{m_2} \cdots D_z^{m_k} F(x, y, \dots, z)$ exists and is continuous if x runs through $U(x) := \langle x_1, \dots, x_{m_1+1} \rangle$, y runs through $U(y) := \langle y_1, \dots, y_{m_2+1} \rangle$, ..., z runs through $U(z) := \langle z_1, \dots, z_{m_k+1} \rangle$. Then we can repeatedly apply (1A.20) and obtain, putting

$$\Omega := [x_1, \dots, x_{m_1+1}]_x [y_1, \dots, y_{m_2+1}]_y \cdots [z_1, \dots, z_{m_k+1}]_z F(x, y, \dots, z), \quad (1A.21)$$

$$\Omega = J_{m_1}(t) J_{m_2}(u) \cdots J_{m_k}(v) D_{X_t}^{m_1} D_{Y_u}^{m_2} \cdots D_{Z_v}^{m_k} F(X_t, Y_u, \dots, Z_v). \quad (1A.22)$$

In this formula we must of course assume that all variables $x_1, \dots, x_{m_1+1}, y_1, \dots, y_{m_2+1}, \dots, z_1, \dots, z_{m_k+1}$ are distinct.

Applying formula (1A.20) to the function x^n , for which

$$[x_1, \dots, x_{n+1}] x^n \equiv 1,$$

we obtain

$$(x^n)^{(n)} = n!, \quad J_n(t) 1 = \frac{1}{n!}. \quad (1A.23)$$

12. Applying (1A.21) and (1A.22) to formula (1A.10) and replacing $F(x, y, \dots, z)$ with $[x, y, \dots, z] f$, we obtain finally

$$\begin{aligned} & [x_1, \dots, x_{m_1+1}, y_1, \dots, y_{m_2+1}, \dots, z_1, \dots, z_{m_k+1}] f \\ &= J_{m_1}(t) J_{m_2}(u) \cdots J_{m_k}(v) D_{X_t}^{m_1} D_{Y_u}^{m_2} \cdots D_{Z_v}^{m_k} [X_t, Y_u, \dots, Z_v] f. \end{aligned} \quad (1A.24)$$

In this formula we assume as above that all variables x_1, \dots, z_{m_k+1} are *distinct*.

As to the condition of differentiability of F , observe that $[X_t, Y_u, \dots, Z_v]f$ is a linear combination of $f(X_t), f(Y_u), \dots, f(Z_v)$ with coefficients which are rational functions of X_t, Y_u, \dots, Z_v . Our differentiability condition is therefore satisfied if $D_x^{m_1}f$ is continuous in $U(x)$, $D_y^{m_2}f$ is continuous in $U(y), \dots, D_z^{m_k}f$ is continuous in $U(z)$. Observe that expression (1A.24) contains derivatives of f of *lower order* than in Hermite's integral formula.

If all variables x_1, \dots, z_{m_k+1} and $f(x)$ are real, we obtain at once from (1A.24) and (1A.23) the *mean value formula*

$$[x_1, \dots, x_{m_1+1}, \dots, z_1, \dots, z_{m_k+1}]f = \frac{1}{m_1! m_2! \cdots m_k!} D_\xi^{m_1} D_\eta^{m_2} \cdots D_\zeta^{m_k} [\xi, \eta, \dots, \zeta] f, \\ \xi \in U(x), \quad \eta \in U(y), \dots, \zeta \in U(z), \quad (1A.25)$$

while in the general case of complex variables we have the *estimate*

$$|[x_1, \dots, z_{m_k+1}]f| \leq \frac{1}{m_1! m_2! \cdots m_k!} \text{Max} |D_\xi^{m_1} D_\eta^{m_2} \cdots D_\zeta^{m_k} [\xi, \eta, \dots, \zeta] f|, \quad (1A.26)$$

where the maximum has to be taken as ξ runs through $U(x)$, η runs through $U(y), \dots, \zeta$ runs through $U(z)$.

For the possibility of dispensing with the continuity of the n th derivatives in the above formulas, see Appendix M.

1B

Confluent Case. Interpolation

CONFLUENT DIVIDED DIFFERENCES

1. In view of definition (1A.2), it is only natural to write

$$[x^2]f := [x, x]f := \lim_{x_1 \wedge x_2 \rightarrow x} [x_1, x_2]f = f'(x). \quad (1B.1)$$

In the general case of (1A.5), if all variables x_1, \dots, x_{m+1} become $=x$, we will define

$$\begin{aligned} [x^{m+1}]f &:= [x, \dots, x]f := \lim [x_1, \dots, x_{m+1}]f \\ (x_1 \wedge x_2 \wedge \dots \wedge x_{m+1} \rightarrow x, \quad x_1 \neq x_2 \neq \dots \neq x_{m+1}), \end{aligned} \quad (1B.2)$$

where in the second term x has to be repeated $m+1$ times. Of course, (1B.2) holds only if the right-hand limit exists.

Assume for the variables (1A.9) that all x_ν become $=x$, all y_μ become $=y, \dots$, all z_λ become $=z$, while $x \neq y \neq \dots \neq z$. Then we will have to define the corresponding *confluent divided difference* by

$$\begin{aligned} [x^{m_1+1}, y^{m_2+1}, \dots, z^{m_k+1}]f &:= \lim [x_1, \dots, y_1, \dots, z_1, \dots, z_{m_k+1}]f \\ (x_v \rightarrow x, y_\mu \rightarrow y, \dots, z_\lambda \rightarrow z, \quad x \neq y \neq \dots \neq z), \end{aligned} \quad (1B.3)$$

where during the limiting process all variables x_1, \dots, z_{m_k+1} are assumed to remain *distinct* and of course the existence of the right-hand limit must be assumed.

2. We can now formulate reasonable conditions ensuring the existence of the limits (1B.2), (1B.3).

Theorem 1B.1. Assume k distinct variables x, y, \dots, z . Assume that f belongs to the continuity class C^{m_1} in a neighborhood of x , to the class C^{m_2} in a neighborhood of y, \dots , to the class C^{m_k} in a neighborhood of z . Then the confluent divided difference, defined by (1B.3) for $k > 1$ and by (1B.2) for $k = 1$, exists and is given by the formula

$$[x^{m_1+1}, y^{m_2+1}, \dots, z^{m_k+1}]f = \frac{1}{m_1! m_2! \cdots m_k!} D_x^{m_1} \cdots D_z^{m_k} [x, y, \dots, z]f. \quad (1B.4)$$

Proof. Observe that in (1A.24), if all $x_v \rightarrow x$, then $X_t \rightarrow x$. Similarly, in the limit implied in formula (1B.3), we have

$$Y_u \rightarrow y, \dots, Z_v \rightarrow z.$$

Since the denominators in the integrand of (1A.24) are products of differences like $X_t - Y_u$ and have therefore limits $\neq 0$, it follows that the integrand in (1A.24) tends to the expression

$$D_x^{m_1} D_y^{m_2} \cdots D_z^{m_k} [x, y, \dots, z] f,$$

which is a constant with respect to the integration variables $t_v, u_\mu, \dots, v_\lambda$. The existence of the limit in (1B.3) and (1B.2) follows immediately and the formula (1B.4) from (1A.23). Theorem 1B.1 is proved.

For a representation of divided differences as quotients of determinants see Appendix L.

A counterexample showing that the continuity conditions imposed on derivatives in the confluent case cannot be dispensed with is given in Appendix M.

CONTINUITY OF CONFLUENT DIVIDED DIFFERENCES

3. It is now easy to see that (1B.3) still holds, if

$$x_v \rightarrow x \ (v = 1, \dots, m_1), \quad y_v \rightarrow y \ (v = 1, \dots, m_2), \dots, z_v \rightarrow z \ (v = 1, \dots, m_k),$$

and if $x \neq y \neq \dots \neq z$, where, however, we can drop the assumption that all x_v remain distinct, all y_v remain distinct, etc. To see this, denote the value of the left-hand expression in (1B.3) by A and consider an open spherical neighborhood $V(A)$ of A and the open neighborhood $V'(A)$ of A with half the radius of $V(A)$. Then, by what has been proved, there exist neighborhoods $U(x), \dots, U(z)$ such that if all x_v lie in $U(x)$, remaining distinct, ..., and all z_v lie in $U(z)$, remaining distinct, the values of

$$[x_1, \dots, y_1, \dots, z_1, \dots, z_{m_k+1}] f(x) \tag{1B.5}$$

lie in $V'(A)$. If we now drop the assumption that the groups $\{x_v\}, \{y_v\}, \dots, \{z_v\}$ consist only of distinct terms, but still assume that each lies in the corresponding neighborhood $U(x), \dots, U(z)$, then they can be approximated by groups of distinct variables and therefore the values of (1B.5) remain in the closure of $V'(A)$, and therefore in $V(A)$. This proves our assertion.

4. Theorem 1B.2. Consider $[x_1, \dots, x_n] f(x)$ where each x_v runs through an open domain Ω_v either on the real line or on the complex plane. We assume further that $f(x)$ is continuous on $\bigcup_v \Omega_v$, that it belongs to the class C^1

on $\bigcup_{v \neq \mu} \Omega_v \cap \Omega_\mu$, to the class C^2 on $\bigcup_{v \neq \mu \neq \lambda} \Omega_v \cap \Omega_\mu \cap \Omega_\lambda$, and so on. Then $[x_1, \dots, x_n] f(x)$ is continuous in the product domain $\Omega_1 \times \Omega_2 \times \dots \times \Omega_n$.

The proof follows from the result of the preceding section.

The product domain $\Omega_1 \times \dots \times \Omega_n$ is also open. If now K is a closed subset of this domain it follows from the above that our divided difference is *uniformly continuous* on K .

5. It is easy to see that formula (1A.13) remains valid if we drop the assumption that all x_v are distinct.

Indeed, if all x_v have the same value x , then the left-hand expression in (1A.13) is, by (1B.4), $f^{(m)}(x)/m!$, while the right-hand integrand in (1A.13) is independent of the t_v and $= f^{(m)}(x)$. In this case formula (1A.13) follows from (1A.23).

If not all x_v have the same value, the neighborhood $U(x) := \langle x_1, \dots, x_{m+1} \rangle$ is either an interval of positive length or a polygon with positive area. There exist therefore systems of *distinct* $y_1^{(\mu)}, \dots, y_{m+1}^{(\mu)}$ from $U(x)$ such that

$$y_v^{(\mu)} \rightarrow x_v \quad (\mu \rightarrow \infty, \quad v = 1, \dots, m+1).$$

Writing formula (1A.13) for the $y_v^{(\mu)}$, we obtain with $\mu \rightarrow \infty$ the formula (1A.13) *using the assumption* that $f^{(m)}(x)$ exists in $U(x)$ and is there continuous with respect to $U(x)$.

Applying now the argument of Sections 11 and 12 of Chapter 1A, we obtain the validity of formulas (1A.22) and (1A.24) using the assumption that $x_v \neq y_\mu \neq \dots \neq z_\lambda$ where however *neither the $m_1 + 1$ variables x_v , nor the $m_2 + 1$ variables y_μ, \dots , nor the $m_k + 1$ variables z_λ need be distinct*. The differentiability conditions remain the same as in Section 12 of Chapter 1A.

In the case of real variables and real f we again obtain formula (1A.25) and in the complex case formula (1A.26), where in both cases neither the x_v , nor the y_μ, \dots , nor the z_λ need be assumed as distinct, while we must assume that $x_v \neq y_\mu \neq \dots \neq z_\lambda$. The differentiability conditions remain the same as in Section 12 of Chapter 1A.

VARIOUS FORMULAS FOR DIVIDED DIFFERENCES

6. In what follows we give different formulas for divided differences, which hold also in the confluent case with the restrictions mentioned if necessary. The proofs of these formulas may be left to the reader as exercises.

$$(1) \quad [x_1, \dots, x_n] f(x+c) = [x_1 + c, \dots, x_n + c] f(x) \quad (c \text{ constant}).$$

$$(2) \quad [cx_1, \dots, cx_n] f(x) = [x_1, \dots, x_n] f(cx) c^{-n+1} \quad (c \neq 0 \text{ constant}).$$

$$(3) \quad [x_1, x_2] f g = f(x_1)[x_1, x_2] g + g(x_2)[x_1, x_2] f \\ = \frac{1}{2}(f(x_1) + f(x_2))[x_1, x_2] g + \frac{1}{2}(g(x_1) + g(x_2))[x_1, x_2] f.$$

$$(4) \quad [x_1, \dots, x_n] \frac{1}{x} = \frac{(-1)^{n-1}}{x_1 \cdots x_n} \quad (x_1 \wedge \cdots \wedge x_n \neq 0).$$

$$(5) \quad [x_1, \dots, x_n] \frac{1}{x^2} = \frac{(-1)^{n-1}}{x_1 \cdots x_n} \left(\frac{1}{x_1} + \cdots + \frac{1}{x_n} \right) \quad (x_1 \wedge \cdots \wedge x_n \neq 0).$$

$$(6) \quad (x_1 - x_2)[x_1, x_2, t_1, \dots, t_n] f + (x_2 - x_3)[x_2, x_3, t_1, \dots, t_n] f \\ + (x_3 - x_1)[x_3, x_1, t_1, \dots, t_n] f = 0.$$

$$(7) \quad [x_0, x_1, \dots, x_n] f = \int f^{(n)}(t_0 x_0 + t_1 x_1 + \cdots + t_n x_n) dt_1 \cdots dt_n \\ (0 \leq t_1 + \cdots + t_n \leq 1, \quad 0 \leq t_v \leq 1, \quad t_0 = 1 - t_1 - \cdots - t_n) \quad (\text{Genocchi}).$$

$$(8) \quad [x_0, x_1, \dots, x_n] f \\ = \int_0^1 \int_0^1 \cdots \int_0^1 t_1^{n-1} t_2^{n-2} \cdots t_{n-1} f^{(n)}(x_0 + (x_1 - x_0)t_1 + (x_2 - x_1)t_1 t_2 \\ + \cdots + (x_n - x_{n-1})t_1 \cdots t_n) dt_1 \cdots dt_n \quad (\text{Genocchi}).$$

7. (9) $[x_1^{m_1}, \dots, x_k^{m_k}] f$

$$= \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{k-2}} dt_{k-1} \varphi(t_1, \dots, t_{k-1}) f^{(m_1 + \cdots + m_k - 1)} \\ \times [(1 - t_1)x_1 + (t_1 - t_2)x_2 + \cdots + t_{k-1}x_k],$$

$$\varphi(t_1, \dots, t_{k-1}) = \frac{(1 - t_1)^{m_1 - 1}}{(m_1 - 1)!} \frac{(t_1 - t_2)^{m_2 - 1}}{(m_2 - 1)!} \cdots \frac{(t_{k-1})^{m_k - 1}}{(m_k - 1)!} \quad (\text{Milne-Thomson}).$$

(10) Assume $f(x) \in C^n$ for real x from the closed interval J and $x_1 \wedge \cdots \wedge x_n \in J$. Then for $x \in J$, $[x_1, \dots, x_n, x] f$ is monotonic in x with $f^{(n)}(x)$ and in the same sense; it is convex in x if $f^{(n)}(x)$ is convex in x (Cauchy).

(11) If $x \neq y$, we have

$$[x^{n+1}, y] f = (y - x)^{-n-1} \left(f(y) - \sum_{v=0}^n \frac{(y - x)^v}{v!} f^{(v)}(x) \right).$$

(12) Formulas (1A.11) and (1A.12) are valid in the general case.

(13) If the k variables x, y, \dots, z are distinct and n, m, \dots, l nonnegative

integers, the following identity holds for any integer $N \geq n+m+\dots+l+k-1$:

$$[x^{n+1}, y^{m+1}, \dots, z^{l+1}] x^N = \sum \binom{p+n}{n} \binom{q+m}{m} \dots \binom{r+l}{l} x^p y^q \dots z^r$$

$$(p \wedge q \dots \wedge r \geq 0, \quad p+q+\dots+r = N-n-m-\dots-l-k+1).$$

(14) If each of the variables x, y, \dots, z is different from each of the variables ξ, η, \dots, ζ and if f satisfies the differentiability conditions resulting from Theorem 1B.1, then

$$[x^n, y^m, \dots, z^l, \xi^v, \eta^u, \dots, \zeta^k] f = [x^n, y^m, \dots, z^l]_u [\xi^v, \eta^u, \dots, \zeta^k] f.$$

(15) Consider k distinct variables x, y, \dots, z , each of which is different from each of the variables t_1, \dots, t_ω , and assume that f satisfies the differentiability conditions implied by Theorem 1B.1. Then

$$[x^n, y^m, \dots, z^l, t_1, \dots, t_\omega] f = [x^n]_u [y^m]_v \dots [z^l]_w [u, v, \dots, w, t_1, \dots, t_\omega] f.$$

(16) If x is different from all variables t_1, \dots, t_ω , then

$$[x^{n+1}]_x [x^{m+1}, t_1, \dots, t_\omega] f = \binom{n+m}{m} [x^{n+m+1}, t_1, \dots, t_\omega] f \quad (n \wedge m > 0).$$

NEWTON'S INTERPOLATION FORMULA

8. Consider a function $f(x)$ of one variable x and $n+1$ *distinct* points x, x_1, \dots, x_n . Then we can write (1A.1) in the form

$$f(x) = f(x_1) + (x-x_1)[x, x_1] f.$$

Applying the same formula to $[x, x_1] f$ as a function of x , we have

$$[x, x_1] f = [x_2, x_1] f + (x-x_2)[x, x_1, x_2] f$$

and generally

$$[x, x_1, \dots, x_{v-1}] f = [x_1, \dots, x_v] f + (x-x_v)[x, x_1, \dots, x_v] f \quad (1B.6)$$

until we have for $v=n$

$$[x, x_1, \dots, x_{n-1}] f = [x_1, \dots, x_n] f + (x-x_n)[x, x_1, \dots, x_n] f.$$

In order to eliminate $[x, x_1] f, [x, x_1, x_2] f, \dots$, we introduce Runge's notation

$$(x-x_1)(x-x_2) \dots (x-x_\mu) =: \mathbb{x}^\mu \quad (\mu = 1, 2, \dots); \quad \mathbb{x}^0 := 1. \quad (1B.7)$$

Obviously we have generally

$$\mathbb{x}^{\mu-1}(x-x_\mu) = \mathbb{x}^\mu \quad (\mu = 1, 2, \dots). \quad (1B.8)$$

9. Multiplying now (1B.6) by \otimes^{v-1} , we obtain by (1B.8)

$$\otimes^{v-1}[x, x_1, \dots, x_{v-1}] f = \otimes^{v-1}[x_1, \dots, x_v] f + \otimes^v[x, x_1, \dots, x_v] f,$$

and summing over $v = 1, 2, \dots, n$, we get finally, as the last $n-1$ terms on the left are canceled out,

$$f(x) = \sum_{v=1}^n \otimes^{v-1}[x_1, \dots, x_v] f + \otimes^n[x, x_1, \dots, x_n] f$$

or, denoting the first sum on the right by $L_{n-1}(f, x)$,

$$f(x) = L_{n-1}(f, x) + R_{n-1}(f, x), \quad (1B.9)$$

where we can write

$$\begin{aligned} L_{n-1}(f, x) &= f(x_1) + (x - x_1)[x_1, x_2] f + (x - x_1)(x - x_2)[x_1, x_2, x_3] f \\ &\quad + \cdots + (x - x_1) \cdots (x - x_{n-1})[x_1, \dots, x_n] f, \end{aligned} \quad (1B.10)$$

and for the "remainder term"

$$R_{n-1}(f, x) = (x - x_1) \cdots (x - x_n)[x, x_1, \dots, x_n] f. \quad (1B.11)$$

10. This is the classical *interpolation formula of Newton*, in the case of distinct interpolation points x_v . For real or not necessarily real x_1, \dots, x_n we can write R_{n-1} correspondingly, by (1A.15), (1A.16), or (1A.17), in one of the forms

$$R_{n-1}(f, x) = (x - x_1) \cdots (x - x_n) \frac{f^{(n)}(\xi)}{n!}, \quad (1B.12)$$

$$R_{n-1}(f, x) = (x - x_1) \cdots (x - x_n) \theta \frac{f^{(n)}(\xi)}{n!}, \quad (1B.13)$$

with $|\theta| \leq 1$ and a convenient ξ from $\langle x, x_1, \dots, x_n \rangle$ if $f(x)$ is assumed to have continuous n th derivatives there; finally, under conditions corresponding to those of Section 9 of Chapter 1A,

$$R_{n-1}(f, x) = \frac{(x - x_1) \cdots (x - x_n)}{n!} (f^{(n)}(\zeta) + \theta D M_{n+1}), \quad (1B.14)$$

where ζ is an arbitrarily chosen point from $\langle x, x_1, \dots, x_n \rangle$, $|\theta| \leq 1$, D the diameter of $\langle x, x_1, \dots, x_n \rangle$, and M_{n+1} an upper bound of $|f^{(n+1)}(x)|$ in $\langle x, x_1, \dots, x_n \rangle$ if $f^{(n+1)}(x)$ is continuous there.

Formula (1B.9) is an *identity*; we can therefore let groups of interpolation points become equal as in Section 1. We replace there the limiting values x, y, \dots, z with t_1, t_2, \dots, t_k . Then R_{n-1} becomes, for instance,

$$R_{n-1}(f, x) = (x - t_1)^{m_1} \cdots (x - t_k)^{m_k} [x, t^{m_1}, \dots, t^{m_k}] f. \quad (1B.15)$$

The corresponding limiting process is allowed if $f(x)$ has continuous n th derivatives in the domain $\langle x, t_1, \dots, t_k \rangle$ and in its neighborhood.

GENERAL INTERPOLATION PROBLEM

11. In order to make clear the significance of the polynomial $L_{n-1}(f, x)$, observe that by (1B.4) $f(x)$ enters into $L_{n-1}(f, x)$ and $R_{n-1}(f, x)$ only with the values

$$f^{(\mu)}(t_\kappa) \quad (\mu = 0, 1, \dots, m_\kappa - 1; \quad \kappa = 1, \dots, k). \quad (1B.16)$$

If therefore a function $\gamma(x)$ satisfies the conditions

$$\gamma^{(\mu)}(t_\kappa) = f^{(\mu)}(t_\kappa) \quad (\mu = 0, \dots, m_\kappa - 1; \quad \kappa = 1, \dots, k) \quad (1B.17)$$

we have

$$L_{n-1}(f, x) = L_{n-1}(\gamma, x).$$

Suppose now that the function $\gamma(x)$ satisfying (1B.17) is a polynomial of degree not exceeding $n-1$. Then, if we apply (1B.13) to $R_{n-1}(\gamma, x)$ it follows that $R_{n-1}(\gamma, x) = 0$ and from $\gamma(x) = L_{n-1}(f, x) + R_{n-1}(\gamma, x)$ that $L_{n-1}(f, x) = \gamma(x)$.

We see in particular that the polynomial $\gamma(x)$ of degree $< n$ satisfying (1B.17) is uniquely determined.

12. We are now going to prove that there always exists a polynomial $\gamma(x)$ satisfying (1B.17). We are going to prove even more.

Assuming arbitrarily n numbers

$$A_\kappa^{(\mu)} \quad (\kappa = 1, \dots, k; \quad \mu = 0, 1, \dots, m_\kappa - 1), \quad (1B.18)$$

we prove that there always exists a polynomial $\gamma(x)$ of degree $< n$ satisfying the n conditions

$$\gamma^{(\mu)}(t_\kappa) = A_\kappa^{(\mu)} \quad (\kappa = 1, \dots, k; \quad \mu = 0, \dots, m_\kappa - 1). \quad (1B.19)$$

13. Indeed, if we write $\gamma(x)$ as

$$\gamma(x) = u_0 x^{n-1} + u_1 x^{n-2} + \dots + u_{n-1},$$

Eqs. (1B.19) represent a set of n linear equations in the unknowns u_i , and we have only to prove that the determinant of this set is not zero. But if this determinant were zero, the corresponding set of homogeneous linear equations could have a nontrivial solution and there would exist a nonzero polynomial $\gamma(x)$ of degree $\leq n-1$ satisfying the conditions

$$\gamma^{(\mu)}(t_\kappa) = 0 \quad (\kappa = 1, \dots, k; \quad \mu = 0, 1, \dots, m_\kappa - 1),$$

that is, divisible by $(x - t_1)^{m_1}, \dots, (x - t_k)^{m_k}$ and therefore by the product $(x - t_1)^{m_1} \cdots (x - t_k)^{m_k}$, which is of degree n . This contradiction proves our assertion.

It follows now in particular that $L_{n-1}(f, x)$ is the unique polynomial $\gamma(x)$ of degree $< n$ satisfying the conditions (1B.17):

$$L_{n-1}^{(\mu)}(f, t_\kappa) = f^{(\mu)}(t_\kappa) \quad (\kappa = 1, \dots, k; \quad \mu = 0, 1, \dots, m_\kappa - 1). \quad (1B.20)$$

14. A function $\gamma(t)$ satisfying the conditions (1B.19) is generally called an *interpolating function to the system of values $A_\kappa^{(\mu)}$* , corresponding to the *interpolation points* (or *interpolation abscissas*) t_κ , each t_κ being taken " m_κ times," that is, with the *multiplicity* m_κ . If $A_\kappa^{(\mu)}$ are the values $f^{(\mu)}(t_\kappa)$ of a function $f(t)$, that is, if $\gamma(t)$ satisfies the conditions (1B.17), $\gamma(t)$ is called an *interpolating function to $f(t)$* .

One could choose $\gamma(x)$ from many different classes of functions, e.g. polynomials, trigonometric functions, rational functions of x , etc. In practice we choose for $\gamma(x)$ functions with very familiar properties.

When the n interpolation points are all equal and $\gamma(x)$ is a polynomial of degree $n-1$, it is the well-known Taylor polynomial.

POLYNOMIAL INTERPOLATION

15. If we have $m_1 + \dots + m_k = n$, the interpolating function $\gamma(t)$ can be chosen as a polynomial of degree $< n$. We speak in this case of the *polynomial interpolation* without any further qualifications although, of course, an interpolation problem (1B.19) could also be considered for polynomials $\gamma(t)$ not necessarily of degree $< n$ but satisfying some other special or general conditions.

If all multiplicities m_κ are equal to 1, the interpolating polynomial $L_{n-1}(f, x)$ can be written in a very elegant way found by Lagrange and derived in any textbook of higher algebra:

Let

$$F(x) = \prod_{v=1}^n (x - t_v).$$

Then we have

$$L_{n-1}(f, x) = \sum_{v=1}^n f(t_v) \frac{F(x)}{(x - t_v) F'(t_v)}. \quad (1B.21)$$

THE REMAINDER FOR A GENERAL INTERPOLATING FUNCTION

16. If $\gamma(x)$ is a general interpolating function satisfying the conditions (1B.17), we can easily obtain an expression for the "remainder" $f(x) - \gamma(x)$

corresponding to (1B.15), if we assume that $f^{(n)}(t)$ as well as $\gamma^{(n)}(t)$ are continuous in $\langle x, t_1, \dots, t_k \rangle$. Indeed, putting $F(x) \equiv f(x) - \gamma(x)$, we see that $L_{n-1}(F, x)$ vanishes identically so that we have

$$f(x) - \gamma(x) = R_{n-1}(F, x) = \prod_{v=1}^n (x - x_v) \prod_{\kappa=1}^k [x, t_\kappa^{m_\kappa}] (f - \gamma) \quad (1B.22)$$

and therefore, *in the real case*, for a ξ from $\langle x, t_1, \dots, t_k \rangle$,

$$f(x) - \gamma(x) = \prod_{\kappa=1}^k (x - t_\kappa)^{m_\kappa} \frac{f^{(n)}(\xi) - \gamma^{(n)}(\xi)}{n!}, \quad \xi \langle x, t_1, \dots, t_k \rangle. \quad (1B.23)$$

TRIANGULAR SCHEMES FOR COMPUTING DIVIDED DIFFERENCES

17. In the praxis the divided differences are computed using a difference scheme similar to that used in the theory of differences with the constant step ω :

x_1	$f(x_1)$			
		$[x_1, x_2] f$		
x_2	$f(x_2)$		$[x_1, x_2, x_3] f$	
		$[x_2, x_3] f$		$[x_1, x_2, x_3, x_4] f$
x_3	$f(x_3)$		$[x_2, x_3, x_4] f$	
		$[x_3, x_4] f$		
x_4	$f(x_4)$			
			$[x_{m-2}, x_{m-1}, x_m] f$	
			$[x_{m-1}, x_m] f$	
x_m	$f(x_m)$			

However, going in this scheme from one column to the next one we must *divide by the difference of the corresponding arguments*, while no such division is usual in the difference scheme in the case of a constant step.

18. In the confluent case a similar scheme can be used. If we consider for instance, $[x_1^3, x_2^2, x_3]f$, the corresponding scheme would be as follows.

$$\begin{array}{lll}
 x_1 f(x_1) & & \\
 f'(x_1) & & \\
 x_1 f(x_1) & \frac{1}{2} f''(x_1) & \\
 f'(x_1) & \frac{1}{2} \frac{\partial^2}{\partial x_1^2} [x_1, x_2]f & \\
 x_1 f(x_1) & \frac{\partial}{\partial x_1} [x_1, x_2]f & \frac{1}{2} \frac{\partial^3}{\partial x_1^2 \partial x_2} [x_1, x_2]f \\
 [x_1, x_2]f & \frac{\partial^2}{\partial x_1 \partial x_2} [x_1, x_2]f & \frac{1}{2} \frac{\partial^3}{\partial x_1^2 \partial x_2} [x_1, x_2, x_3]f \\
 x_2 f(x_2) & \frac{\partial}{\partial x_2} [x_1, x_2]f & \frac{\partial^2}{\partial x_1 \partial x_2} [x_1, x_2, x_3]f \\
 f'(x_2) & \frac{\partial}{\partial x_2} [x_1, x_2, x_3]f & \\
 x_2 f(x_2) & \frac{\partial}{\partial x_2} [x_2, x_3]f & \\
 [x_2, x_3]f & & \\
 x_3 f(x_3) & &
 \end{array}$$

2

Inverse Interpolation. Derivatives of the Inverse Function. One Interpolation Point.

THE CONCEPT OF INVERSE INTERPOLATION

1. If a number a is used as an approximation for a number x , we write $a \approx x$; the same notation is used for approximating functions.

One may approach the problem of finding the zeros of a function $f(x)$ in two different ways. We may equate an interpolating function $T(x)$ of $f(x)$ to zero, $T(x) = 0$, and find the roots of this equation. The question arises whether the roots so obtained will be approximations of the roots of $f(x) = 0$. It is well known that by changing the coefficients of an algebraic equation very slightly, we may get roots which differ considerably from the roots of the original equation (see Appendix A). This problem will be discussed later in greater detail, and we shall establish conditions under which the roots of $T(x) = 0$ are approximations of the roots of $f(x) = 0$ (see Appendices A, B, and K).

The second approach to the problem of finding the roots of $f(x) = 0$ is by using the *inverse function*.

Let $y = f(x)$ be defined in J_x and given in n interpolation points x_v , ($v = 1, 2, \dots, n$):

$$f(x_v) = y_v. \quad (2.1)$$

Let $x = \varphi(y)$ be the inverse function of $y = f(x)$. Then $\varphi(y_v) = x_v$. The problem of finding a zero of $f(x)$ now becomes the problem of evaluating $\varphi(0)$.

Let $T(y)$ be an interpolating polynomial $L_{n-1}(\varphi, y)$ for $\varphi(y)$, so that $T(y_v) = x_v$. We now evaluate $T(0) = \varphi(0)$. An estimate of the error involved here may be obtained from (1B.15) and (1A.15):

$$\begin{aligned}\varphi(y) - T(y) &= \frac{\varphi^{(n)}(\eta)}{n!} \prod_{v=1}^n (y - y_v), \\ \varphi(0) - T(0) &= \frac{\varphi^{(n)}(\eta)}{n!} (-1)^n y_1 y_2 \cdots y_n, \quad \eta \in \langle 0, y_1, \dots, y_n \rangle.\end{aligned} \quad (2.2)$$

Notice that if all our interpolation points are close to the value of the root, then the error will be particularly small.

2. The above procedure has been known for many years, but mathematicians have generally been reluctant to use this approach because the problem of discussing the inverse function and its derivatives has been considered a difficult one. These difficulties are really only superficial. One is usually interested in the solution of $f(x) = 0$ in a certain interval. Generally within this interval $f'(x) \neq 0$. Otherwise, all well-known methods (inverse interpolation and so on) generally fail and one must resort to some special device. We shall therefore assume $f'(x) \neq 0$ in the considered interval, and we shall show that with this assumption the difficulties mentioned above are eliminated.

DARBOUX'S THEOREM ON VALUES OF $f'(x)$

3. We first prove a theorem which gives the right background for our hypotheses.

Theorem 2.1. (Darboux) *Let $f(x)$ be defined and continuous in J_x : $a \leq x \leq b$. Suppose that $f'(x)$ exists in J_x and that $f'(a) = A$, $f'(b) = B$. Then all values between A and B are assumed by $f'(x)$ for $x \in J_x$.*

Remark. This property of $f'(x)$ is sometimes incorrectly given as a definition of a continuous function. The assertion of Theorem 2.1 is, of course, trivial if $f'(x)$ is continuous on J_x . The point of Darboux's theorem is just that the continuity of $f'(x)$ is not necessary. The reader is reminded that the existence of $f'(x)$ in an inner point x_0 of J_x means that both the right-hand and the left-hand derivatives in x_0 exist and have the same value.

Proof of Darboux's Theorem. Without loss of generality we can assume $A < B$, since for $A > B$ it would be sufficient to replace $f(x)$ by $-f(x)$. Let C be any number satisfying $A < C < B$. We will prove that $f'(x)$ assumes the value C somewhere in (J_x) . Consider $F(x) = f(x) - Cx$. Then

$$F'(x) = f'(x) - C, \quad F'(a) = A - C < 0, \quad F'(b) = B - C > 0.$$

We thus have a continuous function which has a negative derivative in a and a positive one in b . Hence, $F(x)$ assumes in (J_x) values which are less than $F(a)$ and $F(b)$; $F(x)$ has therefore in J_x a minimum in a point ξ which is *interior* to J_x , and the derivative must vanish in ξ , $f'(\xi) - C = 0$, $f'(\xi) = C$, Q.E.D.

By Darboux's theorem the assumption that $f'(x) \neq 0$ ($x \in J_x$) implies therefore that either $f'(x) > 0$ for all $x \in J_x$ or $f'(x) < 0$ for all $x \in J_x$.

DERIVATIVES OF THE INVERSE FUNCTION

4. Let $f(x)$ be defined in J_x . Assume that $f'(x)$ exists and does not vanish in J_x . Then $f(x)$ is strictly monotonic in J_x and by the well-known existence theorems the inverse function $x = \varphi(y)$ of $y = f(x)$ exists and has a derivative in the corresponding y -interval,

$$x = \varphi(y), \quad \varphi'(y) = \frac{dx}{dy} = \frac{1}{f'} = \frac{1}{y'}. \quad (2.3)$$

If, moreover, $f''(x) = y''$ exists in J_x , it follows by differentiation of $\varphi'(y)$ that

$$\varphi''(y) = \frac{-y''}{y'^3}. \quad (2.4)$$

We see that with the above assumptions, if $f(x)$ possesses first and second derivatives, so does the inverse function. The problem of finding workable expressions for higher derivatives of the inverse function can become quite complicated. We assume the existence of the first $n+1$ ($n \geq 0$) derivatives of $f(x)$ and get a recurrence formula for obtaining the corresponding derivatives of $\varphi(y)$.

5. Let

$$\varphi^{(k)}(y) = \frac{X_k}{y'^{2k-1}} \quad (k = 1, 2, \dots, n+1). \quad (2.5)$$

Here X_k is a *polynomial* in y' , y'' , ..., $y^{(k)}$. This is true for $n = 0, 1$. We have, in particular, $X_1 = 1$, $X_2 = -y''$. Assume the truth of our assertion for the first n derivatives of $\varphi(y)$. We write (2.5) with $k = n$ and get by differentiation, since $dy'/dy = y''/y'$,

$$\begin{aligned} \frac{d}{dx} X_k(y', \dots, y^{(k)}) &= \sum_{\kappa=1}^k \frac{\partial X_k}{\partial y^{(\kappa)}} y^{(\kappa+1)}, \\ \varphi^{(n+1)}(y) &= (X_n)_x' \frac{1}{y'^{2n}} - (2n-1) X_n \frac{y''}{y'} y'^{-2n}. \end{aligned} \quad (2.6)$$

Multiplying (2.6) by y'^{2n+1} , we obtain from (2.5)

$$X_{n+1} = (X_n)_x' y' - (2n-1) X_n y''. \quad (2.7)$$

6. In Sections 6–8 (and in Appendix C) we write y_v instead of $y^{(v)}$. X_n is then a polynomial in y_1, y_2, \dots, y_n .

By induction we see from (2.7) that

$$X_n = \sum a_{\alpha_1 \alpha_2 \dots \alpha_n} y_1^{\alpha_1} y_2^{\alpha_2} \cdots y_n^{\alpha_n} \quad (2.8)$$

is a *homogeneous* polynomial in y_1, y_2, \dots, y_n of the dimension $n-1$; i.e., we have in each term of (2.8)

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = n - 1. \quad (2.9)$$

Indeed, this is true for X_1 and X_2 . If we assume (2.9) true for an n , we see in using (2.7) that the last right-hand term is homogeneous of the dimension n , while the first term can be written in the form

$$y_1 \sum_v y_{v+1} \frac{\partial}{\partial y_v} X_n \quad (2.10)$$

and therefore also becomes of the dimension n .

Further, if to each variable y_v we assign v as the corresponding *weight*, X_n is *isobaric* of the *total weight* $2n-2$, i.e., we have in each term of (2.8)

$$\alpha_1 + 2\alpha_2 + 3\alpha_3 + \cdots + n\alpha_n = 2n - 2. \quad (2.11)$$

This is true for X_1 and X_2 . Assume (2.11) true for an n . Then, in (2.7), the total weight of the last right-hand term is greater by 2 than that of X_n , while the process (2.10) obviously raises the total weight of each term of X_n also by 2.

7. Finally, the *highest term* in (2.8) and also the only term containing y_n is

$$-y_n y_1^{n-2} \quad (n \geq 2), \quad (2.12)$$

while the *lowest term* in X_n and the only term not containing any y_v with $v > 2$ is

$$(-1)^{n-1} 1 \cdot 3 \cdots (2n-3) y_2^{n-1} \quad (n \geq 2). \quad (2.13)$$

Indeed, our assertions are obviously true for X_2 . Suppose that they are true for an n . A term containing y_{n+1} in (2.8) can be obtained only from (2.12) and in particular from

$$y_1 y_{n+1} \frac{\partial}{\partial y_n} (-y_n y_1^{n-2}) = -y_{n+1} y_1^{n-1}.$$

Further it is clear that the first right-hand term in the expression (2.7) contains in every term at least one y_v with $v > 2$, if the only term of X_n depending on y_1 and y_2 is given by (2.13). The one term in X_{n+1} depending only on y_1 and y_2 is therefore obtained from the second right-hand term in (2.7) by multiplying (2.13) by $-(2n-1)y_2$, and our assertion is proved.

In the next section we give a table of X_1, \dots, X_6 . An explicit formula for X_n as well as the discussion of some special cases will be given in Appendix C.

8. *Table of X_1, \dots, X_6 :*

$$X_1 = 1$$

$$X_2 = -y_2$$

$$X_3 = -y_3 y_1 + 3y_2^2$$

$$X_4 = -y_4 y_1^2 + 10y_3 y_2 y_1 - 15y_2^3$$

$$X_5 = -y_5 y_1^3 + 15y_4 y_2 y_1^2 + 10y_3^2 y_1^2 - 105y_3 y_2^2 y_1 + 105y_2^4$$

$$X_6 = -y_6 y_1^4 + 21y_5 y_2 y_1^3 + 35y_4 y_3 y_1^3 - 210y_4 y_2^2 y_1^2$$

$$-280y_3^2 y_2 y_1^2 + 1260y_3 y_2^3 y_1 - 945y_2^5.$$

ONE INTERPOLATION POINT

9. We consider now the case where we have but one interpolation point x_0 . One generally assumes that if $f(x_0)$ is “small,” then x_0 is close to a zero of $f(x)$. Of course, this statement must be qualified, for if $f(x) = 0$ is multiplied by a small number, the roots are not changed. Thus the smallness of $f(x_0)$ could be the result of such a multiplication. We can say more about this by studying the first derivative of $f(x)$.

Hereafter we will use the symbol $\langle a, b \rangle$ to denote the *closed* interval with the end points a and b and $a \geq b$.

Theorem 2.2. *Let $f(x)$ for an $\eta > 0$ be continuous and differentiable in $J_x: \langle x_0 - \eta, x_0 + \eta \rangle$. Assume for an $m > 0$ that we have $|f(x_0)| \leq \eta m$ and $|f'(x)| \geq m$ everywhere in J_x . Then $f(x) = 0$ has exactly one root in J_x .*

Proof. By Darboux’s theorem, $f'(x)$ has a constant sign throughout J_x , and therefore $f(x)$ is strictly monotonic in J_x and cannot have there more than one root. Without loss of generality, assume $f(x_0) > 0$. Let

$$\text{Min}f(x) = f(\xi) \quad (x \in J_x, \quad \xi \in J_x).$$

We have two cases to consider:

Case I. If $f(\xi) \leq 0$, then the theorem is proved, for if $f(\xi) = 0$, the assertion is evident; and if $f(\xi) < 0$, since $f(x_0) > 0$, there is a point between ξ and x_0 at which $f(x) = 0$.

Case II. $f(\xi) > 0$. By the mean value theorem of differential calculus,

$$f(x_0) - f(\xi) = (x_0 - \xi) f'(\rho), \quad \rho \in J_x,$$

$$f(x_0) > f(x_0) - f(\xi) = |x_0 - \xi| |f'(\rho)|,$$

$$f(x_0) > |x_0 - \xi| m.$$

But, since $f(x)$ is either monotonically increasing or monotonically decreasing in J_x , ξ is one of the end points of J_x . Hence, $f(x_0) > \eta m$, contrary to our hypothesis, and Case II is not possible, Q.E.D.

For example, suppose $|f'(x)| \geq 1/10^2$ for $|x - x_0| \leq 10^{-2}$ and $|f(x_0)| \leq 1/10^5$. Then we can take $\eta = 1/10^3$ and see that $f(x)$ has a zero in the interval $\langle x_0 - 1/10^3, x_0 + 1/10^3 \rangle$. We have therefore an approximation to a zero of $f(x)$ with an error not greater than 0.001.

10. An analogous theorem holds also in the case of an analytic function of a complex variable:

Theorem 2.3. *Let $f(z)$, for an $\eta > 0$, be analytic in the circle K_η ($|z - z_0| \leq \eta$). Assume for an $m > 0$ that we have $|f(z_0)| \leq \eta m$ and, everywhere in K_η , $|f'(z)| \geq m$. Then $f(z) = 0$ has a zero inside K_η .*

Considering instead of $f(z)$ the function

$$g(z) = \frac{1}{\eta} (f(z_0 + \eta z) - f(z_0)),$$

we see that Theorem 2.3 is a corollary of

Theorem 2.3°. *Let $w = g(z)$ be analytic both in the unit circle E , $|z| < 1$, and for $|z| = 1$. Assume that $g(0) = 0$ and that everywhere in E $|g'(z)| > m > 0$. Then $g(z)$ assumes in E every value a with $|a| \leq m$.*

11. Before giving the proof of this theorem we begin with some general considerations.

Suppose that $f(z)$ is regular analytic along a path γ , which is a simple curve in the z -plane. At a point z_0 of γ , if $f'(z_0) \neq 0$, the inverse $z = F(w)$ of $w = f(z)$ exists and is uniquely determined. We call $F(w)$ the local inverse of $f(z)$ in z_0 .

By virtue of $w = f(z)$ the path γ is transformed into a curve Γ in the w -plane and $w_0 = f(z_0)$ lies upon Γ . We will now show that if $f'(z) \neq 0$ along γ then $F(w)$ can be continued analytically along Γ and this analytical continuation coincides with the local inverse of $f(z)$ in the corresponding points of γ .

We have then, along γ ,

$$F(f(z)) = z, \quad (2.14)$$

and it could appear that our assertion is a special case of the so-called principle of permanence of functional equations. However, the following discussion cannot be avoided since, in order to apply this principle of permanence to (2.14), we have to assume that the function $F(w)$ exists as a regular function along Γ , while this fact is just the essential content of our assertion.

12. In order to prove our assertion we can assume without loss of generality that z_0 is the initial point of the path γ . Our assertion is certainly true

for a sufficiently small neighborhood of z_0 and the corresponding neighborhood of w_0 . If it is not true for the whole path γ then there exists a point z_1 on γ such that the assertion is true for the portion of γ between z_0 and z_1 , not necessarily including z_1 , while it is no longer true for a portion of γ beginning with z_0 and containing z_1 in its interior. Since $f'(z_1) \neq 0$, there exists a local inverse $F_1(w)$ of $f(z)$ in z_1 and the analytical continuation of $F_1(w)$ into a sufficiently small neighborhood of $w_1 = f(z_1)$ gives the local inverse of $f(z)$ each time, while on the other hand such a local inverse is also given by $F(w)$ in a portion of γ between z_0 and z_1 immediately adjoining z_1 . Therefore $F_1(w)$ is the immediate analytical continuation of $F(w)$ beyond z_1 , contrary to our assumption. We see that $F(w)$ can be continued up to the end point of Γ .

13. To prove Theorem 2.3° denote by $z = G(w)$ the local inverse of $w = g(z)$ at the origin. We will try to continue $G(w)$ analytically along the path $w = ta$ ($0 \leq t \leq 1$); $G(w)$ obviously can be continued along the segment $w = ta$ ($0 \leq t < t_0 \leq 1$) and remains in modulus < 1 for sufficiently small t_0 .

Let t_1 be the supremum of permissible t_0 . Then $G(w)$ can be continued analytically along the path $w = ta$ ($0 \leq t < t_1$), and we have

$$|G(ta)| < 1 \quad (0 \leq t < t_1)$$

and

$$|G'(ta)| = \frac{1}{|g'(G(ta))|} < \frac{1}{m} \quad (0 \leq t < t_1).$$

We have then

$$G(ta) = \int_0^{ta} G'(w) dw \quad (0 \leq t < t_1) \quad (2.15)$$

and from (2.15) follows the existence of the limit[†]

$$z^* = \lim_{t \uparrow t_1} G(ta) = \int_0^{t_1 a} G'(w) dw \quad (2.16)$$

and further

$$|z^*| < \left| \int_0^{t_1 a} \frac{dw}{m} \right| = |t_1| \frac{|a|}{m} \leq 1.$$

14. Then put $z^* = z_{t_1}$ and consider the path

$$(\gamma) \quad z = z_t := G(ta) \quad (0 \leq t < t_1).$$

[†] We shall denote x approaching ζ *decreasingly* through values larger than ζ (from the right) by $x \downarrow \zeta$. For x approaching ζ *increasingly* through values smaller than ζ (from the left) we write $x \uparrow \zeta$.

Since we have $ta = g(z_t)$ and z^* lies inside the unity circle it follows that

$$t_1 a = g(z^*) = g(z_{t_1})$$

and we obtain as the image of the path $\gamma \cup z_{t_1}$ the segment

$$(\Gamma) \quad w = ta \quad (0 \leq t \leq t_1).$$

Applying the discussion of Sections 11, 12 and replacing $f(z)$ by $g(z)$ and $F(w)$ by $G(w)$ we see that $G(w)$ remains analytic up to the point $w = t_1 a$, so that we can take as t_0 a certain number $> t_1$ unless $t_1 = 1$.

We see that we must have $t_1 = 1$; but now substituting $G(a) = z_0$ for $t = 1$, we obviously have $a = g(z_0)$ and our theorem is proved.

A DEVELOPMENT OF A ZERO OF $f(x)$

15. We can use the values of $\varphi^{(v)}(y)$ to obtain a development of a zero of $f(x)$. Assume that we have an x for which $f(x)$ is already “small” while $f'(x)$ is “not too small,” so that Theorem 2.2 is applicable. Writing then, for the zero ζ given by this theorem, $\zeta = x + h$, we will obtain a development of h in powers of

$$k = -\frac{f(x)}{f'(x)}. \quad (2.17)$$

Indeed, we have, using (2.17),

$$f(x) = -ky', \quad 0 = y - f(x) = y + y'k$$

and therefore

$$\zeta = x + h = \varphi(0) = \varphi(y + y'k),$$

and developing the right-hand expression in powers of k and assuming that $f^{(n+1)}(t)$ is continuous in the interval J_x of Theorem 2.2:

$$x + h = \sum_{v=0}^n \frac{\varphi^{(v)}(y)}{v!} y'^v k^v + \frac{\varphi^{(n+1)}(\sigma)}{(n+1)!} y'^{n+1} k^{n+1}, \quad (2.18)$$

where

$$\sigma = y + \theta' y' k = (1 - \theta') y = \theta y$$

with a θ from $(0, 1)$.

In order to use this formula for the *computation* of h with $n \rightarrow \infty$, we must assume that $f(x)$ is *analytic*. This will be discussed in Chapter 14. On the other hand, the above formula can be used for the discussion of *asymptotic properties* of different approximation procedures.

16. The first right-hand term in (2.18) is x . In the v th term, for $v > 0$, we replace $\varphi^{(v)}(y)$ by its expression from (2.5) and use the relation (2.9). Then this v th term becomes

$$\frac{1}{v!} X_v \left(\frac{y'}{y'}, \frac{y''}{y'}, \dots, \frac{y^{(v)}}{y'} \right) k^v.$$

As to the remainder term in (2.18), in our assumptions the expression $\varphi^{(n+1)}(\sigma) y'^{n+1}/(n+1)!$ remains bounded for $y \rightarrow 0$, that is, for $x \rightarrow \zeta, f(x) \rightarrow 0, k \rightarrow 0$. We obtain therefore the Schröder series[†]

[†] We say that $f = O(g)$ if $\overline{\lim}(f/g)$ is finite for a certain limiting process, while $f = o(g)$ signifies that $f/g \rightarrow 0$. Of course, in using such formulas the limiting process in question must be unambiguously specified. On the other hand, we also write $f = O(g)$ if $|f/g|$ remains bounded for the whole range of values of this expression.

$$\begin{aligned} h &= \zeta - x = \sum_{v=1}^n \frac{1}{v!} X_v \left(\frac{y'}{y'}, \frac{y''}{y'}, \dots, \frac{y^{(v)}}{y'} \right) k^v + O(f(x)^{n+1}), \\ k &= -\frac{f(x)}{f'(x)} \quad (x \rightarrow \zeta, \quad f(x) \rightarrow 0, \quad f'(\zeta) \neq 0). \end{aligned} \tag{2.19}$$

Introducing here the first values of the X_v from Section 8, we obtain

$$\begin{aligned} h &= \zeta - x = k - \frac{1}{2} \frac{y''}{y'} k^2 + \frac{3y'^2 - y'y^{(3)}}{6y'^2} k^3 \\ &\quad + \frac{10y'y''y^{(3)} - y'^2 y^{(4)} - 15y''^3}{24y'^3} k^4 + \dots. \end{aligned} \tag{2.20}$$

3

Method of False Position (Régula Falsi)

DEFINITION OF THE REGULA FALSI

1. Let $y = f(x)$ be defined in J_x . Assume that we are given two distinct interpolation points $x_1, x_2 \in J_x$, $x_1 \neq x_2$, with $f(x_1) = y_1, f(x_2) = y_2, y_1 \neq y_2, y_1 y_2 \neq 0$.

We approximate $f(x)$ by a linear function $L(x)$ which assumes the values y_1 and y_2 in x_1, x_2 :

$$f(x) \supseteq L(x) = \frac{(x-x_1)y_2 - (x-x_2)y_1}{x_2-x_1}. \quad (3.1)$$

To find an approximation to a root of $f(x) = 0$, we solve the linear equation $L(x) = 0$ with respect to x and get

$$x_3 = \frac{x_1 y_2 - x_2 y_1}{y_2 - y_1}. \quad (3.2)$$

Equation (3.2) can also be written in one of the forms

$$x_3 = x_1 - y_1 \frac{x_2 - x_1}{y_2 - y_1}, \quad (3.2a)$$

$$x_3 = x_2 - y_2 \frac{x_2 - x_1}{y_2 - y_1}. \quad (3.2b)$$

2. We ask whether x_3 approximates some solution of $f(x) = 0$. Substituting in our formula (1B.12) for the remainder with $n = 2$, we have for each $x \in J_x$

$$f(x) - L(x) = \frac{1}{2}f''(\xi)(x-x_1)(x-x_2), \quad \xi \in (x, x_1, x_2), \quad (3.3)$$

where by (x, x_1, x_2) we mean the *open* interval having two of these points as end points and the third point not outside. If $x = x_3$, since $L(x_3) = 0$, we get

$$f(x_3) = \frac{1}{2}f''(\xi)(x_3-x_1)(x_3-x_2), \quad \xi \in (x_1, x_2, x_3). \quad (3.4)$$

Substituting (3.2a) and (3.2b) in (3.4), we get

$$f(x_3) = \frac{1}{2} f''(\xi) y_1 y_2 \frac{(x_2 - x_1)^2}{(y_2 - y_1)^2}, \quad \xi \in (x_1, x_2, x_3). \quad (3.5)$$

Since

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'(\xi_0), \quad \xi_0 \in (x_1, x_2),$$

we may rewrite (3.5) as follows:

$$f(x_3) = y_1 y_2 \frac{f''(\xi)}{2f'(\xi_0)^2}, \quad \xi \in (x_1, x_2, x_3), \quad \xi_0 \in (x_1, x_2). \quad (3.6)$$

We see that $f(x_3)$ will be small if x_1 and x_2 are close enough to some zero of $f(x)$, since then y_1 and y_2 are small.

Formula (3.2) is called the *rule of false position*, or *regula falsi*.

USE OF INVERSE INTERPOLATION

3. We shall now obtain a direct estimate of the error of the approximation by x_3 to a zero of $f(x)$ from the theory of inverse interpolation.

Let $x = \varphi(y)$, the inverse function of $y = f(x)$, be defined in the y -interval corresponding to J_x . Assume that there exists a zero ζ of $f(x)$ in J_x and that $f'(x)$ does not vanish in this interval. Then we have $\varphi(0) = \zeta$. We now interpolate the function $\varphi(y)$ by a linear function:

$$\chi(y) = \frac{(y - y_1)x_2 - (y - y_2)x_1}{y_2 - y_1}. \quad (3.7)$$

Our problem is that of evaluating $\varphi(0)$. We approximate this value by

$$\chi(0) = \frac{x_1 y_2 - x_2 y_1}{y_2 - y_1} = x_3.$$

Notice that this value is the same as that obtained by direct interpolation. In either case the approximating curve is a straight line and we obtain the same result whether we interpolate $f(x)$ or $\varphi(y)$.

4. From (2.2) we get

$$\varphi(0) - \chi(0) = y_1 y_2 \frac{\varphi''(\eta)}{2}, \quad \eta \in (0, y_1, y_2).$$

Using (2.4) in this equation, we have, taking in J_x the number ξ corresponding to η ,

$$\begin{aligned}\varphi(0) - \chi(0) &= -y_1 y_2 \frac{f''(\xi)}{2f'(\xi)^3}, \quad \xi \in (\zeta, x_1, x_2), \\ \zeta - x_3 &= -y_1 y_2 \frac{f''(\xi)}{2f'(\xi)^3}, \quad \xi \in (\zeta, x_1, x_2).\end{aligned}\quad (3.8)$$

Now apply the mean value theorem of differential calculus; recalling that $f(\zeta) = 0$, we get

$$\begin{aligned}y_1 &= f(x_1) - f(\zeta) = (x_1 - \zeta) f'(\xi_1), \quad \xi_1 \in (x_1, \zeta), \\ y_2 &= (x_2 - \zeta) f'(\xi_2), \quad \xi_2 \in (x_2, \zeta)\end{aligned}$$

and

$$\begin{aligned}\zeta - x_3 &= \left[\frac{-f''(\xi) f'(\xi_1) f'(\xi_2)}{2f'(\xi)^3} \right] (\zeta - x_1)(\zeta - x_2), \\ \xi &\in (\zeta, x_1, x_2), \quad \xi_1 \in (x_1, \zeta), \quad \xi_2 \in (x_2, \zeta).\end{aligned}\quad (3.9)$$

5. We now discuss the magnitude of the first factor on the right in (3.9). Let $0 \leq m_2 \leq |f''(x)| \leq M_2$, $0 < m_1 \leq |f'(x)| \leq M_1$ throughout J_x . An upper bound of the modulus of the bracketed factor is $K = M_2 M_1^2 / 2m_1^3$; a lower bound is $k = m_2 m_1^2 / 2M_1^3$, and we obtain

$$k |\zeta - x_1| |\zeta - x_2| \leq |\zeta - x_3| \leq K |\zeta - x_1| |\zeta - x_2|. \quad (3.10)$$

If $x_1 \rightarrow \zeta$, $x_2 \rightarrow \zeta$, then, assuming f' and f'' continuous at ζ , we have from (3.9)

$$\frac{\zeta - x_3}{(\zeta - x_1)(\zeta - x_2)} \rightarrow -\frac{f''(\zeta)}{2f'(\zeta)}. \quad (3.11)$$

Since $f''(\zeta)$ and $f'(\zeta)$ are bounded and $f'(\zeta) \neq 0$, we see from (3.11) that a close enough approximation to ζ by x_1 and x_2 induces a considerably better approximation by x_3 .

It may be mentioned that a formula similar to (3.9) can also be obtained from (3.5) by replacing $f(x_3)$ there by $f'(\xi')(x_3 - \zeta)$, y_1 by $(x_1 - \zeta) f'(\xi_1)$, y_2 by $(x_2 - \zeta) f'(\xi_2)$, and $(y_2 - y_1)/(x_2 - x_1)$ by $f'(\xi_3)$. Then we get

$$\zeta - x_3 = \left[-\frac{f''(\xi) f'(\xi_1) f'(\xi_2)}{2f'(\xi') f'(\xi_3)^2} \right] (\zeta - x_1)(\zeta - x_2).$$

Here the bracketed coefficient is less easy to handle than the corresponding coefficient in (3.9), since in the case $\xi' = \xi_3 = \zeta$ the expression $f''(\xi)/f'(\xi)^3$ can be better estimated. On the other hand, the last formula also gives the inequality (3.10).

GEOMETRIC INTERPRETATION (FOURIER'S CONDITIONS)

6. We consider now the case where $y = f(x)$ has a graph as indicated in Fig. 1. Taking x_0 and x_1 as our initial approximations, we obtain x_2 . Taking now x_2 and x_0 as approximations, we obtain x_3 . Continuing this process, we obtain a sequence of points x_1, x_2, x_3, \dots , where generally

$$x_{v+1} = \frac{x_0 f(x_v) - x_v f(x_0)}{f(x_v) - f(x_0)} \quad (v = 1, 2, \dots). \quad (3.12)$$

Does x_v ($v = 1, 2, \dots$) converge?

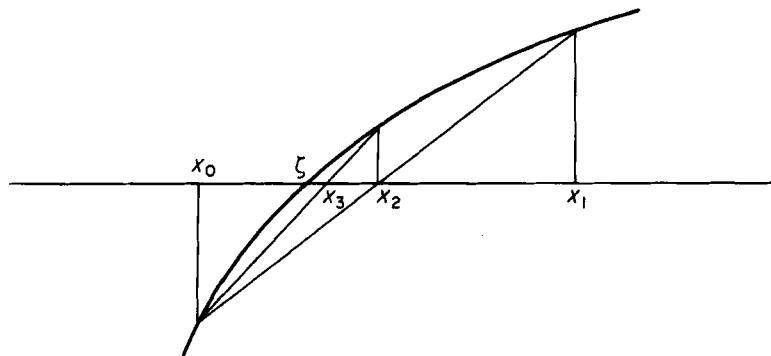


FIGURE 1

7. If the geometric situation is as we have drawn in Fig. 1, then x_v converges indeed. For x_1, x_2, x_3, \dots lie on the concave side of the arc and cannot go beyond ζ ; i.e., we have a monotonic decreasing sequence which is bounded from below by ζ and hence must converge to a limit ζ_0 .

Is ζ_0 the root ζ of $f(x) = 0$? Subtracting ζ_0 from both sides of (3.12) and taking the limit as $v \rightarrow \infty$, we obtain

$$0 = \frac{x_0 f(\zeta_0) - \zeta_0 f(x_0) - \zeta_0 [f(\zeta_0) - f(x_0)]}{f(\zeta_0) - f(x_0)} = f(\zeta_0) \frac{x_0 - \zeta_0}{f(\zeta_0) - f(x_0)}.$$

Now $x_0 \neq \zeta_0$; hence, $f(\zeta_0) = 0$ and ζ_0 is a zero of $f(x)$ in J_x and therefore equal to ζ .

To use the argument outlined above, the curve must have no points of inflection in the considered domain; for instance, it is sufficient to assume that $f''(x) \neq 0$ throughout J_x . As to x_0 , we must take it in such a way that $f(x_0)f''(x_0) > 0$.

If the conditions listed above (the so-called *Fourier conditions*) are not

satisfied, e.g., if x_0 is a point for which $\operatorname{sgn} f(x_0) \neq \operatorname{sgn} f''(x_0)$,[†] we may get a sequence which still converges to ζ but oscillates about ζ . Such a case will be considered in Chapter 5. See also Example 1 in Section 17.

ITERATION WITH SUCCESSIVE ADJACENT POINTS

8. The convergence of the method of false position is considerably improved in the long run if, instead of constantly using x_0 as one of the interpolation points, we use at each step the two last points of the sequence:

$$x_{v+1} = \frac{x_{v-1}f(x_v) - x_v f(x_{v-1})}{f(x_v) - f(x_{v-1})} \quad (v = 1, 2, \dots). \quad (3.13)$$

Applying (3.9) to (3.12), we replace x_1, x_2, x_3 by x_0, x_v, x_{v+1} and obtain

$$\zeta - x_{v+1} = \left[-\frac{f''(\xi)}{2f'(\xi)^3} f'(\xi_1) f'(\xi_2) (\zeta - x_0) \right] (\zeta - x_v), \quad (3.14)$$

$$\xi \in (\zeta, x_0, x_v), \quad \xi_1 \in (x_0, \zeta), \quad \xi_2 \in (x_v, \zeta).$$

If the modulus of the bracketed expression is always $\leq \rho$, $\rho < 1$, and does not tend to zero, we have what is called *linear convergence*. From (3.14) it follows then in the case of formula (3.12)

$$|\zeta - x_v| \leq \rho^{v-1} |\zeta - x_1|, \quad \rho < 1. \quad (3.15)$$

Interpreting this formula in terms of the number of additional true digits which we obtain at each step, we see that if $\rho = 1/10$, we get one digit at each step; if $\rho = 1/3$, we get approximately one digit every second step, etc. On the other hand, the modulus of the bracketed expression in (3.14) becomes certainly considerably < 1 as soon as x_0 is chosen sufficiently near to ζ .

9. If instead of using (3.12) we proceed by (3.13) and use x_3 and x_2 to get x_4 , then x_4 and x_3 to get x_5 , and so on, we get what is called *superlinear convergence*, provided x_0 and x_1 are sufficiently close to ζ . Indeed, multiply (3.10) by K and replace x_1, x_2, x_3 by x_0, x_1, x_2 . Then

$$K|\zeta - x_2| \leq K^2 |\zeta - x_1| |\zeta - x_0|.$$

Put $K|\zeta - x_v| := d_v$; then we have, applying this repeatedly, $d_2 \leq d_0 d_1, \dots, d_{v+1} \leq d_v d_{v-1}$. Assume now that d_0 and d_1 are both $\leq d < 1$. Then

$$d_2 \leq d^2, \quad d_3 \leq d^3, \quad d_4 \leq d^5, \quad d_5 \leq d^8, \dots$$

or generally $d_v \leq d^{\alpha_v}$, where $\alpha_0 = 1, \alpha_1 = 1, \dots, \alpha_{v+1} = \alpha_v + \alpha_{v-1}$. We have here

[†] The symbol $\operatorname{sgn} \alpha$ has the meaning $\alpha/|\alpha|$, if $\alpha \neq 0$, while $\operatorname{sgn} 0 = 0$.

the *Fibonacci* sequence defined by

$$\alpha_{v+1} = \alpha_v + \alpha_{v-1} \quad (v = 1, 2, 3, \dots), \quad \alpha_1 = \alpha_0 = 1. \quad (3.16)$$

10. Equation (3.16) is an example of a homogeneous linear difference equation constant coefficients. We will find a general expression for α_v and then obtain the particular solution of (3.16) corresponding to $\alpha_1 = \alpha_0 = 1$.

We try a solution of the form t^v . Substituting in (3.16), we have $t^{v+1} = t^v + t^{v-1}$. Dividing by t^{v-1} , where $t \neq 0$, we get $t^2 = t + 1$. The roots of this equation are

$$t_1 = \frac{1 + \sqrt{5}}{2}, \quad t_2 = \frac{1 - \sqrt{5}}{2}. \quad (3.17)$$

Consider $\beta_v = c_1 t_1^v + c_2 t_2^v$, where c_1 and c_2 are arbitrary constants. Now β_v satisfies the equation in (3.16) for

$$c_1 t_1^{v+1} + c_2 t_2^{v+1} = c_1 t_1^v + c_2 t_2^v + c_1 t_1^{v-1} + c_2 t_2^{v-1}.$$

Choose c_1 and c_2 such that $\beta_1 = \beta_0 = 1$; i.e., $c_1 t_1 + c_2 t_2 = 1$, $c_1 + c_2 = 1$. We obtain

$$c_1 = \frac{5 + \sqrt{5}}{10} = \frac{t_1}{\sqrt{5}}, \quad c_2 = \frac{5 - \sqrt{5}}{10} = -\frac{t_2}{\sqrt{5}}.$$

Clearly then $\alpha_v = \beta_v$ ($v = 0, 1, 2, \dots$), and we have thus obtained a general expression for

$$\alpha_v = c_1 t_1^v + c_2 t_2^v = \frac{1}{\sqrt{5}}(t_1^{v+1} - t_2^{v+1}), \quad (3.18)$$

where t_1, t_2 are given by (3.17). For $v \rightarrow \infty$, $(1/\sqrt{5})t_1^{v+1}$ will be the dominating term in (3.18) and we have $\alpha_v \sim (1/\sqrt{5})t_1^{v+1}$ or

$$\alpha_v \sim 0.7236 \cdot (1.618)^{v+1}.$$

As a matter of fact, it can be proven that if in the approximation of Section 9 we have $\zeta - x_v \rightarrow 0$ and $f^{(3)}(x)$ exists and remains bounded, then we always have

$$\frac{|\zeta - x_{v+1}|}{|\zeta - x_v|^{t_1}} = \left| \frac{2f'(\zeta)}{f''(\zeta)} \right|^{t_2} + O(t_2^v) \quad (v \rightarrow \infty). \quad (3.19)$$

The proof is given in Chapter 12, Sections 11 and 12.

HORNER UNITS AND EFFICIENCY INDEX

11. The computational work involved in computing a function or one of its derivatives will be called in this book a *horner* (*Horner unit*). We can thus

say that at the expense of one horner the number of true digits obtained from our computation is multiplied in the average by $1.618 \dots$.[†]

If we obtain by a procedure a sequence x_v convergent to ζ and have to spend m_v horners for the passage from x_v to x_{v+1} , we call the limit

$$\lim^{m_v} \sqrt{\frac{\ln|x_{v+1}-\zeta|}{\ln|x_v-\zeta|}},$$

if this limit exists, the *efficiency index* of the procedure. We see then that the efficiency index of the *regula falsi* used in the sense of formula (3.13) is $1.618 \dots$.[‡]

THE ROUNDING-OFF RULE

12. Since in using (3.2) the error of x_3 is $O(y_1 y_2)$, it may appear plausible that the values of y_1 and y_2 both have to be computed with the same degree of precision and that, in particular, if y_1 is computed first, its computation must be resumed after the order of magnitude of y_2 is known, and carried to the required degree of precision. As a matter of fact, however, the precision of the order $O(y_1 y_2)$ is not necessarily sufficient in computing y_1 and y_2 , and on the other hand, the precision necessary and sufficient in computing y_i is just $O((x_1 - x_2)y_i)$ ($i = 1, 2$), so that neither value has to be computed anew.

13. Indeed, it follows at once from (3.2) that for $i = 1, 2$

$$\frac{\partial x_3}{\partial y_i} = \pm \frac{y_1 y_2}{y_i} \frac{x_1 - x_2}{(y_1 - y_2)^2}.$$

Assume now that y_1 and y_2 are the *exact values* of $f(x_1), f(x_2)$, and let $y_1 + \delta_1$, $y_2 + \delta_2$ be the approximate values which have to be introduced into (3.2). Then

[†] Of course the use of a Horner unit is very much a matter of convention. Its significance varies from one problem to another and also according to the number of decimal places with which the computation is performed.

A similar idea is used in the theory of partial differential equations of the first order in order to compare the different methods of solution of such equations. The ‘index of simplicity’ is usually the sum of orders of all ordinary differential systems used in the method, although, for instance, obviously special ordinary differential equations of the first order can be much more difficult than a certain differential system of the third order.

[‡] The reason why the method of the false position is applied in the standard texts according to formula (3.12) is apparently a certain prejudice against *extrapolation* versus *interpolation*, a prejudice which is certainly justified in dealing with empirical data but is very much less so if we are working on an analytic expression.

we obtain for δ_1 and δ_2 the conditions

$$\delta_i \frac{\partial x_3}{\partial y_i} = \pm \frac{y_1 y_2}{y_i} \frac{x_1 - x_2}{(y_1 - y_2)^2} \delta_i = O(y_1 y_2), \quad \delta_i = O\left(y_i \frac{(y_1 - y_2)^2}{x_1 - x_2}\right).$$

On the other hand, as $x_1 \rightarrow \zeta$, $x_2 \rightarrow \zeta$, we have $y_1 - y_2 \sim f'(\zeta)(x_1 - x_2)$ and the above conditions become

$$\delta_i = O((x_1 - x_2)y_i). \quad (3.20)$$

The condition (3.20) can also be written in the form

$$\delta_i = O(y_i^2 - y_1 y_2). \quad (3.21)$$

From (3.21) we easily see that, if for a fixed positive ε , $|y_2| \leq (1-\varepsilon)|y_1|$, then the conditions (3.21) are equivalent to

$$\delta_1 = O(y_1^2), \quad \delta_2 = O(y_1 y_2) \quad (|y_2| \leq (1-\varepsilon)|y_1|). \quad (3.22)$$

It may finally be mentioned that if $y_1 y_2 < 0$, then the conditions (3.22) are sufficient without any further assumption about y_2 .

14. In the above discussion, we stated only the relationship between the “input” into (3.2) and the “output.” The problem, which of the theoretically equivalent formulas (3.2), (3.2a), (3.2b) has to be used, is quite a different one.

If we use (3.2) directly, then the numerator must be computed with the precision $O[y_1 y_2(y_1 - y_2)] = O[y_1 y_2(x_1 - x_2)]$. The situation is considerably better if we use one of the formulas (3.2a) and (3.2b), and in particular (3.2b), if $|y_2| \leq |y_1|$. Indeed, in this case the product $y_2(x_2 - x_1)/(y_2 - y_1)$ must be computed with a precision $O(y_1 y_2)$, and therefore the quotient $(x_2 - x_1)/(y_2 - y_1)$, which tends to $1/f'(\zeta)$, need only be computed with the precision $O(y_1)$.

LOCATING THE ZERO WITH THE REGULA FALSI

15. In our preceding discussion, we assumed that a zero exists in the considered interval and also that x_1 and x_0 are near enough to the zero so that d_1 and d_0 are less than *one*. Often we do not know whether a zero exists in the interval and we just hope that in the course of the computation we get close to such a point.

We have shown in Theorem 2.2 that y is a root of $f(x) = 0$ if $|f(y)|$ is sufficiently small; i.e., we can locate a root in the interval $(y \pm \eta)$ where $|f(y)| \leq \eta m_1$, m_1 being the minimum of $|f'(x)|$ in this interval.

Now let x_1 and x_2 be the starting values used to obtain the value x_3 by the rule of false position (3.2). Then from (3.4) we have

$$f(x_3) = \frac{1}{2}f''(\xi)(x_3 - x_1)(x_3 - x_2), \quad \xi \in (x_1, x_2, x_3),$$

and can try to apply Theorem 2.2 to $x = x_3$. We may avoid computing $f(x_3)$ if we are interested only in the *existence* of a root of $f(x) = 0$. We have then only to check whether

$$\frac{1}{2}M_2|x_3 - x_1||x_3 - x_2| \leq \eta m_1, \quad (3.23)$$

where

$$M_2 = \sup_{x \in (x_1, x_2, x_3)} |f''(x)|, \quad m_1 = \inf |f'(x)| (x_3 - \eta \leq x \leq x_3 + \eta).$$

16. We illustrate this on a classical equation originally considered by Newton:

$$f(x) \equiv x^3 - 2x - 5 = 0, \quad f'(x) = 3x^2 - 2, \quad f''(x) = 6x,$$

$$x_1 = 1, \quad y_1 = -6, \quad x_2 = 2, \quad y_2 = -1, \quad x_3 = 2.2.$$

$f'(x)$ and $f''(x)$ are both monotonic increasing functions in the interval (x_1, x_2, x_3) . Hence, $f'(x)$ assumes a minimum at the left end point of any subinterval and $f''(x)$ assumes a maximum at the right end point of (x_1, x_2, x_3) . Substituting in (3.23), we obtain

$$\frac{1}{2} \cdot 13.2 \cdot 0.24 = 1.584 < \eta m_1.$$

If we try $\eta = 0.15$, we have $m_1 = f'(2.05) = 10.6$ and the inequality $1.584 < 0.15 \cdot 10.6 = 1.59$ is satisfied. Our zero lies in the interval $(2.05, 2.35)$.

We now choose instead

$$x_1 = 1.8, \quad y_1 = -2.768, \quad x_2 = 2, \quad y_2 = -1, \quad x_3 = 2.113.$$

Substituting in (3.23) as before, we have

$$\frac{1}{2} \cdot 12.678 \cdot 0.113 \cdot 0.313 = 0.2242 \leq \eta m_1.$$

If $\eta = 0.03$, then $m_1 = f'(2.083) = 11.016$; $0.2242 \dots \leq (0.03)(11.01)$, and we can say that x_3 is an approximation to a zero of $f(x) = 0$ with an error less than 0.03; i.e., there is a root in the interval (2.113 ± 0.03) .

EXAMPLES OF COMPUTATION BY THE REGULA FALSI

17. We obtain the following forms of (3.12) and (3.13) by subtracting x_v from both sides:

$$x_{v+1} = x_v - f(x_v) \frac{x_v - x_0}{f(x_v) - f(x_0)}, \quad (3.24a)$$

$$x_{v+1} = x_v - f(x_v) \frac{x_v - x_{v-1}}{f(x_v) - f(x_{v-1})}. \quad (3.24b)$$

In using these formulas, if we want to compute k digits of x_{v+1} after the point and if $f(x_v)$ begins with k_1 zeros after the point, it is sufficient to get only $k - k_1$ digits after the decimal point in the cofactor of $f(x_v)$.

Example 1. The following is an example of the computation by formula (3.24a) with linear convergence:

$$f(x) \equiv x^3 - 2x - 5 = 0.$$

The correct value of the root to 13 decimal places is

$$\zeta = 2.0945\ 51481\ 5423.$$

EXAMPLE 1

(1) v	(2) x_v	(3) $\zeta - x_v$
0	2.	
1	3.	-0.9054
2	2.0588 23529 4	0.0357
3	2.0965 58636 2	-0.0 ₂ 201 ^a
4	2.0944 40519 3	0.0 ₃ 111
5	2.0945 57621 8	-0.0 ₅ 614
6	2.0945 51139 9	0.0 ₆ 342
7	2.0945 51500 6	-0.0 ₇ 191

^a The subscript denotes the number of zeros following the decimal point.

Example 2. The following is an example of the computation by formula (3.24b) with supralinear convergence: $f(x) \equiv x^3 - 2x - 5 = 0$. The root correct to 24 decimal places is

$$\zeta = 2.0945\ 51481\ 54232\ 65914\ 82387.$$

We have

$$f'(\zeta) = 11.1614377, \quad f''(\zeta) = 12.56730888, \quad \frac{f''(\zeta)}{2f'(\zeta)} = 0.5629789. \quad (3.25)$$

We give a further discussion of the *regula falsi* from another point of view in Chapter 5.

For generalization of the *regula falsi* to systems of equations, see Appendix D.

EXAMPLE 2

(1) v	(2) x_v	(3) $\zeta - x_v$	(4) $\frac{x_{v+1} - \zeta}{(x_v - \zeta)(x_{v-1} - \zeta)}$
0	2.		
1	3.	-0.9054	
2	2.0588 23529 4	0.0357	0.411
3	2.0812 63659 65	0.0133	0.574
4	2.0948 24184 27	0.0 ₃ 2727	0.565
5	2.0945 49431 75	0.0 ₅ 205	0.5624
6	2.0945 51481 228	0.0 ₉ 314	0.5629 779
7	2.0945 51481 54232 69542 1	0.0 ₁₅ 3627	^a

^a All calculations for x_7 were made with double precision, i.e., with 20 decimal places.

4

Iteration

A CONVERGENCE CRITERION FOR AN ITERATION

1. Let $\psi(x)$ be defined in J_x and let $x_1 \in J_x$. Form

$$x_2 = \psi(x_1), \quad x_3 = \psi(x_2), \dots, x_{v+1} = \psi(x_v), \dots, \quad (4.1)$$

and assume that the points so obtained lie in J_x . If, in particular, $\psi(x_1) = x_1$, the whole sequence x_v ($v = 1, 2, \dots$) consists of the repetition of x_1 .

A number ζ such that $\psi(\zeta) = \zeta$ is called a *fixed point of the iteration* or a *center of the iteration* (4.1). The function $\psi(x)$ is, in this connection, called an *iterating function*.

Theorem 4.1. *Let $F(x)$ be defined in J_x and*

$$F(x) \equiv x - \psi(x), \quad (4.2)$$

where $\psi(x)$ is used as an iterating function to form the sequence (4.1). Assume that in (4.1) $x_v \rightarrow \zeta$, where all x_v and ζ lie in J_x , and that $\psi(x)$ is continuous in ζ . Then $F(\zeta) = 0$.

Proof. From the continuity of $\psi(x)$ in ζ , from $x_{v+1} = \psi(x_v)$, and from $x_v \rightarrow \zeta$, it follows that $\zeta = \psi(\zeta)$. Substituting in (4.2), we get

$$\zeta - \zeta = F(\zeta) = 0, \quad \text{Q.E.D.}$$

POINTS OF ATTRACTION AND REPULSION

2. We shall use $V(\zeta_0)$ to denote a symmetric neighborhood of ζ_0 , for example, $\zeta_0 - \eta < x < \zeta_0 + \eta$, $\eta > 0$. Suppose that $\zeta_0 = \psi(\zeta_0)$. ζ_0 is called a *point of attraction* if in (4.1) for every starting point x_1 within a sufficiently close neighborhood of ζ_0 we have $x_v \rightarrow \zeta_0$. ζ_0 is called a *point of definite repulsion* if in (4.1) for all points x_1 within a sufficiently close neighborhood of ζ_0 we have $x_v \nrightarrow \zeta_0$ (unless one of the x_v becomes equal to ζ_0).[†]

[†] The more special termini of a *point of attraction* and a *point of definite repulsion* for a *one-sided* approximation are defined in a corresponding way, replacing the "sufficiently close neighborhood" of ζ_0 by a sufficiently close *one-sided* neighborhood.

Theorem 4.2. Let $\psi(x)$ be an iterating function defined in J_x . Assume that $\zeta_0 = \psi(\zeta_0)$, $\zeta_0 \in (J_x)$, and that $\psi'(\zeta_0)$ exists. Then ζ_0 is a point of attraction or a point of definite repulsion depending on whether we have $|\psi'(\zeta_0)| < 1$ or $|\psi'(\zeta_0)| > 1$. In the first case we have

$$\frac{x_{v+1} - \zeta_0}{x_v - \zeta_0} \rightarrow \psi'(\zeta_0). \quad (4.3)$$

3. Proof. Part I. Suppose $|\psi'(\zeta_0)| < 1$. Then, if we choose a p with $|\psi'(\zeta_0)| < p < 1$ we have for any x within a convenient $V(\zeta_0)$

$$\left| \frac{\psi(x) - \psi(\zeta_0)}{x - \zeta_0} \right| \leq p. \quad (4.4)$$

Let an $x_1 \in V(\zeta_0)$ be the starting point. Then

$$\left| \frac{x_2 - \zeta_0}{x_1 - \zeta_0} \right| \leq p < 1.$$

Hence, the distance from x_2 to ζ_0 is less than the distance of x_1 from ζ_0 , and we have $x_2 \in V(\zeta_0)$. Applying repeatedly (4.4) to x_2, \dots, x_v , we get $|x_3 - \zeta_0| \leq p|x_2 - \zeta_0|, \dots,$

$$|x_{v+1} - \zeta_0| \leq p^v |x_1 - \zeta_0| \rightarrow 0 \quad (v \rightarrow \infty). \quad (4.5)$$

Hence ζ_0 is a point of attraction. From $x_v \rightarrow \zeta_0$ follows

$$\frac{x_{v+1} - \zeta_0}{x_v - \zeta_0} = \frac{\psi(x_v) - \psi(\zeta_0)}{x_v - \zeta_0} \rightarrow \psi'(\zeta_0).$$

4. Proof. Part II. Suppose that $|\psi'(\zeta_0)| > 1$. Then, if we choose a p with $|\psi'(\zeta_0)| > p > 1$, we have for any x from a convenient $V(\zeta_0)$

$$\left| \frac{\psi(x) - \psi(\zeta_0)}{x - \zeta_0} \right| \geq p. \quad (4.5a)$$

Hence for an $x_1 \in V(\zeta_0)$, we get $|x_2 - \zeta_0| \geq p|x_1 - \zeta_0|$; i.e., x_2 is farther away from ζ_0 than x_1 . The same argument holds for any x_v as long as it lies in $V(\zeta_0)$ and is $\neq \zeta_0$. Hence $x_v \rightarrow \zeta_0$, unless one of the x_v is $= \zeta_0$, and ζ_0 is a point of definite repulsion, Q.E.D.

5. Remarks. (a) If we get outside the neighborhood $V(\zeta_0)$, the inequality (4.5a) may not hold, and it is entirely possible that our next point may be ζ_0 . It is possible to construct a nonanalytic function $\psi(x)$ such that outside a certain neighborhood of ζ_0 the function is everywhere equal to ζ_0 . Even a nonconstant analytic function $\psi(x)$ may be constructed which is equal to ζ_0 at an enumerable number of points outside a neighborhood of ζ_0 .

(b) In the case $|\psi'(\zeta_0)| < 1$, the convergence will be the better the smaller $|\psi'(\zeta_0)|$ is, provided we start in a sufficiently small neighborhood of ζ_0 . More generally, the rapidity of convergence depends on the smallness of the expression

$$\left| \frac{\psi(x) - \psi(\zeta_0)}{x - \zeta_0} \right|.$$

(c) Since the original problem is to find ζ_0 , we are still faced with the problem of determining whether our desired solution is a point of attraction or repulsion. If throughout the considered interval $|\psi'(x)| \leq p < 1$ and if there is a root in the interval, then this root will be a point of attraction and may be obtained by the method of iteration. We would then have (4.4). If p is sufficiently small, we have fast convergence. In practice we certainly say the convergence is "good" if p is less than $\frac{1}{10}$. If p is greater than $\frac{1}{2}$, the convergence is certainly "slow." In any case, we shall try to work out a scheme to speed up the convergence considerably.

IMPROVING THE CONVERGENCE

6. Suppose $|\psi'(x)| < 1$ in the considered interval. What does this condition signify in terms of $F(x) = x - \psi(x)$? We have $\psi'(x) = 1 - F'(x)$. Now $|1 - F'(x)| < 1$ implies that $0 < F'(x) < 2$. Hence, if we use the above criterion for convergence, our method of iteration applies primarily only to functions $F(x)$ which are monotonically *increasing*, but not too fast. [If $F(x)$ is monotonically *decreasing*, we replace $F(x)$ by $-F(x)$.]

Assume that $F'(x)$ is *positive* in the considered interval, and $M \geq F'(x) \geq m > 0$. The roots of $F(x) = 0$ do not change if we multiply $F(x) = 0$ by a constant $c \neq 0$ and $F'(x)$ remains positive if $c > 0$. We can choose c so that the convergence is ensured even for $M > 2$, and in general improved. For $F_c := cF(x)$, we have for the corresponding iterating function $\psi_c(x) = x - cF(x)$, and $\psi_c'(x)$ lies between $1 - cM$ and $1 - cm$. If we start with $c = 0$ and successively and continuously increase c , we find that $|1 - cM|$ and $|1 - cm|$ diminish as long as $1 - cM$ remains positive. First $1 - cM$ becomes zero and then $|1 - cM|$ begins to increase. In drawing a graph, we see that $\text{Max}(|1 - cM|, |1 - cm|)$ is the least if we choose c so that

$$1 - cm = -(1 - cM),$$

$$c = \frac{2}{m+M}, \quad (4.6)$$

$$|\psi_c'(x)| \leq \frac{M-m}{M+m}. \quad (4.7)$$

If $F'(x)$ does not change very fast [i.e., $F''(x)$ is small], this bound will be rather small. In any case, if $F''(x)$ is bounded in a neighborhood of ζ , we can make the bound in (4.7) as small and the convergence as fast as we want, by appropriately narrowing the interval around the zero.

In some cases the value of $\psi'(\zeta_0) = \alpha$ is known from theoretical discussions. If then $\alpha \neq 0$, $\alpha \neq 1$, we obtain at once a considerable improvement of convergence, replacing $\psi(x)$ by

$$\psi^*(x) = \frac{1}{1-\alpha}(\psi(x) - \alpha x). \quad (4.8)$$

Indeed, we verify at once that ζ_0 is a fixed point for $\psi^*(x)$ too, as

$$\frac{1}{1-\alpha}(\zeta_0 - \alpha \zeta_0) = \zeta_0,$$

and the derivative $\psi''(\zeta)$ is

$$\frac{1}{1-\alpha}(\alpha - \alpha) = 0.$$

If α is not known we can still try to use the expression (4.8), replacing α by a suitable approximation. Indeed we have, if $x_0 \rightarrow \zeta_0$, $x_1 \rightarrow \zeta_0$, $x_2 \rightarrow \zeta_0$, and $\psi'(x)$ is continuous in ζ_0 ,

$$\frac{x_2 - x_1}{x_1 - x_0} = \frac{\psi(x_1) - \psi(x_0)}{x_1 - x_0} \rightarrow \alpha.$$

In (4.8) therefore we replace α by $(x_2 - x_1)/(x_1 - x_0)$ and obtain

$$\frac{x_1 - x_0}{x_2 - 2x_1 + x_0} \left(\frac{x_2 - x_1}{x_1 - x_0} x - \psi(x) \right).$$

Putting here $x = x_0$ we get a new approximation for ζ_0 :

$$x^* = \frac{x_0 x_2 - x_1^2}{x_2 - 2x_1 + x_0} = \frac{x_0 \psi(\psi(x_0)) - \psi(x_0)^2}{\psi(\psi(x_0)) - 2\psi(x_0) + x_0} =: \Psi(x_0).$$

In this way we obtain with x^* a new approximation for ζ_0 which is in most cases considerably better than x_2 . The iterating function $\Psi(x)$ has been discovered by Steffensen, who obtained it in a different way (cf. Appendix E).

7. Take, as an example for the procedure of Section 6, the equation

$$x + \log x = 0.5$$

where $\log x$ is the *common logarithm* of x ; this equation is dealt with by Whittaker and Robinson, in *The Calculus of Observations*, as Example 2 in

Section 43, by means of the recurrence formula

$$x_{v+1} = 0.5 - \log x_v$$

and starting from the value $x_0 = 0.68$ obtained directly from the logarithm table. By iterating 7 times and by repeatedly using the arithmetical mean $\frac{1}{2}(x_v + x_{v+1})$ to speed up the convergence, they get the approximate value $x = 0.672382$, which is correct to five decimal places.

Now in the interval $\langle 0.67, 0.68 \rangle$ the function $F(x) = x + \log x - 0.5$ has the first derivative $F'(x) = 1 + \mu/x$ ($\mu = \log e = 0.43429448$), which lies between $M = 1.64821$ and $m = 1.63866$. We obtain from (4.6) $c = 0.6085$ and take $F_c(x) := cF(x)$,

$$\psi_c(x) = x - F_c(x) = 0.3915x - 0.6085 \log x + 0.30425.$$

Then we get the iteration formula $x_{v+1} = \psi_c(x_v)$ where $|\psi_c'| < 0.0029$ in the interval $\langle 0.67, 0.68 \rangle$. We have here

$$x_0 = 0.68, \quad x_1 = 0.67239, \quad x_2 = 0.672383185.$$

To check x_2 , we compute $F_c(x_2) = 21.2(\pm 6) \cdot 10^{-9}$ [†] and from Theorem 2.2 we get now, since $|F_c'(x) - 1| < 0.003$, the expression for the exact solution ζ : $\zeta = x_2 \pm 22 \cdot 10^{-9}$. We obtain here in three steps an approximation with an error $< 3 \cdot 10^{-8}$.

8. We come to the value (4.6) of c by the following plausibility argument. We want to choose c in such a way as to make $|\psi_c'| = |1 - cF'|$ as small as possible in the root ζ . But then the most appropriate value for $1/c$ would be the *arithmetical mean* between the maximum and the minimum of F' . In practice we merely divide F by a certain value $F'(\xi)$, choosing ξ appropriately, and we have then in any case convergence as long as $F'(\xi) > \frac{1}{2} \operatorname{Max} F'(x)$. Of course, if we already know that x_v is much nearer to ζ than x_{v-1} , it is best to choose at the v th step $c_v = 1/F'(x_v)$, that is to say, to compute x_{v+1} by the formula

$$x_{v+1} = x_v - \frac{F(x_v)}{F'(x_v)};$$

we come in this way to the Newton–Raphson formula.

In our above discussion of the iteration method, the assumption that $F(x)$ is monotonically increasing (or monotonically decreasing) is quite essential. In the following chapter, we discuss a method which works even if $F'(x)$ changes its sign in the considered interval, as long as $|F'(x)|$ remains bounded.

[†] The meaning of this notation is that the number in parentheses indicates the multiple of the last decimal unit given. Thus $21.2(\pm 6) \cdot 10^{-9}$ designates a number contained between $21.8 \cdot 10^{-9}$ and $20.6 \cdot 10^{-9}$.

9. The condition $|\psi(\zeta_0)| < 1$ of Theorem 4.2 is certainly satisfied if we have $\psi'(\zeta_0) = 0$, which is the same as

$$\frac{\psi(x) - \zeta_0}{x - \zeta_0} \rightarrow 0 \quad (x \rightarrow \zeta_0). \quad (4.9)$$

Here ζ_0 is certainly a point of attraction, but usually more can be said about the convergence than in the case of formula (4.5). Indeed, replacing x in (4.9) by x_v , we get

$$\frac{x_{v+1} - \zeta_0}{x_v - \zeta_0} \rightarrow 0 \quad (v \rightarrow \infty). \quad (4.10)$$

Generally, if we have (4.9), there will exist a number $k > 1$ such that

$$\limsup_{x \rightarrow \zeta_0} \frac{|\psi(x) - \zeta_0|}{|x - \zeta_0|^k} < \infty, \quad k > 1. \quad (4.11)$$

Then we have also, replacing x here by x_v ,

$$\limsup_{v \rightarrow \infty} \frac{|x_{v+1} - \zeta_0|}{|x_v - \zeta_0|^k} < \infty, \quad k > 1. \quad (4.12)$$

10. If we have a sequence x_v going to ζ_0 and if we have a relation of the type (4.12), then we say that the convergence of the x_v to ζ_0 is *supralinear*, and in particular that its *degree of convergence* is at least k , while the upper bound of all numbers k for which (4.12) holds is then the *exact degree of convergence* of the sequence x_v .

If on the other hand we have the relation $\limsup_{v \rightarrow \infty} |x_{v+1} - \zeta_0| / |x_v - \zeta_0| < 1$, we speak of at least *linear convergence*. These concepts are due to Schröder (1869).

A sufficient criterion for the degree of convergence being k is given by the following:

If we have

$$\psi'(\zeta_0) = \dots = \psi^{(k-1)}(\zeta_0) = 0, \quad \psi^{(k)}(\zeta_0) \neq 0,$$

and if $\psi(\zeta_0) = \zeta_0$, then, as $x \rightarrow \zeta_0$,

$$\frac{\psi(x) - \zeta_0}{(x - \zeta_0)^k} \rightarrow \frac{\psi^{(k)}(\zeta_0)}{k!} \quad (x \rightarrow \zeta_0).$$

11. If the condition $|\psi'(\zeta_0)| < 1$ of Theorem 4.2 is satisfied, there exists a neighborhood $U(\zeta_0)$ such that we get convergence of x_v to ζ_0 , starting with every point of this neighborhood. This is a result of typically *local character*, since, from our data, nothing more about this neighborhood $U(\zeta_0)$ can be asserted than just its existence. The following theorem contains an assertion of *global character*.

Theorem 4.3. Let $\psi(x)$ be an iterating function defined in J_x ($\zeta_0 - \eta < x < \zeta_0 + \eta$) with a fixed point ζ_0 . Assume further that $\psi'(x)$ exists everywhere in J_x and that we have, for a fixed m , $0 < m < 1$,

$$|\psi'(x)| \leq 1 - m < 1 \quad (x \in J_x). \quad (4.13)$$

Then, for every starting point from J_x we obtain a sequence x_v converging to ζ_0 and the convergence is at least linear.

12. Proof. We have, if an x_v lies in J_x , by the mean value theorem,

$$\frac{x_{v+1} - \zeta_0}{x_v - \zeta_0} = \frac{\psi(x_v) - \psi(\zeta_0)}{x_v - \zeta_0} = \psi'(\xi), \quad \xi \in (x_v, \zeta_0),$$

and therefore, by (4.13),

$$|x_{v+1} - \zeta_0| \leq (1-m)|x_v - \zeta_0|.$$

Since therefore x_{v+1} lies nearer to ζ_0 than x_v , we see that for every x_0 from J_x , the whole sequence x_v lies in J_x . On the other hand, we have from the above inequality

$$|x_v - \zeta_0| \leq (1-m)^v |x_0 - \zeta_0| \quad (v = 0, 1, \dots),$$

and we see that indeed the x_v tend to ζ_0 at least linearly. This proves our theorem.

13. Remarks. (1) In the above wording of Theorem 4.3, the interval J_x is assumed as *open*. It is easy to see that the theorem remains true if one or both end points of J_x are added to J_x , provided $\psi(x)$ is assumed continuous in the whole interval J_x , while, as to the conditions concerning $\psi'(x)$, it is sufficient if they hold *inside* J_x .

(2) The interval J_x in Theorem 4.3 has been assumed *symmetric* with respect to ζ_0 . The theorem remains, however, true if J_x is assumed as a *one-sided* neighborhood of ζ_0 , that is, one of the intervals between ζ_0 and $\zeta_0 \pm \eta$. However, in this case we must *assume* that $\psi(x)$ remains in J_x , whenever x lies in J_x . On the other hand, this assumption is not necessary if $\psi'(x)$ is non-negative inside J_x .

(3) Theorem 4.3 also remains true in the case of an analytic function $\psi(x)$ of a complex variable. We must then replace J_x by the circle $|x - \zeta_0| < \eta$. In the proof, instead of using the mean value theorem, we can write

$$x_{v+1} - \zeta_0 = \psi(x_v) - \psi(\zeta_0) = \int_{\zeta_0}^{x_v} \psi'(x) dx$$

and obtain at once

$$|x_{v+1} - \zeta_0| \leq (1-m)|x_v - \zeta_0|.$$

14. In Theorem 4.3 the existence of the fixed point ζ_0 has been *assumed*. However, in the hypothesis (4.13) the existence of a fixed point in J_x can be *proved*, using Theorem 2.2, if ζ_0 can be considered in a certain sense as an “approximate fixed point”:

Theorem 4.4. *For the interval J_x ($|x - x_0| \leq \eta$) assume that $\psi'(x)$ exists and satisfies the relation*

$$|\psi'(x)| \leq 1 - m, \quad 0 < m < 1 \quad (|x - x_0| \leq \eta). \quad (4.14)$$

Assume further that $|\psi(x_0) - x_0| \leq \eta m$.

Then there exists in J_x exactly one fixed point ζ_0 of $\psi(x)$, $\psi(\zeta_0) = \zeta_0$.

Proof. Define $f(x)$ as $\psi(x) - x$. Then we have $|f(x_0)| \leq \eta m$ and by virtue of (4.14), everywhere in J_x

$$|f'(x)| = |1 - \psi'(x)| \geq 1 - (1 - m) = m.$$

Then the conditions of Theorem 2.2 are satisfied and $f(x)$ has a unique zero ζ_0 in J_x , for which indeed $\psi(\zeta_0) = \zeta_0$.

15. The analog of Theorem 4.4 for *functions of a complex variable* is the following:

Theorem 4.5. *Let $\psi(z)$ be defined and analytic in the circle K_ρ ($|z - z_0| \leq \rho$). Suppose that for an m , $0 < m < 1$, we have*

$$\sigma \equiv |\psi(z_0) - z_0| < \rho m, \quad (4.15)$$

while in the whole circle K_ρ

$$|\psi'(z)| \leq 1 - m \quad (|z - z_0| \leq \rho). \quad (4.16)$$

Then there exists a ζ_0 inside K_ρ with

$$\psi(\zeta_0) = \zeta_0, \quad |\zeta_0 - z_0| < \rho; \quad (4.17)$$

there is no other fixed point of $\psi(z)$ in K_ρ and, for all z_1 lying in a circle K_ε ($|z - z_0| \leq \varepsilon$) with a convenient ε the sequence (4.1) for $x_1 = z_1$ converges uniformly to ζ_0 .

Proof. From the continuity of $\psi(z)$ it follows that there exist positive numbers ε, δ such that for every z in K_ε we have

$$\frac{|\psi(z) - z|}{m} \leq \rho - \delta \quad (|z - z_0| \leq \varepsilon). \quad (4.18)$$

ε can be chosen $< \text{Min}(\rho, \delta)$.

For a general z from K_ε consider

$$\psi_0(z) = z, \quad \psi_1(z) = \psi(z), \quad \psi_2(z) = \psi(\psi(z)), \dots, \psi_{n+1}(z) = \psi(\psi_n(z))$$

as long as the values of $\psi_1(z), \psi_2(z), \dots, \psi_n(z)$ remain, for all z from K_ε , in K_ρ . Then we have, for $v = 1, 2, \dots, n+1$,

$$\psi_v'(z) = \psi'(\psi_{v-1}(z))\psi'(\psi_{v-2}(z)) \cdots \psi'(\psi_1(z))\psi'(z)$$

and by (4.16)

$$|\psi_v'(z)| \leq (1-m)^v \quad (v = 1, 2, \dots, n+1). \quad (4.19)$$

We have therefore for $v = 1, 2, \dots, n$, integrating along the straight line joining z and $\psi(z)$,

$$\psi_{v+1}(z) - \psi_v(z) = \psi_v(\psi(z)) - \psi_v(z) = \int_z^{\psi(z)} \psi_v'(z) dz,$$

$$|\psi_{v+1}(z) - \psi_v(z)| \leq (1-m)^v |\psi(z) - z| < m(\rho - \delta)(1-m)^v.$$

Therefore, for a general z from K_ε it follows that

$$\psi_{n+1}(z) = z + \sum_{v=0}^n (\psi_{v+1}(z) - \psi_v(z)), \quad (4.20)$$

$$|\psi_{n+1}(z) - z_0| < \varepsilon + \sum_{v=0}^n m(\rho - \delta)(1-m)^v < \varepsilon + m(\rho - \delta) \sum_{v=0}^{\infty} (1-m)^v,$$

$$|\psi_{n+1}(z) - z_0| < \varepsilon + \rho - \delta < \rho.$$

We see that

$$|\psi_{n+1}(z) - z_0| < \rho \quad (|z| \leq \varepsilon),$$

and therefore all $\psi_v(z)$ ($v = 1, 2, \dots$) lie inside K_ρ for $|z| \leq \varepsilon$. But then from (4.20) we have

$$\lim_{n \rightarrow \infty} \psi_n(z) - z_0 = z - z_0 + \sum_{v=0}^{\infty} (\psi_{v+1}(z) - \psi_v(z))$$

and this series is majorized, for $|z - z_0| \leq \varepsilon$, by

$$\varepsilon + \sum_{v=0}^{\infty} m(\rho - \delta)(1-m)^v = \varepsilon + \rho - \delta < \rho$$

so that $\zeta_0 = \lim_{n \rightarrow \infty} \psi_n(z)$ exists and lies inside K_ρ . From $\psi_{n+1}(z) = \psi(\psi_n(z))$ follows (4.17) for $n \rightarrow \infty$.

If there existed another fixed point ζ_1 of $\psi(z)$ in K_ρ we would have

$$\zeta_1 - \zeta_0 = \psi(\zeta_1) - \psi(\zeta_0) = \int_{\zeta_0}^{\zeta_1} \psi'(z) dz,$$

$$|\zeta_0 - \zeta_1| \leq (1-m)|\zeta_1 - \zeta_0|,$$

and it follows that $\zeta_1 = \zeta_0$. Theorem 4.5 is proved.

5

Further Discussion of Iterations. Multiple Zeros

ITERATIONS BY MONOTONIC ITERATING FUNCTIONS

1. Theorem 5.1. Let $f(x)$ be continuous in $J_0: \langle x_0, x_0 + d \rangle$, $f(x_0) \neq 0$, and d so chosen that $f(x_0)d < 0$. Put $\psi(x) = x - f(x)$ and assume further that

$$\frac{f(y) - f(x)}{y - x} \leq 1 \quad (x \in J_0, \quad y \in J_0, \quad x \neq y). \quad (5.1)$$

Form x_v ($v = 0, 1, \dots$) by the iteration $x_{v+1} = \psi(x_v)$, as long as $x_v \in J_0$. Then

(a) If $f(\zeta) = 0$, $\zeta \in J_0$, and $f(x) \neq 0$ in (x_0, ζ) , we have $|x_v - \zeta| \downarrow 0$ as $v \rightarrow \infty$, and all x_v lie in J_0 and even in $\langle x_0, \zeta \rangle$.

(b) If $f(x) \neq 0$ in J_0 , there exists an n_0 such that x_{n_0} does not lie in J_0 .

2. Proof. We begin by showing that $\psi(x)$ is a monotonically increasing function in J_0 . Indeed, we have

$$\frac{\psi(y) - \psi(x)}{y - x} = \frac{y - x - [f(y) - f(x)]}{y - x} = 1 - \frac{f(y) - f(x)}{y - x} \geq 0.$$

Therefore, for $y > x$, $\psi(y) - \psi(x) \geq 0$, $\psi(y) \geq \psi(x)$.

Consider now $x_1 = \psi(x_0) = x_0 - f(x_0)$. We have obviously

(α) If $d > 0$, then $f(x_0) < 0$ and $x_0 < x_1$.

(β) If $d < 0$, then $f(x_0) > 0$ and $x_0 > x_1$.

Now, in case (α) $\psi(x_0) \leq \psi(x_1)$, $x_1 \leq x_2$, and in case (β) $\psi(x_0) \geq \psi(x_1)$, $x_1 \geq x_2$. Repeating this argument and comparing x_2 with x_3 , x_3 with x_4 , etc., we see that in case (α) we have a monotonic increasing sequence $x_0 < x_1 \leq x_2 \leq x_3 \leq \dots$ and in case (β) a monotonic decreasing sequence $x_0 > x_1 \geq x_2 \geq x_3 \geq \dots$.

If at the v th stage the equality holds, i.e., $x_v = x_{v+1} = x_v - f(x_v)$, then $f(x_v) = 0$; furthermore, $x_v = x_k$ ($k = v+1, v+2, \dots$) and our sequence converges to x_v .

3. If none of the x_v gets out of the interval, we have a monotonic sequence contained in J_0 , and this sequence must converge to a limit ζ_0 . Since J_0 is closed, $\zeta_0 \in J_0$. But we have $x_{v+1} = x_v - f(x_v)$. Taking the limit of both sides as $v \rightarrow \infty$, we have, from the continuity of $f(x)$ in J_0 , $\zeta_0 = \zeta_0 - f(\zeta_0)$ and so $f(\zeta_0) = 0$. This proves the assertion (b).

Assume now that the hypothesis of (a) is satisfied. Suppose $d > 0$. Then $x_1 > x_0$ and, since $x_0 < \zeta$, $x_1 = \psi(x_0) \leq \psi(\zeta) = \zeta$; we see that x_1 lies between x_0 and ζ . By repeating the same argument, we see that the x_v increase and are contained in J_0 between x_0 and ζ . Their limit is a zero of $f(x)$ and ζ is the closest zero; therefore $x_v \uparrow \zeta$. As the argument is completely symmetric for $d < 0$, the assertion (a) is proved and hence the whole theorem.

4. If $f(x)$ has a finite first derivative in J_0 , then (5.1) can be replaced by

$$f'(x) \leq 1 \quad (x \in J_0). \quad (5.2)$$

Indeed, if (5.1) holds, we obtain (5.2) in letting y go to x . Suppose, on the other hand, that (5.2) holds. By the mean value theorem of differential calculus, we have

$$\frac{f(y) - f(x)}{y - x} = f'(\xi), \quad \xi \in (y, x), \quad (5.3)$$

and (5.1) follows immediately.

5. Remarks. If (5.2) or the condition $f(x_0)d < 0$ is not satisfied, we can consider $cf(x) = 0$ instead of the equation $f(x) = 0$. By choosing c sufficiently small and of appropriate sign, we have $cf'(x) \leq 1$ and $dcf(x_0) < 0$.

As we have explained in Chapter 4, the convergence generally can be speeded up by choosing c conveniently and considering the equation $cf(x) = 0$. However, this certainly does not work if $f'(\zeta_0) = 0$, that is to say, if we have a multiple root. In this case, Theorem 5.1 can be still applied, and although $\psi'(\zeta_0) = 1$, we still get convergence, but it is very slow. We now consider this case in detail.

MULTIPLE ZEROS

6. Let ζ be a zero of (exact) multiplicity k of $f(x)$ and assume that $f^{(k)}(x)$ is continuous in ζ . Then

$$f(\zeta) = 0, \quad f'(\zeta) = 0, \dots, f^{(k-1)}(\zeta) = 0, \quad f^{(k)}(\zeta) \neq 0. \quad (5.4)$$

By developing $f(x)$ in powers of $x - \zeta$, we obtain

$$f(x) = \frac{(x-\zeta)^k}{k!} f^{(k)}(\zeta + \theta(x-\zeta)), \quad 0 < \theta < 1. \quad (5.5)$$

Since $f^{(k)}(x)$ is continuous in ζ , we have, as $x \rightarrow \zeta$,

$$\frac{f(x)}{(x-\zeta)^k} \rightarrow \frac{f^{(k)}(\zeta)}{k!} \neq 0. \dagger \quad (5.6)$$

In the following discussion, we assume only that $f(x)$ vanishes in ζ in such a way that for $k > 1$, $f(x)/|x-\zeta|^k \rightarrow A \neq 0$ (either for $x \uparrow \zeta$ or for $x \downarrow \zeta$). It is not necessary that k be an integer. Let $k = 1 + \alpha$, where $\alpha > 0$. In the case of Theorem 5.1, we have either $x_v \uparrow \zeta$ or $x_v \downarrow \zeta$ where the x_v are defined by $x_{v+1} = x_v - f(x_v)$. We are now going to prove a result about the distances of the x_v from ζ .

Theorem 5.2. *Let $f(x)$ be defined in a one-sided neighborhood J_0 of ζ : $\langle \zeta, \zeta + d \rangle$ and be continuous at ζ . Let $f(\zeta) = 0$. Suppose that the sequence x_v defined by the iteration formula $x_{v+1} = x_v - f(x_v)$ lies in J_0 and tends to ζ through J_0 , and that, if x tends to ζ through J_0 , we have for an $\alpha > 0$*

$$\frac{f(x)}{|x-\zeta|^{1+\alpha}} \rightarrow A \neq 0 \quad (5.7)$$

with a finite $A \neq 0$. Then the following relation holds:

$$v^{1/\alpha} |\zeta - x_v| \rightarrow \frac{1}{|\alpha A|^{1/\alpha}} \quad (v \rightarrow \infty). \quad (5.8)$$

7. Proof. We can assume without loss of generality that A and the values of $f(x)$ in J_0 are > 0 . By (5.7) $f'(\zeta)$ exists as a one-sided derivative. We have

$$\frac{\zeta - x_{v+1}}{\zeta - x_v} = \frac{\zeta - x_v - [f(\zeta) - f(x_v)]}{\zeta - x_v} \rightarrow 1 - f'(\zeta). \quad (5.9)$$

But by (5.7)

$$\frac{f(x) - f(\zeta)}{x - \zeta} = \frac{f(x)}{x - \zeta} \rightarrow 0 \quad (5.10)$$

as $x \rightarrow 0$ through J_0 , i.e., $f'(\zeta) = 0$ and

$$\frac{\zeta - x_{v+1}}{\zeta - x_v} \rightarrow 1. \quad (5.11)$$

Introduce the positive numbers

$$q_v = A |\zeta - x_v|^\alpha \quad (5.12)$$

† As a matter of fact, the continuity of $f^{(k)}(x)$ in ζ is not necessary for (5.6). It is sufficient that $f^{(k)}(\zeta)$ exists. A more detailed study of the situation implied by a formula of the type (5.6) is to be found in Sections 2-12 of Appendix N.

and consider

$$d_v = \frac{1}{q_{v+1}} - \frac{1}{q_v} = \frac{1}{q_{v+1}} \left(1 - \frac{q_{v+1}}{q_v} \right). \quad (5.13)$$

We shall prove that d_v has a finite nonvanishing limit as $v \rightarrow \infty$. Let

$$p_v = \frac{f(x)}{x_v - \zeta}. \quad (5.14)$$

From the recurrence formula

$$x_{v+1} - \zeta = (x_v - \zeta) - f(x_v),$$

where $f(x_v) > 0$ and all $x_v - \zeta$ have the same sign, it follows that all $x_v - \zeta$ are > 0 , since otherwise, $x_v - \zeta$ cannot tend to 0. We see that

$$x_v - \zeta > 0. \quad (5.15)$$

From (5.15) follows

$$p_v > 0,$$

$$p_v = \frac{f(x_v)}{x_v - \zeta} \sim \frac{A |\zeta - x_v|^{1+\alpha}}{x_v - \zeta} = A |\zeta - x_v|^{\alpha}, \quad (5.16)$$

$$p_v \sim |q_v| = q_v, \quad p_v \rightarrow 0. \quad (5.17)$$

From (5.15) we have further

$$\frac{q_{v+1}}{q_v} = \left(\frac{x_{v+1} - \zeta}{x_v - \zeta} \right)^{\alpha} = (1 - p_v)^{\alpha} = 1 - \alpha p_v + O(p_v^2) \quad (v \rightarrow \infty), \quad (5.18)$$

$$\frac{1 - q_{v+1}/q_v}{p_v} \rightarrow \alpha. \quad (5.19)$$

Since $p_v \sim q_v \sim q_{v+1}$, we have from (5.19)

$$\frac{1}{q_{v+1}} \left(1 - \frac{q_{v+1}}{q_v} \right) \rightarrow \alpha,$$

and from (5.13) follows

$$d_v \rightarrow \alpha. \quad (5.20)$$

8. In order to complete our proof, we need the following well-known theorem by Cauchy:

Let $a_v \rightarrow \alpha$ ($v = 1, 2, \dots$). Then the sequence

$$\frac{a_1}{1}, \quad \frac{a_1 + a_2}{2}, \quad \dots, \quad \frac{a_1 + \dots + a_n}{n}, \quad \dots$$

converges to α .

Now identify a_v with d_v . Then

$$\frac{d_1 + \cdots + d_n}{n} \rightarrow \alpha$$

and since

$$d_1 + \cdots + d_n = \left(\frac{1}{q_2} - \frac{1}{q_1} \right) + \cdots + \left(\frac{1}{q_{n+1}} - \frac{1}{q_n} \right) = \frac{1}{q_{n+1}} - \frac{1}{q_1},$$

we have

$$\frac{1/q_{v+1} - 1/q_1}{v} \rightarrow \alpha, \quad \frac{1}{vq_{v+1}} \rightarrow \alpha.$$

But $v/(v+1) \rightarrow 1$, hence $1/q_{v+1}(v+1) \rightarrow \alpha$, i.e.,

$$\frac{1}{vq_v} \rightarrow \alpha, \quad vq_v \rightarrow \frac{1}{\alpha}, \quad (5.21)$$

and by (5.12)

$$v|\zeta - x_v|^\alpha \rightarrow \frac{1}{\alpha A},$$

$$v^{1/\alpha} |\zeta - x_v| \rightarrow \frac{1}{|\alpha A|^{1/\alpha}}, \quad \text{Q.E.D.} \quad (5.22)$$

9. If α is at least 1, i.e., if we have a multiple zero of at least multiplicity 2, then we have in (5.22) a power of v which is at most 1 and $|\zeta - x_v|$ goes to 0 very slowly. (The efficiency index in this case is = 1.) Thus, this method is not one which would be used by computers. The result is still useful in that it tells us that in the case of multiple roots we should look for other methods.[†]

CONNECTION OF THE REGULA FALSI WITH THE THEORY OF ITERATION

10. Let us consider the following case of the iteration by the *regula falsi*:

$$x_{v+1} = \frac{af(x_v) - x_v f(a)}{f(x_v) - f(a)}. \quad (5.23)$$

This is an example of an iteration where the iterating function $\psi(x)$ is

$$\psi(x) = \frac{af(x) - xf(a)}{f(x) - f(a)} \quad (5.24)$$

[†] See, for example, Chapter 8 and Appendix E.

and, for a ζ with $f(\zeta) = 0$,

$$\psi'(\zeta) = 1 - f'(\zeta) \frac{a-\zeta}{f(a)-f(\zeta)}. \quad (5.25)$$

From Lemma 4.2 we see that ζ will be a point of attraction if $|\psi'(\zeta)| < 1$, i.e., if

$$0 < \frac{f'(\zeta)}{[f(a)-f(\zeta)]/(a-\zeta)} < 2. \quad (5.26)$$

The case where $|\psi'(\zeta)| = 1$ is an exceptional case and will not be treated here. Notice that the slope of the chord from a to ζ must have the same sign as $f'(\zeta)$. Furthermore, the modulus of the slope of the chord must be greater than $\frac{1}{2}|f'(\zeta)|$. We illustrate this by Fig. 2.

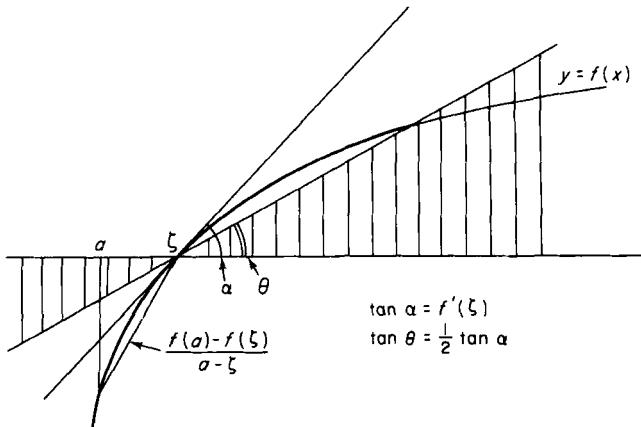


FIGURE 2

In Fig. 2, our point $(a, f(a))$ must so be chosen that the chord from it to ζ will not fall within the hatched area. In other words, a must be so chosen that the point whose coordinates are a and $f(a)$ falls on the darkened portion of the curve. This gives us a rule by which we can choose our point a at a greater distance from ζ and still be assured of convergence. If our chord falls within the hatched area, $|\psi'(\zeta)| > 1$ and ζ is a point of definite repulsion.

Observe finally that if $\psi'(\zeta) < 0$, the sequence x_n , if it converges to ζ_0 , converges alternatively.

6

The Newton–Raphson Method

THE IDEA OF THE NEWTON–RAPHSON METHOD

1. Assume that we are given two coincident interpolation points $x_1 = a$, $x_2 = a$ and that we know $f(a)$ and $f'(a)$. We wish to find a linear interpolating polynomial which approximates $f(x)$ in such a way that it is equal to $f(a)$ at a and its first derivative is equal to $f'(a)$ at a . This polynomial is precisely the Taylor's polynomial in $x - a$, i.e.,

$$f(a) + (x-a)f'(a). \quad (6.1)$$

We equate to zero and solve for x :

$$x = a - \frac{f(a)}{f'(a)} \quad (f'(a) \neq 0). \quad (6.2)$$

This expression gives an approximation to a root of $f(x) = 0$ which lies in a neighborhood of a . This idea was originally due to Newton, but Raphson was the first to express it in the form (6.2), and today (6.2) is known in England as the Newton–Raphson formula.

In (6.2) if $a = x_0$ is a first approximation, then $x = x_1$ will be the second approximation. Substituting x_1 for a in (6.2), we get a third approximation x_2 . Generally,

$$x_{v+1} = x_v - \frac{f(x_v)}{f'(x_v)}. \quad (6.3)$$

In this way we obtain a sequence of numbers x_v ($v = 0, 1, \dots$). We consider the following problems: Can we conclude from the properties of this sequence that a zero of $f(x)$ exists in a neighborhood of the x_v ? Does the sequence tend to this zero, and if so, how good is the approximation by x_v ?

THE USE OF INVERSE INTERPOLATION

2. We could discuss these questions by considering the linear interpolating polynomial in (6.1), but here again we get our results more quickly by studying the inverse function.

Assume that $f'(x) \neq 0$ in the considered neighborhood of x_0 . Let $y = f(x)$, $x = \varphi(y)$, $y_0 = f(x_0)$, $\varphi(y_0) = x_0$, $\varphi'(y_0) = 1/f'(x_0)$. Then our interpolating polynomial for $\varphi(y)$ is

$$L(y) = x_0 + (y - y_0) \frac{1}{f'(x_0)}. \quad (6.4)$$

Substituting in (1B.23), we get

$$\varphi(y) - L(y) = \frac{\varphi''(\eta)}{2}(y - y_0)^2. \quad (6.5)$$

Using x_0 as our initial approximation and substituting 0 for y in (6.5), we have, putting $\varphi(0) = \zeta$,

$$\zeta - x_1 = \left[-\frac{f''(\xi)}{2f'(\xi)^3} \right] f(x_0)^2, \quad \xi \in (\zeta, x_0). \quad (6.6)$$

3. In order to apply (6.6), we must have an upper bound for $|f''(x)|$ and a lower bound for $|f'(x)|$. Notice that $\zeta - x_1$ is quadratic in $f(x_0)$. The value of $f(x_0)$ was computed in determining x_1 and so we can immediately get an upper bound for the distance of x_1 to ζ . Let $|f''(x)| \leq M_2$, $|f'(x)| \geq m_1 > 0$; then

$$|\zeta - x_1| \leq \frac{M_2}{2m_1^3} f(x_0)^2 =: k. \quad (6.7)$$

If in our neighborhood of ζ , $f''(x)$ does not change its sign, then the sign of the bracketed expression in (6.6) is fixed and we can find a new neighborhood of x_1 in which ζ lies. If the sign of this bracketed expression is positive, we are sure that ζ lies in the interval $\langle x_1, x_1 + k \rangle$. If the sign is negative, then ζ lies in the interval $\langle x_1 - k, x_1 \rangle$.

In this discussion we have assumed that a root ζ exists in the neighborhood of x_0 . Often in practice we must proceed with our computation before this is known. This is indeed justified, for often after the first steps of the computation we can tell whether there is a zero in the considered neighborhood and whether our sequence is convergent. We deal with this in more detail in Chapter 7.

In (6.6) we can introduce the distance of x_0 to ζ . For we have $f(x_0) - f(\zeta) = (x_0 - \zeta)f'(\xi_0)$, $\xi_0 \in (x_0, \zeta)$. Substituting this in (6.6), we obtain

$$\zeta - x_1 = -\frac{1}{2} \frac{f''(\xi)}{f'(\xi)^3} f'(\xi_0)^2 (\zeta - x_0)^2, \quad \xi_0 \wedge \xi \in (x_0, \zeta), \quad (6.8)$$

$$\frac{\zeta - x_1}{(\zeta - x_0)^2} \rightarrow -\frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)} \quad (x_0 \rightarrow \zeta). \quad (6.9)$$

We see that our approximation will, generally speaking, be improved

quadratically at each step. If the product of the factors multiplying $(\zeta - x_0)^2$ in (6.8) is absolutely ≤ 1 , and if the approximation x_0 has n exact decimal places then x_1 will have at least $2n$ exact decimal places.

COMPARISON OF REGULA FALSI AND NEWTON-RAPHSON METHOD

4. In Chapter 3, Section 11, we showed that the efficiency index of the *regula falsi* used according to (3.13) is $1.618\dots$.

In applying the Newton–Raphson formula, it is necessary to compute f and f' for each step; i.e., each step of computation is made at the expense of two horners; therefore, $\sqrt{2} = 1.414\dots$ is the efficiency index of the Newton–Raphson method. We see that the *regula falsi* used according to (3.13) is *better* than the Newton–Raphson method.

This advantage is restricted only to the solution of ordinary equations. The *regula falsi* is useful only for this purpose, and its applications do not easily extend to other theories.[†] On the other hand, the principle of the Newton–Raphson method can be extended to the theories of differential equations, integral equations, functional equations, and many other branches of analysis.

The success of the method depends of course on the choice of x_0 . In this respect not much is known.

[†] Cf., however, the papers listed in the bibliographical note to Appendix D.

7

Fundamental Existence Theorems in the Newton–Raphson Iteration

ERROR ESTIMATES A PRIORI AND A POSTERIORI

1. It may be possible to obtain information and error estimates in terms of the starting data; e.g., if $f(x)$ is continuous in $\langle a, b \rangle$ and changes its sign in this interval, then a zero exists in $\langle a, b \rangle$. Such an estimate would be called an estimate *a priori*. In practice we often begin our computation and just hope that we get a zero. If we use the results of the computation, we can often obtain estimates which are much better than the *a priori* estimates. Such estimates are called *a posteriori*. *A priori* estimates are sometimes 100 to 1000 times too large. *A posteriori* estimates may be of the right order of magnitude.

FUNDAMENTAL EXISTENCE THEOREMS

2. We give in what follows a theorem on a real function of a real variable in its simplest form and prove it in Sections 4–8. We then formulate the analogous theorem in the complex case (Theorem 7.2). Both theorems, however, are special cases of a more refined and more sophisticated theorem on the Newton–Raphson method in Banach spaces (Chapter 38), so that we do not give a special proof of Theorem 7.2.

In the following theorems *we do not make any assumptions about the existence of a zero*.

3. Theorem 7.1. *Let $f(x)$ be a real function of the real variable x , $f(x_0)f'(x_0) \neq 0$, and put $h_0 = -f(x_0)/f'(x_0)$, $x_1 = x_0 + h_0$. Consider the interval $J_0: \langle x_0, x_0 + 2h_0 \rangle$ and assume that $f''(x)$ exists in J_0 , that $\text{Sup}_{J_0} |f''(x)| = M$, and*

$$2|h_0|M \leq |f'(x_0)|. \quad (7.1)$$

Form, starting with x_0 , the sequence x_v by the recurrence formula

$$x_{v+1} = x_v - \frac{f(x_v)}{f'(x_v)} \quad (v = 0, 1, \dots).$$

Then all x_v lie in J_0 and we have

$$x_v \rightarrow \zeta \quad (v \rightarrow \infty), \quad (7.2)$$

where ζ is the only zero in J_0 . Unless $\zeta = x_0 + 2h_0$, ζ is a simple zero.[†] Further, we have the relations

$$(a) \quad \frac{|x_{v+1} - x_v|}{|x_v - x_{v-1}|^2} \leq \frac{M}{2|f'(x_v)|} \quad (v = 1, 2, \dots),$$

$$(b) \quad |\zeta - x_{v+1}| \leq \frac{M}{2|f'(x_v)|} |x_v - x_{v-1}|^2 \quad (v = 1, 2, \dots).$$

4. Proof. We have

$$\begin{aligned} f'(x) - f'(x_0) &= \int_{x_0}^x f''(x) dx, \\ |f'(x) - f'(x_0)| &\leq |x - x_0|M \end{aligned} \quad (7.3)$$

and by (7.1)

$$|f'(x_1) - f'(x_0)| \leq |h_0|M \leq \frac{|f'(x_0)|}{2}. \quad (7.4)$$

Furthermore, from (7.4),

$$\begin{aligned} |f'(x_1)| &\geq |f'(x_0)| - |f'(x_1) - f'(x_0)| \geq |f'(x_0)| - \frac{|f'(x_0)|}{2}, \\ |f'(x_1)| &\geq \frac{|f'(x_0)|}{2}. \end{aligned} \quad (7.5)$$

Now, integrating by parts, we have

$$\begin{aligned} \int_{x_0}^{x_1} (x_1 - x) f''(x) dx &= -(x_1 - x_0) f'(x_0) + f(x_1) - f(x_0) \\ &= -h_0 f'(x_0) - f(x_0) + f(x_1) = f(x_1), \\ f(x_1) &= \int_{x_0}^{x_1} (x_1 - x) f''(x) dx. \end{aligned} \quad (7.6)$$

We introduce here a new variable of integration t , putting

$$\begin{aligned} x &= x_0 + th_0, \\ x_1 - x &= x_1 - x_0 - th_0 = h_0 - th_0 = h_0(1-t), \quad dx = h_0 dt; \\ f(x_1) &= h_0^2 \int_0^1 (1-t) f''(x_0 + th_0) dt. \end{aligned}$$

[†] It can be proved that if ζ lies at $x_0 + 2h_0$, then $f(x)$ is a quadratic polynomial with the double root ζ . See Chapter 40.

5. It now follows that

$$|f(x_1)| \leq |h_0|^2 \int_0^1 (1-t) |f''(x_0 + th_0)| dt$$

and since $1-t \geq 0$,

$$\begin{aligned} |f(x_1)| &\leq M |h_0|^2 \int_0^1 (1-t) dt = \frac{|h_0|^2 M}{2}, \\ |f(x_1)| &\leq \frac{1}{2} |h_0|^2 M. \end{aligned} \quad (7.7)$$

Let $h_1 = -f(x_1)/f'(x_1)$, $x_2 = x_1 + h_1$. Applying (7.5) and (7.7), we get

$$\begin{aligned} |h_1| &\leq \frac{|h_0|^2 M}{|f'(x_0)|}, \\ \frac{M|h_1|}{|f'(x_1)|} &\leq \frac{|h_0|^2 M^2}{|f'(x_0)| |f'(x_1)|} \leq \frac{|h_0|^2 M^2}{|f'(x_0)| \frac{1}{2} |f'(x_0)|}, \\ \frac{2M|h_1|}{|f'(x_1)|} &\leq \frac{2^2 |h_0|^2 M^2}{|f'(x_0)|^2} = \left(\frac{2|h_0|M}{|f'(x_0)|}\right)^2. \end{aligned} \quad (7.8)$$

By (7.1), the expression in parentheses is ≤ 1 . Hence

$$2|h_1|M \leq |f'(x_1)|. \quad (7.9)$$

Now by (7.8)

$$\begin{aligned} \frac{|h_1|}{|h_0|} &\leq \frac{1}{2} \left(\frac{2|h_0|M}{|f'(x_0)|} \right) \leq \frac{1}{2}, \\ |h_1| &\leq \frac{1}{2} |h_0|. \end{aligned} \quad (7.10)$$

From (7.10), we see that the point x_2 will not get beyond the distance $\frac{1}{2}|h_0|$ from x_1 and will remain in J_0 . Further, the interval $J_1: (x_1, x_1 + 2h_1)$ lies in J_0 .

6. We let generally

$$x_{v+1} = x_v + h_v, \quad h_v = -\frac{f(x_v)}{f'(x_v)}. \quad (7.11)$$

Relation (7.9) shows that the hypotheses of our theorem remain true if we replace x_0 and h_0 by x_1 and h_1 , respectively, and consequently the hypotheses remain true for x_v and h_v ($v = 0, 1, \dots$).

Let J_v ($v = 0, 1, \dots$) be the interval $(x_v, x_v + 2h_v)$. We have a sequence of intervals $J_{v+1} \in J_v$ (J_{v+1} is contained in J_v) with the radius of J_{v+1} (the radius of an interval is half its length) at the most equal to one-half the radius of J_v . We know that such a sequence converges to a point ζ . Since each of the

intervals lies in J_0 and J_0 is closed, ζ lies in J_0 :

$$x_v \rightarrow \zeta, \quad \zeta \in J_0. \quad (7.12)$$

To prove that ζ is a root of $f(x)$, multiply (7.11) by $f'(x_v)$: $f'(x_v)x_{v+1} = f'(x_v)x_v - f(x_v)$. Taking the limit on both sides as $x_v \rightarrow \zeta$, we have $f'(\zeta)\zeta = f'(\zeta)\zeta - f(\zeta)$ and

$$f(\zeta) = 0. \quad (7.13)$$

7. By (7.3) we have for all x from J_0

$$|f'(x) - f'(x_0)| \leq |x - x_0|M \leq 2|h_0|M.$$

Assume that $|x - x_0| < |h_0|$, i.e., that x lies *inside* (J_0). Then

$$|f'(x) - f'(x_0)| < 2|h_0|M \leq |f'(x_0)|, \quad (7.14)$$

and we see that $f'(x) \neq 0$ for x inside (J_0). ζ is therefore a simple root of $f(x)$ if it lies inside (J_0).

We prove now that ζ is the only root in J_0 . But $f'(x)$ does not vanish in J_0 . Hence $f'(x)$ is strictly monotonically increasing or decreasing in J_0 and thus has only one root.

8. The assertions (a) and (b) of Theorem 7.1 can now be easily deduced:
(a) is equivalent to

$$|h_v| \leq \frac{M|h_{v-1}|^2}{2|f'(x_v)|} \quad (v = 1, 2, \dots).$$

We use (7.7) and the fact that our starting assumptions are true for all v ($v = 0, 1, \dots$). We have therefore

$$|f(x_v)| \leq \frac{M}{2}|h_{v-1}|^2$$

and by (7.11)

$$|h_v| \leq \frac{M|h_{v-1}|^2}{2|f'(x_v)|},$$

which is (a).

To prove (b), we notice that ζ lies in an interval with center x_{v+1} and radius $|h_v|$, i.e., $|\zeta - x_{v+1}| \leq |h_v|$, and use (a). Our theorem is completely proved.

9. Theorem 7.2. Let $f(z)$ be a complex function of the complex variable z in a neighborhood of z_0 , $f(z_0)f'(z_0) \neq 0$, $h_0 = -f(z_0)/f'(z_0)$, $z_1 = z_0 + h_0$. Consider the circle K_0 : $|z - z_1| \leq |h_0|$, and assume $f(z)$ analytic in K_0 . $\max_{K_0} |f''(z)| = M$ and $2|h_0|M \leq |f'(z_0)|$. Form, starting with z_0 , the sequence

z_v by the recurrence formula

$$z_{v+1} = z_v - \frac{f(z_v)}{f'(z_v)} \quad (v = 0, 1, \dots).$$

Then all z_v lie in K_0 and we have

$$z_v \rightarrow \zeta \quad (v \rightarrow \infty),$$

where ζ is the only zero of $f(z)$ in K_0 . ζ is a simple root unless it lies on the boundary of K_0 . Further, we have the relations valid for $v = 1, 2, \dots$:

$$\frac{|z_{v+1} - z_v|}{|z_v - z_{v-1}|^2} \leq \frac{M}{2|f'(z_v)|}, \quad (7.15)$$

$$|\zeta - z_{v+1}| \leq \frac{M}{2|f'(z_v)|} |z_v - z_{v-1}|^2, \quad (7.16)$$

$$|z_{v+1} - z_v| \leq \frac{1}{2} |z_v - z_{v-1}|, \quad (7.17)$$

$$2|z_{v+1} - z_v| M \leq |f'(z_v)|, \quad (7.18)$$

where the equality sign only holds if $f(z)$ is a quadratic polynomial with a double zero.

The formulas (7.17) and (7.18) correspond to the relations (7.10) and (7.9) obtained in the course of the proof of Theorem 7.1.

10. It is clear that a zero exists if we begin computing by the Newton-Raphson method and if at one of the steps the inequality $2|h_v|M \leq |f'(x_v)|$ holds. It is essential that $f'(x_0)$ is not zero. It is just as dangerous if $f'(\zeta) = 0$, for if we are sufficiently close to ζ , $|f'(x_0)|$ will be very small. On the other hand, if ζ is a simple zero and if x_0 is sufficiently close to ζ , our conditions will certainly be satisfied.

If ζ is a multiple zero and we use the Newton-Raphson method, we find as in Chapter 5 the convergence too slow for computing purposes or we do not get any convergence at all. A modification which can be used in this case is discussed in the following chapter.

A complete convergence discussion in the case of a quadratic polynomial is given in Appendix F.

For some modifications and an improvement of the Newton-Raphson method, see Appendix G. See also Appendix I, Section 6.

8

An Analog of the Newton–Raphson Method for Multiple Roots

CONVERGENCE OF SCHRÖDER'S ITERATION FOR MULTIPLE ROOTS

1. An analog of the Newton–Raphson formula, which applies in the case of multiple roots, has been given by Schröder [5].

Suppose that we have a zero of *exact multiplicity p*. We then replace the Newton–Raphson formula by Schröder's formula:

$$x_{v+1} = x_v - p \frac{f(x_v)}{f'(x_v)} \quad (v = 1, 2, \dots).^{\dagger} \quad (8.1)$$

With this modification, the difficulties due to the multiplicity will vanish and the convergence remains superlinear.

2. We say that ζ is a zero of $f(x)$ of *exact multiplicity p* if the derivatives $f^{(k)}(\zeta)$ ($x = 1, \dots, p$) exist and

$$f(\zeta) = f'(\zeta) = \dots = f^{(p-1)}(\zeta) = 0, \quad f^{(p)}(\zeta) \neq 0. \quad (8.2)$$

We assume from now on in this chapter that $p > 1$. We obtain then, developing $f(x)$ and $f'(x)$ at ζ ,

$$f(x) = A(x - \zeta)^p(1 - \varepsilon), \quad f'(x) = pA(x - \zeta)^{p-1}(1 - \varepsilon_1) \quad (8.3)$$

where ε and ε_1 , like $\varepsilon_2, \dots, \varepsilon', \varepsilon_1', \dots$ in the following, denote expressions tending to 0 as $x \rightarrow \zeta$, and A is given by

$$A := \frac{1}{p!} f^{(p)}(\zeta). \quad (8.4)$$

From both formulas (8.3) it follows that

$$\frac{f(x)}{f'(x)} = \frac{x - \zeta}{p} (1 + \varepsilon_2) \quad (8.5)$$

[†] Equation (8.1) is obtained applying (6.3) to $f(x)^{1/p}$.

where

$$\varepsilon_2 = \frac{1+\varepsilon}{1+\varepsilon_1} - 1 = \frac{\varepsilon - \varepsilon_1}{1+\varepsilon_1}.$$

3. Choose now a positive integer $p' \leq p$. If from a given approximation x to ζ we form the next approximation x' by the formula

$$x' = x - p' \frac{f(x)}{f'(x)}$$

it follows from (8.5) as $x \rightarrow \zeta$ that

$$x' - \zeta = \left(1 - \frac{p'}{p}\right)(x - \zeta) - \frac{p'}{p} \varepsilon_2 (x - \zeta), \quad (8.6)$$

$$x' - \zeta = \left(1 - \frac{p'}{p}\right)(x - \zeta) + o(x - \zeta). \quad (8.7)$$

It follows immediately from formula (8.7) that if we apply the original Newton-Raphson formula ($p' = 1$) in our case we will have for the sequence x_v , if we start in a sufficiently close neighborhood of ζ , $(x_{v+1} - \zeta)/(x_v - \zeta) \rightarrow 1 - 1/p$. This is *linear convergence* and is very weak for large values of p .

A similar result is also obtained if formula (8.1) is applied, replacing p with p' , $1 < p' < p$. Here, too, we obtain only linear convergence.

On the other hand, if formula (8.1) is applied, using the *exact* multiplicity p , formula (8.6) becomes

$$x' - \zeta = -\varepsilon_2 (x - \zeta) \quad (8.8)$$

and it follows that

$$x' - \zeta = o(x - \zeta) \quad (x \rightarrow \zeta),$$

so that in this case we certainly have a superlinear convergence.

4. The convergence of (8.1) becomes quadratic if we assume a little more about the derivative $f^{(p)}(x)$, namely, that $f^{(p)}(x)$ exists in a neighborhood of ζ and satisfies at ζ the Lipschitz condition

$$f^{(p)}(x) - f^{(p)}(\zeta) = O(|x - \zeta|) \quad (x \rightarrow \zeta). \quad (8.9)$$

In this case both ε and ε_1 in (8.3) are $O(|x - \zeta|)$, the same holds for ε_2 , and from (8.8) it follows that

$$x' - \zeta = O(|x - \zeta|^2) \quad (x \rightarrow \zeta). \quad (8.10)$$

Here we have indeed quadratic convergence, if we start in a sufficiently close neighborhood of ζ .

5. In order to discuss the asymptotic behavior of $x' - \zeta$ as $x \rightarrow \zeta$, assume now that $f^{(p+1)}(\zeta)$ also exists. Then formulas (8.3) can be refined. Developing $f(x)$ and $f'(x)$ at ζ , we obtain

$$\begin{aligned} f(x) &= A(x-\zeta)^p + B(x-\zeta)^{p+1}(1+\varepsilon'), \\ f'(x) &= pA(x-\zeta)^{p-1} + (p+1)B(x-\zeta)^p(1+\varepsilon_1') \end{aligned} \quad (8.11)$$

where B is given by

$$B := \frac{1}{(p+1)!} f^{(p+1)}(\zeta). \quad (8.12)$$

Comparing (8.11) with (8.3), we obtain

$$\varepsilon = \frac{B}{A}(x-\zeta)(1+\varepsilon'), \quad \varepsilon_1 = \frac{p+1}{p} \frac{B}{A}(x-\zeta)(1+\varepsilon_1')$$

and therefore

$$\varepsilon_2 = \frac{1+\varepsilon}{1+\varepsilon_1} - 1 = -\frac{(x-\zeta)B/(pA) + o(x-\zeta)}{1+\varepsilon_1} = -\left(\frac{B}{pA} + \varepsilon_3'\right)(x-\zeta).$$

It now follows from (8.8) that

$$x' - \zeta = \left(\frac{B}{pA} + o(1)\right)(x-\zeta)^2.$$

Inserting the values of A and B from (8.4) and (8.12), we finally get

$$x' - \zeta = \frac{1}{p(p+1)} \frac{f^{(p+1)}(\zeta)}{f^{(p)}(\zeta)} (x-\zeta)^2 + o(|x-\zeta|^2). \quad (8.13)$$

ERROR ESTIMATES A PRIORI

6. While (8.13) gives good information about the behavior of $x' - \zeta$ in terms of $x - \zeta$, if we apply the Schröder formula (8.1), it cannot of course be used for error estimates, since the values of the derivatives entering into (8.13) are usually unknown. In order to derive explicit inequalities corresponding to (8.13) we now assume further that $f^{(p+1)}$ exists and is continuous in a neighborhood of ζ . In this case our discussion is based on the following identity:

$$F(x) := \frac{1}{(p-1)!} \int_x^\zeta (x-t)^{p-1} (t-\zeta) f^{(p+1)}(t) dt = pf(x) - (x-\zeta) f'(x), \quad (8.14)$$

which holds for $p \geq 1$ if $f(x)$ has a zero of multiplicity $\geq p$ at ζ .

We first prove (8.14) for $p = 1$. In this case $F(x)$ becomes, integrating by parts,

$$\int_x^\zeta (t - \zeta) f''(t) dt = (t - \zeta) f'(t) \Big|_x^\zeta - \int_x^\zeta f'(t) dt = f(x) - (x - \zeta) f'(x),$$

and this is (8.14) with $p = 1$.

We can therefore assume, proving (8.14), that $p > 1$ and that this formula is true if p is replaced with $p - 1$. On the other hand, both sides of this formula vanish for $x = \zeta$. It is therefore sufficient to prove that the derivatives on both sides are equal. But the derivative on the left is

$$\frac{-1}{(p-1)!} (x-t)^{p-1} (t-\zeta) f^{(p+1)}(t) \Big|_{t=x} + \frac{1}{(p-2)!} \int_x^\zeta (x-t)^{p-2} (t-\zeta) f^{(p+1)}(t) dt,$$

where the first term vanishes. Since the derivative of the right-hand expression in (8.14) is $(p-1) f'(x) - (x - \zeta) f''(x)$, the differentiated formula becomes

$$\frac{1}{(p-2)!} \int_x^\zeta (x-t)^{p-2} (t-\zeta) f^{(p+1)}(t) dt = (p-1) f'(x) - (x - \zeta) f''(x),$$

and this is the formula (8.14) in which p is replaced with $p - 1$ and $f(x)$ with $f'(x)$. Formula (8.14) is proved.

7. $F(x)$ becomes, if a new integration variable τ is introduced by $t = \zeta + \tau(x - \zeta)$,

$$F(x) = -\frac{(x-\zeta)^{p+1}}{(p-1)!} \int_0^1 (1-\tau)^{p-1} \tau f^{(p+1)}(\zeta + \tau(x-\zeta)) d\tau.$$

In the above integral the cofactor of $f^{(p+1)}$ is ≥ 0 . We can therefore write in the real case

$$F(x) = -\frac{(x-\zeta)^{p+1}}{(p-1)!} T f^{(p+1)}(\xi_1), \quad \xi_1 \in \langle x, \zeta \rangle,$$

and in the general case

$$F(x) = \theta^* \frac{(x-\zeta)^{p+1}}{(p-1)!} T M_{p+1}, \quad M_{p+1} = \max |f^{(p+1)}(u)| \quad (|u - \zeta| \leq |x - \zeta|)$$

where

$$T := \int_0^1 (1-\tau)^{p-1} \tau d\tau = -\frac{(1-\tau)^p \tau}{p} \Big|_0^1 + \frac{1}{p} \int_0^1 (1-\tau)^p d\tau = \frac{1}{p(p+1)}.$$

It follows therefore, replacing x with x_v , that in the real case

$$F(x_v) = -\frac{(x_v - \zeta)^{p+1}}{(p+1)!} f^{(p+1)}(\xi_1), \quad \xi_1 \in \langle x_v, \zeta \rangle \quad (8.15)$$

and in the general case

$$F(x_v) = \theta * \frac{(x_v - \zeta)^{p+1}}{(p+1)!} M_{p+1}, \quad M_{p+1} = \text{Max} |f^{(p+1)}(u)| \quad (|u - \zeta| \leq |x_v - \zeta|). \quad (8.16)$$

On the other hand, developing $f'(x_v)$ at ζ , we obtain, since $f^{(p)}(\zeta)$ is the first derivative $\neq 0$, in the real case

$$f'(x_v) = \frac{(x_v - \zeta)^{p-1}}{(p-1)!} f^{(p)}(\xi_2), \quad \xi_2 \in \langle x_v, \zeta \rangle \quad (8.17)$$

and in the general case

$$f'(x_v) = \frac{(x_v - \zeta)^{p-1}}{(p-1)!} \left[f^{(p)}(\zeta) + \theta * \frac{(x_v - \zeta)}{p} M_{p+1} \right] \quad (8.18)$$

where M_{p+1} has the same meaning as in (8.16).

8. It now follows that in the real case, by (8.14) and (8.15),

$$\frac{(x_v - \zeta)^{p+1}}{(p+1)!} f^{(p+1)}(\xi_1) = f'(x_v) \left[x_v - \zeta - p \frac{f(x_v)}{f'(x_v)} \right] = f'(x_v)(x_{v+1} - \zeta)$$

and, using (8.17),

$$x_{v+1} - \zeta = (x_v - \zeta)^2 \frac{f^{(p+1)}(\xi_1)}{p(p+1)f^{(p)}(\xi_2)}. \quad (8.19)$$

In the general case we have similarly from (8.14), (8.16), and (8.18)

$$|x_{v+1} - \zeta| \leq \frac{M_{p+1}|x_v - \zeta|^2}{(p+1)[p|f^{(p)}(\zeta)| - M_{p+1}|x_v - \zeta|]} \quad (8.20)$$

where M_{p+1} is defined in (8.16).

RECURRENT ERROR ESTIMATES

If we now assume that

$$|x_v - \zeta| \leq \frac{1}{2} \frac{|f^{(p)}(x_v)|}{M_{p+1}}, \quad (8.21)$$

it follows that

$$|f^{(p)}(\zeta)| \geq |f^{(p)}(x_v)| - |x_v - \zeta|M_{p+1} \geq \frac{1}{2}|f^{(p)}(x_v)| \quad (8.22)$$

and therefore

$$\frac{p|f^{(p)}(\zeta)|}{M_{p+1}} - |x_v - \zeta| \geq \frac{p}{2M_{p+1}}|f^{(p)}(x_v)| - \frac{1}{2} \frac{|f^{(p)}(x_v)|}{M_{p+1}} = \frac{p-1}{2M_{p+1}}|f^{(p)}(x_v)|.$$

Using this inequality in (8.20), we obtain

$$|x_{v+1} - \zeta| \leq \frac{2M_{p+1}|x_v - \zeta|^2}{(p^2-1)|f^{(p)}(x_v)|}. \quad (8.23)$$

Combining (8.23) with (8.21), it follows further that

$$|x_{v+1} - \zeta| \leq \frac{|x_v - \zeta|}{p^2-1} \leq \frac{|x_v - \zeta|}{3}. \quad (8.24)$$

9. The relation (8.23) is still not completely satisfactory, since M_{p+1} as well as $f^{(p)}(x_v)$ depend on v . It is, however, easy to deduce a formula in which these expressions are replaced with constants independent of v .

Suppose that relation (8.21) holds for $v = v_0$. Then we can write

$$\begin{aligned} |x_{v_0} - \zeta| &\leq \frac{\Delta}{2M^*}, \quad \Delta := |f^{(p)}(x_{v_0})|, \quad M^* := \text{Max } |f^{(p+1)}(u)| \\ &\quad (|u - \zeta| \leq |x_{v_0} - \zeta|). \end{aligned} \quad (8.25)$$

Then we will prove that formulas (8.21) and (8.22) hold for all $v \geq v_0$ and that we have generally, instead of (8.23),

$$|x_{v+1} - \zeta| \leq \frac{6M^*}{(p^2-1)\Delta} |x_v - \zeta|^2 \quad (v \geq v_0). \quad (8.26)$$

Indeed, it follows from (8.22) that $|f^{(p)}(\zeta)| \geq \Delta/2$. We therefore obtain, if $|u - \zeta| \leq |x_{v_0} - \zeta|/3$,

$$\begin{aligned} |f^{(p)}(u)| &\geq |f^{(p)}(\zeta)| - M^*|u - \zeta| \geq \frac{\Delta}{2} - \frac{1}{3}M^*|x_{v_0} - \zeta| \geq \frac{\Delta}{2} - \frac{\Delta}{6} = \frac{\Delta}{3}, \\ 3|f^{(p)}(u)| &\geq \Delta. \end{aligned} \quad (8.27)$$

It then follows that

$$\begin{aligned} |u - \zeta| &\leq \frac{1}{3}|x_{v_0} - \zeta| \leq \frac{\Delta}{6M^*} \leq \frac{|f^{(p)}(u)|}{2M^*}, \\ |u - \zeta| &\leq \frac{|f^{(p)}(u)|}{2M^*} \quad (|u - \zeta| \leq \frac{1}{3}|x_{v_0} - \zeta|). \end{aligned} \quad (8.28)$$

Since by (8.24) x_{v_0+1} can be taken as u in (8.28), we see that the relation (8.21) holds also for $v = v_0 + 1$. Indeed, M_{p+1} corresponding to $v_0 + 1$ is certainly $\leq M^*$, as the interval in which M^* is $\text{Max } |f^{(p)}(u)|$ contains the corresponding interval for x_{v_0+1} . We can therefore continue indefinitely and it follows that for $v \geq v_0$ we have generally

$$|x_{v+1} - \zeta| \leq \frac{1}{3}|x_v - \zeta| \quad (v \geq v_0) \quad (8.29)$$

and further formula (8.23). But then we obtain from (8.23) using (8.27)

$$|x_{v+1} - \zeta| \leq \frac{2M^*}{p^2 - 1} \frac{|x_v - \zeta|^2}{\frac{1}{3}\Delta}$$

and this is the assertion (8.26).

EVALUATION OF EXACT MULTIPLICITY

10. The discussion above assumes the knowledge that there exists a zero ζ of exact multiplicity p . This fact is not, however, usually known in advance and only has sense if the data are given with absolute precision. A numerical discussion can only assert the existence of a cluster of p zeros lying very close together.

It is therefore of interest to give a rule which allows us to obtain in such a case a suitable value of p from numerical computation.

Assume that the Newton-Raphson formula is applied to a zero ζ of exact multiplicity p twice, starting with an x already sufficiently close to ζ . If we denote the two values which were obtained in this way by x' and x'' , it follows from (8.7) with $p' = 1$ that, as $x \rightarrow \zeta$,

$$\begin{aligned} \frac{x' - \zeta}{x - \zeta} &= 1 - \frac{1}{p} + o(1), & \frac{x' - x}{x - \zeta} &= -\frac{1}{p} + o(1), \\ \frac{x'' - x'}{x' - \zeta} &= -\frac{1}{p} + o(1), & \frac{x'' - x'}{(1 - 1/p + o(1))(x - \zeta)} &= -\frac{1}{p} + o(1), \\ \frac{x'' - x'}{x' - x} &= 1 - \frac{1}{p} + o(1), \\ p &= \frac{x - x'}{x - 2x' + x''} + o(1). \end{aligned} \quad (8.30)$$

Using Gaussian brackets, we can therefore write

$$p = \left[\frac{1}{2} + \frac{x - x'}{x'' - 2x' + x} \right]. \quad (8.31)$$

11. We can therefore replace the rule (8.1) with the rule

$$x_{2v+1} := x_{2v} - \frac{f(x_{2v})}{f'(x_{2v})}, \quad (8.32)$$

$$u_{2v+1} := x_{2v+1} - \frac{f(x_{2v+1})}{f'(x_{2v+1})}, \quad (8.33)$$

$$x_{2v+2} := x_{2v+1} - \left[\frac{1}{2} + \frac{x_{2v} - x_{2v+1}}{u_{2v+1} - 2x_{2v+1} + x_{2v}} \right] (x_{2v+1} - u_{2v+1}), \quad (8.34)$$

where, however, if the value of the Gaussian brackets is ≤ 1 , Eq. (8.34) is to be replaced with

$$x_{2v+2} := u_{2v+1} - \frac{f(u_{2v+1})}{f'(u_{2v+1})}. \quad (8.35)$$

The efficiency index is then $\sqrt[3]{4}$. This is $< \sqrt{2}$ but > 1 . However, if p is = 1 we have again the classical Newton-Raphson case.

If $f^{(p+2)}(\zeta)$ also exists, it is easily seen that the term $o(1)$ in (8.30) is $O(|x - \zeta|)$. Therefore, in this case, if we use in Schröder's formula instead of p the quotient $(x - x')/(x - 2x' - x'')$, the error is of the order $O(|x - \zeta|^2)$ and the quadratic convergence remains. In this way a certain simplification of the procedure can be obtained. Then the efficiency index is again $\sqrt[3]{4}$. However, in this case the efficiency index remains generally the same also in the case $p = 1$.

The above rule also usually gives improved convergence if we have a *cluster of zeros*. In this case the value of p in (8.31) can change according as we come into the closer neighborhood of subclusters or finally single zeros.

Formula (8.26) gives an estimate *a priori* which will usually be too conservative. However, in this case the feedback techniques developed in Appendix Q can be used (see Appendix Q, Section 2.5).

9

Fourier Bounds for the
Newton–Raphson Iteration

1. Consider a function $f(x)$ in the interval $J_0: x_0 \leq x \leq y_0$. In this discussion, we assume that the so-called “Fourier conditions” are satisfied in J_0 , i.e., $\zeta \in J_0$, where $f(\zeta) = 0$, $f'(x)f''(x) \neq 0$ for $x \in J_0$, $f(x_0)f(y_0) < 0$, and $f(x_0)f''(x_0) > 0$ with continuous $f''(x)$. Assume without loss of generality

$$f(x_0) < 0 < f(y_0), \quad f'(x) > 0, \quad f''(x) < 0 \quad (x \in J_0) \quad (9.0)$$

(see Fig. 3).

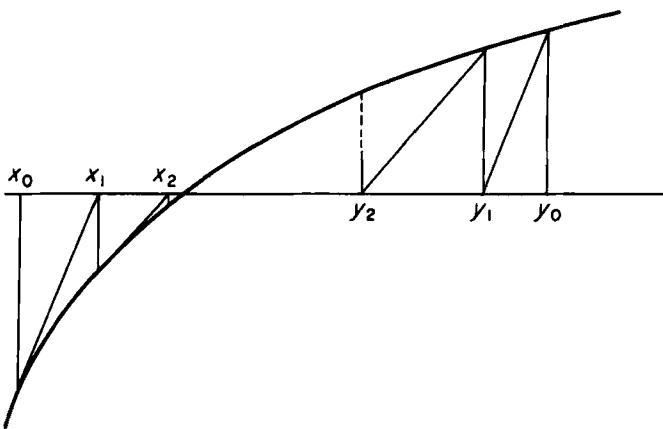


FIGURE 3

The sequence x_v ($v = 0, 1, \dots$) is defined by the Newton–Raphson formula (6.3). The sequence y_v , introduced by Fourier, is defined by

$$y_{v+1} = y_v - \frac{f(y_v)}{f'(x_v)} \quad (v = 0, 1, \dots). \quad (9.1)$$

Notice that y_{v+1} will be the point of intersection of the line through the point $[y_v, f(y_v)]$ with the slope $f'(x_v)$, and the x -axis.

It is evident from Fig. 3 that the x_v tend monotonically to a number $\xi \leq \zeta$

and the y_v to a number $\eta \geq \zeta$. Then we have from (6.3) and (9.1) as $v \rightarrow \infty$, since ζ is the unique root of $f(x)$ in J_0 ,

$$-\frac{f(\zeta)}{f'(\zeta)} = 0, \quad -\frac{f(\eta)}{f'(\zeta)} = 0, \quad \zeta = \eta = \zeta,$$

and we see that $x_v \uparrow \zeta$, $y_v \downarrow \zeta$, and in particular

$$(y_v - x_v) \downarrow 0 \quad (v \rightarrow \infty). \quad (9.2)$$

We shall now prove that, assuming $f''(x)$ as continuous in J_0 ,

$$\frac{y_{v+1} - x_{v+1}}{(y_v - x_v)^2} \rightarrow -\frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)} \quad (v \rightarrow \infty). \quad (9.3)$$

2. We introduce λ_v by

$$\lambda_v = \frac{y_v - \zeta}{\zeta - x_v}. \quad (9.4)$$

From (9.3) it will follow that the rule for diminishing of the distance between x_v and y_v is quadratic. The limitation of ζ by y_v would be sufficiently accurate if we had $\lambda_v \rightarrow 1$. We shall show, however, that $\lambda_v \rightarrow \infty$ and even $\lambda_{v+1}/\lambda_v^2 \rightarrow 1$. From (9.1) we have

$$\frac{y_{v+1} - \zeta}{y_v - \zeta} = 1 - \frac{1}{f'(x_v)} \frac{f(y_v) - f(\zeta)}{y_v - \zeta}, \quad (9.5)$$

$$f'(x_v) \frac{y_{v+1} - \zeta}{y_v - \zeta} = f'(x_v) - \frac{f(y_v) - f(\zeta)}{y_v - \zeta}. \quad (9.6)$$

We easily verify the identity

$$\int_0^1 f'[y_v + t(\zeta - y_v)] dt = \frac{f(\zeta) - f(y_v)}{\zeta - y_v}. \quad (9.7)$$

3. Since obviously

$$f'(x_v) = \int_0^1 f'(x_v) dt, \quad (9.8)$$

we have from (9.6) and (9.7)

$$f'(x_v) \frac{y_{v+1} - \zeta}{y_v - \zeta} = - \int_0^1 [f'(y_v + t(\zeta - y_v)) - f'(x_v)] dt. \quad (9.9)$$

On the other hand, we have obviously for the integrand

$$f'[y_v + t(\zeta - y_v)] - f'(x_v) = \int_0^{y_v - x_v + t(\zeta - y_v)} f''(x_v + w) dw. \quad (9.10)$$

We introduce here a new variable of integration u defined by

$$w = u[y_v - x_v + t(\zeta - y_v)]$$

and obtain

$$\begin{aligned} f'[y_v + t(\zeta - y_v)] - f'(x_v) \\ = [y_v - x_v + t(\zeta - y_v)] \int_0^1 f''\{x_v + u[y_v - x_v + t(\zeta - y_v)]\} du, \end{aligned} \quad (9.11)$$

$$\begin{aligned} f'(x_v) \frac{y_{v+1} - \zeta}{y_v - \zeta} \\ = - \int_0^1 [y_v - x_v + t(\zeta - y_v)] \int_0^1 f''\{x_v + u[y_v - x_v + t(\zeta - y_v)]\} du dt. \end{aligned} \quad (9.12)$$

From (9.4) we have

$$1 + \lambda_v = \frac{y_v - x_v}{\zeta - x_v}, \quad \frac{1 + \lambda_v}{\lambda_v} = \frac{y_v - x_v}{y_v - \zeta}. \quad (9.13)$$

Hence, dividing both sides of (9.12) by $y_v - \zeta$, we have

$$f'(x_v) \frac{y_{v+1} - \zeta}{(y_v - \zeta)^2} = - \int_0^1 \left(\frac{1 + \lambda_v}{\lambda_v} - t \right) \int_0^1 f''\{x_v + u[y_v - x_v + t(\zeta - y_v)]\} du dt. \quad (9.14)$$

4. Since $[(1 + \lambda_v)/\lambda_v] - t > 0$ for $0 \leq t \leq 1$, we can apply the generalized mean value theorem of the integral calculus to (9.14) and obtain

$$\frac{\zeta - y_{v+1}}{(\zeta - y_v)^2} = \frac{f''(\xi)}{f'(x_v)} \int_0^1 \int_0^1 \left(\frac{1 + \lambda_v}{\lambda_v} - t \right) du dt, \quad \xi \in (x_v, y_v) \quad (9.15)$$

$$\frac{\zeta - y_{v+1}}{(\zeta - y_v)^2} = \frac{f''(\xi)}{f'(x_v)} \left(\frac{1 + \lambda_v}{\lambda_v} - \frac{1}{2} \right) = \frac{1}{2} \frac{f''(\xi)}{f'(x_v)} \left(1 + \frac{2}{\lambda_v} \right). \quad (9.16)$$

By our assumptions

$$\kappa := -\frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)} \quad (9.17)$$

is $\neq 0$; then by (9.16)

$$\frac{\lambda_v}{2 + \lambda_v} \frac{\zeta - y_{v+1}}{(\zeta - y_v)^2} \rightarrow -\kappa \quad (v \rightarrow \infty). \quad (9.18)$$

From (6.9) we have

$$\frac{\zeta - x_{v+1}}{(\zeta - x_v)^2} \rightarrow \kappa \quad (v \rightarrow \infty), \quad (9.19)$$

and dividing (9.18) by (9.19) gives

$$\frac{-(\zeta - y_{v+1})/(\zeta - x_{v+1})}{[(\zeta - y_v)/(\zeta - x_v)]^2} \sim 1 + \frac{2}{\lambda_v} \quad (v \rightarrow \infty). \quad (9.20)$$

But here the left-hand expression is by (9.4) equal to $\lambda_{v+1}/\lambda_v^2$. Therefore, we get

$$\frac{\lambda_{v+1}/\lambda_v}{2 + \lambda_v} \rightarrow 1 \quad (v \rightarrow \infty). \quad (9.21)$$

By (9.21) we have in any case $\lim \lambda_{v+1}/\lambda_v \geq 2$ and therefore

$$\lambda_v \rightarrow \infty. \quad (9.22)$$

Hence from (9.21)

$$\frac{\lambda_{v+1}}{\lambda_v^2} \rightarrow 1. \quad (9.23)$$

We have now from (9.13)

$$\frac{y_{v+1} - x_{v+1}}{(y_v - x_v)^2} = \frac{1 + \lambda_{v+1}}{(1 + \lambda_v)^2} \frac{\zeta - x_{v+1}}{(\zeta - x_v)^2}$$

and by using (9.22), (9.23), and (9.19), we obtain finally

$$\frac{y_{v+1} - x_{v+1}}{(y_v - x_v)^2} \rightarrow \kappa = -\frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)}, \quad \text{Q.E.D.}$$

10

Dandelin Bounds for the Newton–Raphson Iteration

1. We assume, as in Chapter 9, replacing y_0 there by z_0 , that in the interval J_0 : $x_0 \leq x \leq z_0$, the Fourier conditions and in particular the conditions (9.0) are satisfied. Let x_0 and z_0 be the starting points, where the sequence x_v ($v = 0, 1, \dots$) is defined by the Newton–Raphson formula (6.3) and the sequence z_v is obtained by applying the *regula falsi* to x_{v-1} and z_{v-1} ; i.e.,

$$z_{v+1} = z_v - f(z_v) \frac{z_v - x_v}{f(z_v) - f(x_v)} \quad (v = 0, 1, \dots). \quad (10.1)$$

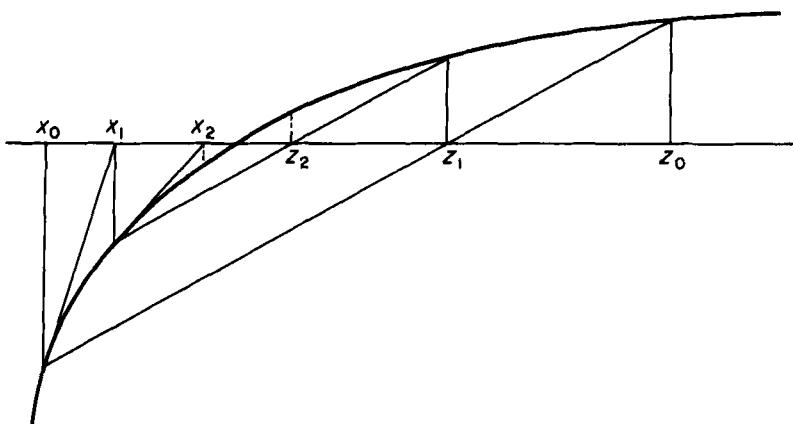


FIGURE 4

From Fig. 4 we see immediately that the z_v decrease monotonically to a certain number η and it now follows from (10.1) that

$$\frac{f(z_v)}{[f(z_v) - f(x_v)]/(z_v - x_v)}$$

tends to zero. But here the denominator remains between two positive bounds $[\text{Min } f'(x), \text{Max } f'(x)]$, and therefore $f(\eta) = 0$, $\eta = \zeta$, $z_v \downarrow \zeta$. x_v and z_v are *Dandelin's bounds* for ζ .

We introduce μ_v by

$$\mu_v = \frac{z_v - \zeta}{\zeta - x_v} \quad (v = 0, 1, \dots) \quad (10.2)$$

and put $J_1 = \langle x_0, y_0 \rangle$ for a fixed $y_0 > \zeta$. Then we will prove

Theorem 10.1. *If $f^{(3)}(x)$ is continuous in J_0 the sequence μ_v tends to a limit $\Lambda(z_0)$ with $v \rightarrow \infty$, uniformly for $\zeta < z_0 \leq y_0$, where $\Lambda(z_0)$ is a continuous function of z_0 , and decreases monotonically to 0 as z_0 tends to ζ .*

2. Proof. We have identically

$$\frac{z_{v+1} - \zeta}{z_v - \zeta} = 1 - \frac{f(z_v) - f(\zeta)}{z_v - \zeta} \frac{z_v - x_v}{f(z_v) - f(x_v)} \quad (10.3)$$

and can write

$$\begin{aligned} \frac{z_{v+1} - \zeta}{z_v - \zeta} &= \frac{N}{D}, \\ N &= \frac{f(x_v) - f(z_v)}{x_v - z_v} - \frac{f(\zeta) - f(z_v)}{\zeta - z_v}, \quad D = \frac{f(z_v) - f(x_v)}{z_v - x_v}. \end{aligned} \quad (10.4)$$

Formula (9.7) is an identity true for all $\zeta \neq y_v$; hence, we can write

$$\frac{f(a) - f(b)}{a - b} = \int_0^1 f'[b + t(a - b)] dt = \int_0^1 f'[a + t(b - a)] dt \quad (a \neq b). \quad (10.5)$$

Applying (10.5) to D and N in (10.4), we have

$$D = \int_0^1 f'[x_v + t(z_v - x_v)] dt, \quad (10.6)$$

$$N = \int_0^1 \{f'[z_v + t(x_v - z_v)] - f'[z_v + t(\zeta - z_v)]\} dt. \quad (10.7)$$

Now apply (10.5) to the integrand of (10.7); we obtain

$$N = -(\zeta - x_v) \int_0^1 \int_0^1 f''[z_v + t(x_v - z_v) + ut(\zeta - x_v)] t dt du. \quad (10.8)$$

3. Now put

$$\frac{z_{v+1} - \zeta}{(z_v - \zeta)(\zeta - x_v)} = \frac{N/(\zeta - x_v)}{D} = \frac{N^*}{D}, \quad N^* = \frac{N}{\zeta - x_v} \quad (10.9)$$

and consider

$$S = \int_0^1 \int_0^1 f''(x_v + uth) t du dt, \quad (10.10)$$

where $h = \zeta - x_v$. We introduce, instead of u , a new variable of integration, y , by $y = x_v + uth$, and have

$$hS = \int_0^1 \int_{x_v}^{x_v+th} f''(y) dy dt = \int_0^1 [f'(x_v + th) - f'(x_v)] dt,$$

$$h^2 S = \int_0^1 f'(x_v + th) h dt - hf'(x_v).$$

Since the integrand is $d[f(x_v + th)]/dt$, and $x_v + h = \zeta$, we have

$$\begin{aligned} h^2 S &= f(\zeta) - f(x_v) - hf'(x_v) = -[f(x_v) + hf'(x_v)], \\ -\frac{h^2 S}{f'(x_v)} &= \frac{f(x_v)}{f'(x_v)} + h, \\ -\frac{h^2 S}{f'(x_v)} &= x_v - x_{v+1} + \zeta - x_v = \zeta - x_{v+1}, \\ -\frac{S}{f'(x)} &= \frac{\zeta - x_{v+1}}{(\zeta - x_v)^2}. \end{aligned} \quad (10.11)$$

Dividing (10.9) by (10.11), we have from (10.2)

$$\frac{(z_{v+1} - \zeta)/(\zeta - x_{v+1})}{(z_v - \zeta)/(\zeta - x_v)} = -\frac{N^*}{S} \frac{f'(x_v)}{D} = \frac{\mu_{v+1}}{\mu_v}. \quad (10.12)$$

4. We will have to show that μ_v tends to a limit. First we state two well-known theorems on infinite products which we will need but shall not prove here.

Theorem A. If $\sum_{v=0}^{\infty} |c_v - 1|$ converges and no c_v is $= 0$, then $\prod_{v=0}^{\infty} c_v$ converges to a limit $\neq 0$.

Theorem B. If the factors c_v are continuous functions of a point and if $\sum_{v=0}^{\infty} |c_v - 1|$ converges uniformly for all points in a domain, then $\prod_{v=0}^{\infty} c_v$ converges to a continuous function in this domain.

We write now

$$\mu_n = \mu_0 \prod_{v=0}^{n-1} \frac{\mu_{v+1}}{\mu_v} = \mu_0 \frac{\prod_{v=0}^{n-1} Q_v}{\prod_{v=0}^{n-1} q_v}, \quad (10.13)$$

where by (10.12)

$$Q_v = -\frac{N^*}{S}, \quad q_v = \frac{D}{f'(x_v)} \quad (v = 0, 1, \dots). \quad (10.14)$$

5. Consider first

$$Q_v - 1 = \frac{-N^* - S}{S}. \quad (10.15)$$

From (10.8), (10.9), and (10.10) follows

$$-N^* - S$$

$$= \int_0^1 \int_0^1 \{f''[z_v + t(x_v - z_v) + ut(\zeta - x_v)] - f''[x_v + ut(\zeta - x_v)]\} t dt du.$$

Applying (10.5) to the integrand, we have

$$-N^* - S$$

$$= (z_v - x_v) \int_0^1 \int_0^1 \int_0^1 t(1-t) f^{(3)}[x_v + ut(\zeta - x_v) + w(1-t)(z_v - x_v)] dw dt du.$$

Since $t(1-t)$ is nonnegative in $\langle 0, 1 \rangle$, we apply the generalized mean value theorem of the integral calculus and obtain

$$-N^* - S = (z_v - x_v) f^{(3)}(\xi) \int_0^1 \int_0^1 \int_0^1 (1-t) t dt dw du, \quad \xi \in (x_v, z_v).$$

Integrating, we have

$$-N^* - S = \frac{f^{(3)}(\xi)}{6} (z_v - x_v), \quad \xi \in (x_v, z_v). \quad (10.16)$$

From (10.10) the generalized mean value theorem leads to

$$S = f''(\eta) \int_0^1 \int_0^1 t dt du, \quad S = \frac{f''(\eta)}{2}, \quad \eta \in (x_v, z_v). \quad (10.17)$$

From (10.15), (10.16), and (10.17) follows

$$Q_v - 1 = \frac{z_v - x_v}{3} \frac{f^{(3)}(\xi)}{f''(\eta)}, \quad \xi, \eta \in (x_v, z_v). \quad (10.18)$$

6. On the other hand, by (10.14) and (10.6),

$$q_v - 1 = \frac{D - f'(x_v)}{f'(x_v)}, \quad (10.19)$$

$$D - f'(x_v) = \int_0^1 \{f'[x_v + t(z_v - x_v)] - f'(x_v)\} dt. \quad (10.20)$$

Applying (10.5) to (10.20), we obtain by the generalized mean value theorem

$$\begin{aligned} D - f'(x_v) &= (z_v - x_v) \int_0^1 \int_0^1 f''[x_v + ut(z_v - x_v)] t dt du \\ &= (z_v - x_v) f''(\xi_1) \int_0^1 \int_0^1 t dt du, \quad \xi_1 \in (x_v, z_v). \end{aligned}$$

$$D - f'(x_v) = \frac{z_v - x_v}{2} f''(\xi_1), \quad \xi_1 \in (x_v, z_v),$$

and from (10.19)

$$q_v - 1 = \frac{z_v - x_v f''(\xi_1)}{2 f'(x_v)}, \quad \xi_1 \in (x_v, z_v). \quad (10.21)$$

7. We introduce M_k and m_k by

$$\max_{(x_0, y_0)} |f^{(k)}(x)| = M_k, \quad \min_{(x_0, y_0)} |f^{(k)}(x)| = m_k \quad (k = 1, 2, 3). \quad (10.22)$$

From (10.18) and (10.21) we have then

$$|Q_v - 1| \leq |z_v - x_v| K, \quad |q_v - 1| \leq |z_v - x_v| K, \quad (10.23)$$

where

$$K = \max\left(\frac{1}{3} \frac{M_3}{m_2}, \frac{1}{2} \frac{M_2}{m_1}\right). \quad (10.24)$$

8. But, on the other hand, from (6.9) and (9.17), we have

$$\frac{\zeta - x_{v+1}}{(\zeta - x_v)^2} \rightarrow \kappa, \quad \frac{\zeta - x_{v+1}}{\zeta - x_v} \sim \kappa(\zeta - x_v) \rightarrow 0.$$

Hence $\sum_{v=0}^{\infty} |\zeta - x_v|$ is convergent.

Similarly, from (3.11) and (9.17) we have

$$\frac{\zeta - z_{v+1}}{(\zeta - z_v)(\zeta - x_v)} \rightarrow \kappa, \quad \frac{\zeta - z_{v+1}}{\zeta - z_v} \sim \kappa(\zeta - x_v) \rightarrow 0,$$

and $\sum_{v=0}^{\infty} |\zeta - z_v|$ is convergent. From

$$z_v - x_v = z_v - \zeta + \zeta - x_v, \quad |z_v - x_v| \leq |\zeta - x_v| + |\zeta - z_v|$$

the convergence of $\sum_{v=0}^{\infty} |z_v - x_v|$ now follows.

9. We now show that $\sum_{v=0}^{\infty} |z_v - x_v|$ is uniformly convergent as a function of z_0 if, for a fixed y_0 , $\zeta < z_0 \leq y_0$. Notice that the points x_v do not depend on z_v . We will keep x_0 fixed and discuss what happens as $z_0 \downarrow \zeta$. Now all z_v are strictly increasing functions of z_0 . The initial value of z_0 is $y_0 = z_0'$; denote corresponding values of z_v by z_v' . Then

$$z_v \leq z_v' \quad (v = 0, 1, \dots),$$

and hence

$$|z_v - x_v| \leq |z_v' - x_v|.$$

Therefore by (10.23)

$$|Q_v - 1| \leq K|z_v' - x_v|, \quad |q_v - 1| \leq K|z_v' - x_v|.$$

This holds for $\zeta < z_0 \leq y_0$. Hence $\sum_{v=0}^{\infty} |Q_v - 1|$, $\sum_{v=0}^{\infty} |q_v - 1|$ converge uniformly if z_0 varies between ζ and y_0 . We see by Theorem B that both products $\prod_{v=0}^{\infty} Q_v$, and $\prod_{v=0}^{\infty} q_v$ converge uniformly for these z_0 , and from (10.13) it follows that $\lim_{v \rightarrow \infty} \mu_v$ exists uniformly for all z_0 between ζ and y_0 . Now the z_v are rational and continuous functions of z_0 . By (10.2) the μ_v are also continuous functions of z_0 and we see that $\Lambda(z_0)$ is a continuous function of z_0 for $\zeta < z_0 \leq y_0$.

10. If $z_0 = \zeta$, then each $z_v = \zeta$ ($v = 0, 1, \dots$) and it follows that for $z_0 \downarrow \zeta$ each $\mu_v(z_0)$ tends to 0. As each $\mu_v(z_0)$ is a monotonically increasing function of z_0 , so is $\Lambda(z_0)$ in the interval $\zeta < z_0 \leq y_0$. Therefore, as $\Lambda(z_0) \geq 0$, it tends to limit Λ_0 as $z_0 \downarrow \zeta$. On the other hand, for any positive ε there exists a $\mu_k(z_0)$ such that we have

$$|\Lambda(z_0) - \mu_k(z_0)| \leq \varepsilon \quad (\zeta < z_0 \leq y_0).$$

We have therefore

$$\Lambda(z_0) \leq \varepsilon + \mu_k(z_0)$$

and, as $z_0 \downarrow \zeta$, $\Lambda_0 \leq \varepsilon$. We see that $\Lambda_0 = 0$, i.e., $\Lambda(z_0) \downarrow 0$ ($z_0 \downarrow \zeta$). It follows that the error from the z_v side can be made an arbitrarily small part of the error from the x_v side if we start with a sufficiently close point z_0 .

11

Three Interpolation Points

INTERPOLATION BY LINEAR FRACTIONS

1. Let $f(x)$ be defined in J_x . Assume that we are given *three distinct interpolation points*

$$x_v \in J_x, \quad f(x_v) = f_v \quad (v = 0, 1, 2). \quad (11.1)$$

First we build up a function which interpolates $f(x)$ in these three points. If we choose a polynomial as our interpolating function, it will usually be quadratic; as such it has at least two zeros, and we will not know which of them to take (cf., however, the discussion in Chapter 17). We avoid this difficulty by considering instead a function which has only *one* simple zero, namely,

$$w = \frac{\alpha x + \beta}{\gamma x + \delta}, \quad \alpha\delta - \beta\gamma \neq 0. \quad (11.2)$$

Notice that in (11.2) we have three essential constants. We know from projective geometry that if w is related to x by (11.2), then the points x, x_0, x_1, x_2 have the same cross ratio as the points w, f_0, f_1, f_2 ; i.e.,

$$\frac{(x-x_1)/(x-x_0)}{(x_2-x_1)/(x_2-x_0)} = \frac{(w-f_1)/(w-f_0)}{(f_2-f_1)/(f_2-f_0)}. \quad (11.3)$$

We assume that $f_0 \neq f_1 \neq f_2$ and introduce Δ by

$$\Delta = \frac{(f_2-f_1)/(f_2-f_0)}{(x_2-x_1)/(x_2-x_0)} \quad (f_0 \neq f_1 \neq f_2, \quad x_0 \neq x_1 \neq x_2); \quad (11.4)$$

then we have from (11.3)

$$\frac{w-f_1}{w-f_0} = \Delta \frac{x-x_1}{x-x_0} \quad (f_0 \neq f_1 \neq f_2, \quad x_0 \neq x_1 \neq x_2). \quad (11.5)$$

In this way we obtain our interpolating function w when the three interpolation points are distinct.

TWO COINCIDENT INTERPOLATION POINTS

2. We consider from now on the case where two of our x_i are coincident; i.e., $x_2 = x_0$, $x_1 \neq x_0$. We assume further that we know f_0, f_1 and $f'_0 = f'(x_0)$ and that $f_0 \neq f_1$. We can get the corresponding interpolating function by going to the limit in Δ . Rewrite (11.4) as follows:

$$\Delta = \frac{(f_2 - f_1)/(x_2 - x_1)}{(f_2 - f_0)/(x_2 - x_0)}. \quad (11.6)$$

Then as $x_2 \rightarrow x_0$, we have as the limits of Δ and (11.5)

$$\Delta^* = \frac{1}{f'_0} \frac{f_1 - f_0}{x_1 - x_0} \quad (f_0 \neq f_1, \quad x_0 = x_2 \neq x_1), \quad (11.7)$$

$$\frac{w - f_1}{w - f_0} = \Delta^* \frac{x - x_1}{x - x_0} \quad (f_0 \neq f_1, \quad x_0 = x_2 \neq x_1). \quad (11.8)$$

It is immediately clear that $w(x)$ given by (11.8) satisfies the conditions $w(x_0) = f_0$, $w(x_1) = f_1$. Differentiating both sides of (11.8) with respect to x , we verify at once also that $w'(x_0) = f'_0$.

3. As in previous discussions, we shall use the inverse function to obtain estimates of error. Putting $w = f(x)$, let $x = \Phi(w)$ be the inverse function of $f(x)$. We have then obviously

$$\Phi(f_0) = x_0, \quad \Phi(f_1) = x_1, \quad \Phi'(f_0) = \frac{1}{f'_0}.$$

On the other hand, if we solve (11.8) with respect to x , we obtain a function $\varphi(w)$ given by

$$\frac{w - f_1}{w - f_0} = \Delta^* \frac{\varphi - x_1}{\varphi - x_0}. \quad (11.9)$$

Solving (11.9), we have

$$\varphi(w) = \frac{(x_0 - x_1 \Delta^*) w + x_1 f_0 \Delta^* - x_0 f_1}{(1 - \Delta^*) w + f_0 \Delta^* - f_1}. \quad (11.10)$$

It is easily verified that generally for constant $\alpha, \beta, \gamma, \delta$

$$\left(\frac{\alpha w + \beta}{\gamma w + \delta} \right)^{(3)} = 6\gamma^2 \frac{\alpha \delta - \beta \gamma}{(\gamma w + \delta)^4}. \quad (11.11)$$

From (11.10) and (11.11) it follows that

$$\varphi^{(3)}(w) = 6(1 - \Delta^*)^2 \frac{\Delta^*(f_1 - f_0)(x_1 - x_0)}{N^4}, \quad (11.12)$$

where

$$N = (1 - \Delta^*)w + f_0 \Delta^* - f_1. \quad (11.13)$$

ERROR ESTIMATES

4. Let $\varphi(w) \sqsupseteq \Phi(w)$. If we apply (1B.23), our interpolating function is $\varphi(w)$; consequently the factor $f^{(n)}(\xi) - \gamma^{(n)}(\xi)$ in (1B.23) must now be replaced by $[\Phi(\eta) - \varphi(\eta)]^{(n)}$. Then we have

$$\Phi(f) - \varphi(f) = \frac{1}{6}(f-f_0)^2(f-f_1)[\Phi^{(3)}(\eta) - \varphi^{(3)}(\eta)], \quad \eta \in (f, f_0, f_1). \quad (11.14)$$

For $f=0$, denoting by x_2 the new approximation $\varphi(0)$ to our root ζ , we have from (11.14)

$$\zeta - x_2 = -\frac{f_0^2 f_1}{6} [\Phi^{(3)}(\eta) - \varphi^{(3)}(\eta)], \quad \eta \in (0, f_0, f_1). \quad (11.15)$$

Taking $w=0$ in (11.10), we have

$$x_2 = \frac{x_1 f_0 \Delta^* - x_0 f_1}{f_0 \Delta^* - f_1}. \quad (11.16)$$

Assume $x_0, x_1 \rightarrow \zeta$. Then $f_0, f_1 \rightarrow 0$ and hence $\eta \rightarrow 0$. Under this assumption, if $f'(\zeta) \neq 0$, we have from (11.7)

$$\Delta^* = \frac{1}{f'_0} \frac{f_1 - f_0}{x_1 - x_0} \rightarrow \frac{f'(\zeta)}{f'(\zeta)} = 1. \quad (11.17)$$

We consider first

$$\frac{\Delta^* - 1}{x_1 - x_0} = \frac{f(x_1) - [f(x_0) + f'(x_0)(x_1 - x_0)]}{f'(x_0)(x_1 - x_0)^2}. \quad (11.18)$$

The bracketed expression in (11.18) gives the first two terms of Taylor's expansion of $f(x)$ around x_0 . Hence, the numerator of the right-hand side of (11.18) is equal to the remainder term $\frac{1}{2}(x_1 - x_0)^2 f''(\xi_1)$, $\xi_1 \in (x_0, x_1)$, and we have

$$\frac{\Delta^* - 1}{x_1 - x_0} = \frac{1}{2} \frac{f''(\xi_1)}{f'(x_0)}, \quad \xi_1 \in (x_0, x_1). \quad (11.19)$$

If $x_0, x_1 \rightarrow \zeta$, then

$$\frac{\Delta^* - 1}{x_1 - x_0} = \frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)}. \quad (11.20)$$

Consider now

$$\frac{f_0 \Delta^* - f_1}{x_1 - x_0} = \frac{f_0(\Delta^* - 1)}{x_1 - x_0} - \frac{f_1 - f_0}{x_1 - x_0}. \quad (11.21)$$

Under the assumptions $x_0 \rightarrow \zeta$, $x_1 \rightarrow \zeta$, $f_0 \rightarrow 0$ and using (11.20), we get

$$\frac{f_0 \Delta^* - f_1}{x_1 - x_0} \rightarrow -f'(\zeta), \quad (11.22)$$

and hence from (11.13), since $w = 0$,

$$\frac{N}{x_1 - x_0} \rightarrow -f'(\zeta). \quad (11.23)$$

5. From (11.12) we have, by (11.17), (11.20), and (11.23),

$$\varphi^{(3)}(\eta) = \frac{6\Delta^*[(\Delta^* - 1)/(x_1 - x_0)]^2(f_1 - f_0)/(x_1 - x_0)}{[N/(x_1 - x_0)]^4} \rightarrow \frac{3}{2} \frac{f''(\zeta)^2 f'(\zeta)}{f'(\zeta)^2 f''(\zeta)^4}, \quad (11.24)$$

$$\varphi^{(3)}(\eta) \rightarrow \frac{3}{2} \frac{f''(\zeta)^2}{f'(\zeta)^5}. \quad (11.25)$$

From (2.5) and the table in Chapter 2, Section 8, follows

$$\Phi^{(3)}(\eta) = 3f''^2 f'^{-5} - f^{(3)} f'^{-4}. \quad (11.26)$$

On the other hand, as $\eta \rightarrow 0$, $\Phi(\eta)$ tends to ζ . Hence

$$\Phi^{(3)}(\eta) \rightarrow 3f''(\zeta)^2 f'(\zeta)^{-5} - f^{(3)}(\zeta) f'(\zeta)^{-4}. \quad (11.27)$$

Further we have

$$f(x_0) = f(x_0) - f(\zeta) = f'(\xi_0)(x_0 - \zeta), \quad \xi_0 \in (x_0, \zeta), \quad (11.28)$$

$$f(x_1) = f'(\xi_1)(x_1 - \zeta), \quad \xi_1 \in (x_1, \zeta). \quad (11.29)$$

Substituting these results in (11.15), we obtain

$$\frac{\zeta - x_2}{(x_0 - \zeta)^2 (x_1 - \zeta)} \rightarrow \frac{1}{4} f''(\zeta)^2 f'(\zeta)^{-2} - \frac{1}{6} f^{(3)}(\zeta) f'(\zeta)^{-1}. \quad (11.30)$$

Equation (11.30) shows that we have here an *approximation of the third order*.

USE IN ITERATION PROCEDURE

6. Our procedure obviously can be considered as a combination of the *regula falsi* and the Newton–Raphson method. If we have already applied both

methods, the Newton-Raphson method in the point x_0 and the *regula falsi* in the points x_0 and x_1 , then the application of (11.16) requires no further horner and the results obtained using those methods *once* can be considerably improved.

On the other hand, if we want to use this method of approximation as an *iteration* method, we would have to use consecutively the following triplets of interpolation points:

$$(x_0, x_1, x_1), \quad (x_1, x_2, x_2), \quad (x_2, x_3, x_3), \dots,$$

where in the first triplet x_1 is used *twice* and x_0 only *once*.

7. If we put

$$\ln \frac{1}{|x_\mu - \zeta|} = y_\mu \quad (11.31)$$

we obtain then from (11.30), observing that the values of x_0 and x_1 must be interchanged, the relation

$$y_{\mu+2} = y_\mu + 2y_{\mu+1} + k_\mu, \quad (11.32)$$

where k_μ are bounded, if the expression to the right in (11.30) is $\neq 0$. But then we will prove in the following chapter that if $x_\mu \rightarrow \zeta$, $y_\mu \rightarrow \infty$, we have

$$\frac{y_{\mu+1}}{y_\mu} \rightarrow 1 + \sqrt{2} = 2.414\dots, \quad (11.33)$$

and since we spend two horners at each step, the efficiency index of this iteration is $\sqrt{2.414\dots} = 1.55\dots$. If we compare this with the efficiency indices of the *regula falsi*, 1.618..., and of the Newton-Raphson method, 1.414..., we see that this new iteration method is better than the Newton-Raphson iteration method but not as good as the *regula falsi* used as an iteration method.

Example. If we apply the above method to the equation $x^2 - 2x - 1 = 0$ discussed in Section 17 of Chapter 3, starting with the triple (x_0, x_1, x_1) , $x_0 = 3$, $x_1 = 2$, we obtain the set of values[†] given in the accompanying table.

v	$10^{-2}(x_v - 2.09)$						$ x_v - \zeta $	y_{v+1}/y_v
2	0.026045	1977	4011	2994	3		$1.4937 \cdot 10^{-3}$	2.759
3	0.024551	4320	3811	0802	6		$4.95 \cdot 10^{-8}$	2.585
4	0.024551	4815	4232	6592	3		$9 \cdot 10^{-19}$	2.47
5	0.024551	4815	4232	6591	4		—	—

[†] Computed by Mr. Allen Reiter in the Mathematics Research Center of U.S.A., Madison, Wisconsin.

12

Linear Difference Equations

INHOMOGENEOUS AND HOMOGENEOUS DIFFERENCE EQUATIONS

1. In this chapter we shall study the so-called *linear difference equations of the order n with constant coefficients*, i.e., a sequence of recurrence formulas

$$a_0 z_{\mu+n} + a_1 z_{\mu+n-1} + \cdots + a_n z_\mu = k_{\mu+n} \quad (a_0 = 1; \mu = 0, 1, \dots), \quad (12.1)$$

where the a_0, \dots, a_n are fixed constants, while $k_{\mu+n}$ is a given sequence. The problem is then usually to determine all z_μ ($\mu \geq 0$) if the initial values z_0, z_1, \dots, z_{n-1} are given. This is obviously always possible step by step.

If all $k_{\mu+n} = 0$ we have the *homogeneous difference equation*, written in y :

$$a_0 y_{\mu+n} + a_1 y_{\mu+n-1} + \cdots + a_n y_\mu = 0 \quad (a_0 = 1; \mu = 0, 1, \dots). \quad (12.2)$$

2. The general solution of Eq. (12.1) in its classical form depends on the following polynomial, the so-called *characteristic polynomial* of (12.1) and (12.2):

$$\varphi(x) = x^n + a_1 x^{n-1} + \cdots + a_n. \quad (12.3)$$

We consider simultaneously with $\varphi(x)$ the polynomial

$$\psi(x) = x^n \varphi(1/x) = 1 + a_1 x + \cdots + a_n x^n. \quad (12.4)$$

If the constants k_v are given, we can consider the power series

$$K(x) = \sum_{v=n}^{\infty} k_v x^v \quad (12.5)$$

as given, and the problem of determination of the z_v by (12.1) can be considered as the problem of determination of the corresponding *generating series*

$$Z(x) = \sum_{v=0}^{\infty} z_v x^v. \quad (12.6)$$

3. If we then multiply $Z(x)$ by $\psi(x)$, the coefficients of x^v in the product are for $v \geq n$, in virtue of (12.1), just the corresponding k_v . Therefore, (12.1)

is equivalent to the identity

$$\psi(x)Z(x) = K(x) + P_{n-1}(x), \quad (12.7)$$

where $P_{n-1}(x)$ is a polynomial of degree $n-1$, which can be chosen arbitrarily.

Solving (12.7) with respect to $Z(x)$, we have for all x with $\psi(x) \neq 0$

$$Z(x) = \frac{K(x)}{\psi(x)} + \frac{P_{n-1}(x)}{\psi(x)}, \quad (12.8)$$

and we see that the z_v are obtained in developing the right-hand expression in (12.8) in ascending powers of x .

GENERAL SOLUTION OF THE HOMOGENEOUS EQUATION

4. In the case of the homogeneous equation (12.2), we have $K(x) = 0$ and the development of $Z(x)$ is obtained, e.g., by decomposing $P_{n-1}(x)/\psi(x)$ in partial fractions.

We denote the n zeros of $\varphi(x)$ by u_1, \dots, u_n and order them in such a way that

$$|u_1| \geq |u_2| \geq \dots \geq |u_n|.$$

Assume first that all u_v are different. Then the right-hand expression in (12.8) in the case $K(x) = 0$ is

$$\sum_{\kappa=1}^n \frac{\gamma_\kappa}{1-u_\kappa x},$$

and we obtain

$$z_\mu = \sum_{\kappa=1}^n \gamma_\kappa u_\kappa^\mu \quad (\mu = 0, 1, \dots). \quad (12.9)$$

Here the n constants γ_κ are the n “integration constants” of (12.2) and can be chosen arbitrarily.

5. If $\varphi(x)$ has multiple zeros and, e.g., u is a zero of $\varphi(x)$ with a multiplicity $m > 1$, then the corresponding part of the decomposition of $P_{n-1}(x)/\psi(x)$ is given by

$$\sum_{\kappa=1}^m \frac{\gamma_\kappa}{(1-ux)^\kappa}.$$

Developing this, we obtain

$$\sum_{v=0}^{\infty} u^v (b_0 + b_1 v + \dots + b_{m-1} v^{m-1}) x^v,$$

where coefficients b_0, \dots, b_{m-1} can assume arbitrary values if γ_κ are chosen conveniently. Therefore, in this case, b_0, \dots, b_{m-1} can be considered as constants of integration. Proceeding in the same manner for all zeros of $\varphi(x)$, we obtain finally, if v_1, \dots, v_k are the *different* zeros of $\varphi(x)$ with the corresponding multiplicities m_1, \dots, m_k , as the complete solution of (12.2) the expression

$$z_\mu = \sum_{\kappa=1}^k v_\kappa^\mu Q_\kappa(\mu), \quad (12.10)$$

where $Q_\kappa(x)$ ($\kappa = 1, \dots, k$) are arbitrary polynomials in x of degree $m_\kappa - 1$.

LEMMA ON DIVISION OF POWER SERIES

6. We shall now use (12.8) in order to discuss the asymptotic behavior of the z_v in some important cases. We first prove the following:

Lemma 12.1. *Suppose that for positive constants s and γ the coefficients of a general power series*

$$K(x) = \sum_{v=0}^{\infty} k_v x^v$$

satisfy the condition $|k_v| \leq \gamma s^v$ ($v = 0, 1, \dots$). Then if $0 < |\xi| < s$, we have for the coefficients of

$$\frac{K(x)}{1-\xi x} = \sum_{v=0}^{\infty} l_v x^v,$$

$|l_v| \leq \gamma_1 s^v$ ($v = 0, 1, \dots$) with a convenient $\gamma_1 = \gamma s / |s - |\xi||$. The same is true if $|\xi| > s$, assuming that $K(1/\xi) = 0$.

Proof. We can assume without loss of generality $s = 1$, since otherwise we could replace x by x/s . If $|\xi| < 1$, we have

$$l_v = k_v + k_{v-1}\xi + \dots + k_0\xi^v,$$

$$|l_v| \leq \gamma(1 + |\xi| + \dots + |\xi|^v) < \frac{\gamma}{1 - |\xi|}.$$

If $|\xi| > 1$ and $K(1/\xi) = 0$, we have

$$l_v = \xi^v \left(k_0 + \frac{k_1}{\xi} + \dots + \frac{k_v}{\xi^v} \right) = -\xi^v \left(\frac{k_{v+1}}{\xi^{v+1}} + \frac{k_{v+2}}{\xi^{v+2}} + \dots \right),$$

$$|l_v| \leq \gamma \left(\frac{1}{|\xi|} + \frac{1}{|\xi|^2} + \dots \right) = \frac{\gamma}{|\xi| - 1}.$$

7. Corollary 1. Suppose that $K(x)$ satisfies the condition of Lemma 12.1 and let ξ_1, \dots, ξ_n be n numbers such that each $|\xi_v|$ is $\neq 0, \neq s$. Suppose further that for each ξ_v with $|\xi_v| > s$, $K(x)$ has $1/\xi_v$ as a zero, the multiplicity of which is at least equal to the number of those ξ_v which are $= \xi_v$. Then in

$$\frac{K(x)}{\prod_{v=1}^n (1 - \xi_v x)} = \sum_{v=0}^{\infty} t_v x^v$$

we have $t_v = O(s^v)$.

Corollary 2. If $|\xi| < s$, then for a positive integer m , the development of $1/[(1 - \xi x)(1 - sx)^m]$ is majorized by the development of $\gamma_1/(1 - sx)^m$ for $\gamma_1 = s/(s - |\xi|)$:

$$\frac{1}{(1 - \xi x)(1 - sx)^m} \ll \frac{s/(s - |\xi|)}{(1 - sx)^m} \quad (m = 1, 2, \dots).^{\dagger}$$

Indeed for $m = 1$ the assertion can be written in the form

$$\frac{(1 - sx)^{-1}}{1 - \xi x} \ll \gamma_1 (1 - sx)^{-1}$$

and this follows from the first assertion of the lemma for $\gamma = 1$. On the other hand, such a majoration relation remains obviously true if multiplied on both sides by the same power series with positive coefficients. And in multiplying our relation on both sides by the development of $1/(1 - sx)^{m-1}$, which has positive coefficients, we obtain the assertion of Corollary 2.

ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF (12.1)

8. We are now going to prove:

Theorem 12.1. Suppose that we have in (12.8) for the above zeros u_1 and u_2 of $\varphi(x)$

$$|u_1| > 1 > |u_2| \tag{12.11}$$

and for a positive constant $s < |u_1|$

$$k_v = O(s^v) \quad (v \rightarrow \infty). \tag{12.12}$$

Then we have for a convenient constant α

$$z_v/u_1^v \rightarrow \alpha \quad (v \rightarrow \infty). \tag{12.13}$$

[†] The symbol \ll is the symbol of majorization introduced by Henri Poincaré. Its meaning is that for each power x^v the modulus of the coefficient of x^v to the left does not exceed the corresponding coefficient to the right.

Further we have, if $s > |u_2|$,

$$z_v = \alpha u_1^v + O(s^v) \quad (v \rightarrow \infty). \quad (12.14)$$

If $|u_1| > s = |u_2|$ and m is the maximal multiplicity of the zeros of $\varphi(x)$ with modulus $= |u_2|$, we have

$$z_v = \alpha u_1^v + O(v^{m-1} u_2^v) \quad (v \rightarrow \infty). \quad (12.15)$$

If $s < |u_2|$, we have, m having the same meaning as above,

$$z_v = \alpha u_1^v + O(v^{m-1} u_2^v) \quad (v \rightarrow \infty). \quad (12.16)$$

9. Proof. Assume that we have altogether $k-1$ roots u_k with the modulus $= |u_2|$:

$$|u_2| = \dots = |u_k| > |u_{k+1}|.$$

Then we can assume without loss of generality that we have $s > |u_{k+1}|$. Applying Corollary 1 of Lemma 12.1 to (12.4), we have

$$Z(x) = \frac{K_1(x)}{\prod_{k=1}^k (1 - u_k x)}, \quad (12.17)$$

$$K_1(x) = \sum_{v=0}^{\infty} k_v' x^v,$$

with $k_v' = O(s^v)$ ($v \rightarrow \infty$). Since $K_1(x)$ has the radius of convergence $\geq 1/s > 1/|u_1|$, $Z(x)$ has in $1/u_1$ at the most a single pole and can therefore be written in the form

$$Z(x) = \frac{\alpha}{1 - u_1 x} + P(x), \quad (12.18)$$

where $P(x)$ has the radius of convergence $> 1/|u_1|$. From (12.17) and (12.18) we obtain then

$$P(x) = \frac{K_2(x)}{(1 - u_1 x) \prod_{k=2}^k (1 - u_k x)}, \quad (12.19)$$

$$K_2(x) = \sum_{v=0}^{\infty} k_v'' x^v = K_1(x) - \alpha \prod_{k=2}^k (1 - u_k x), \quad (12.20)$$

where we have again $k_v'' = O(s^v)$ and obviously $K_2(1/u_1) = 0$. Therefore, by Lemma 12.1 for $\xi = u_1$, the v th coefficient of

$$\frac{K_2(x)}{1 - u_1 x} = K_3(x) = \sum_{v=0}^{\infty} k_v^{(3)} x^v$$

is $O(s^v)$, and we have

$$Z(x) - \frac{\alpha}{1-u_1 x} = \frac{K_3(x)}{\prod_{k=2}^t (1-u_k x)}. \quad (12.21)$$

10. If now $s > |u_2|$, we can apply here to the right-hand quotient Corollary 1 to the above lemma and see that the v th coefficient of the development of this quotient is $O(s^v)$. Equations (12.13) and (12.14) follow from this immediately.

Assume now that $s \leq |u_2|$. Denote the different ones among u_2, \dots, u_t by $u', \dots, u^{(t)}$ and let m_τ be generally the multiplicity of $u^{(\tau)}$, where obviously $m = \max m_\tau$. Then we have, decomposing in partial fractions,

$$\frac{1}{\prod_{k=2}^t (1-u_k x)} = \sum_{\tau=1}^t \frac{p_\tau(x)}{(1-u^{(\tau)}x)^{m_\tau}},$$

where each $p_\tau(x)$ is a polynomial of degree $m_\tau - 1$. Introducing this in (12.21) we obtain

$$Z(x) - \frac{\alpha}{1-u_1 x} = \sum_{\tau=1}^t \frac{K^{(\tau)}(x)}{(1-u^{(\tau)}x)^{m_\tau}}. \quad (12.22)$$

Here we have for $\tau = 1, \dots, t$

$$K^{(\tau)}(x) = \sum_{v=0}^{\infty} k_{\tau,v} x^v = K_3(x) p_\tau(x)$$

and therefore again

$$k_{\tau,v} = O(s^v) \quad (v \rightarrow \infty; \quad \tau = 1, \dots, t).$$

Since therefore each $K^{(\tau)}(x)$ has a majorant in the form $\gamma^{(\tau)}(1-sx)^{-1}$, the development of the right-hand side in (12.22) is majorized by

$$\frac{\gamma}{(1-sx)(1-|u_2|x)^m}$$

for a convenient γ :

$$Z(x) - \frac{\alpha}{1-u_1 x} \ll \frac{\gamma}{(1-sx)(1-|u_2|x)^m}. \quad (12.23)$$

If now $s = |u_2|$, the coefficient of x^v in the development to the right in (12.23) is $\leq \gamma \binom{m+v}{v} |u_2|^v$, and we have (12.13), (12.15).

If finally $s < |u_2|$, it follows from Corollary 2 to the above lemma (by replacing there ξ by s and s by $|u_2|$) that the expression to the right in (12.23) is majorized by $\gamma_1 (1-|u_2|x)^{-m}$, and here the coefficient of x^v is $O(v^{m-1} |u_2|^v)$ with $v \rightarrow \infty$; we have (12.13) and (12.16). Our theorem is proved.

ASYMPTOTIC BEHAVIOR OF ERRORS IN THE REGULA FALSI ITERATION

11. We now apply our result to the situation discussed in Section 10 of Chapter 3 in order to prove (3.19). Introducing the notations

$$\delta_v = |x_v - \zeta|, \quad y_v = \ln \frac{1}{\delta_v}, \quad (12.24)$$

we obtain from (3.14), replacing x_0 by x_{v-1} there,

$$y_{v+1} - y_v - y_{v-1} = \ln \left| \frac{2f'(\xi)^3}{f''(\xi)f'(\xi_1)f'(\xi_2)} \right|. \quad (12.25)$$

Here ξ, ξ_1, ξ_2 lie in the smallest interval containing ζ, x_{v-1}, x_v , and we have therefore

$$\max(|\xi - \zeta|, |\xi_1 - \zeta|, |\xi_2 - \zeta|) \leq \max(\delta_{v-1}, \delta_v). \quad (12.26)$$

The expression to the right in (12.25) converges to

$$\delta = \ln \left| \frac{2f'(\zeta)}{f''(\zeta)} \right|; \quad (12.27)$$

we assume, of course that $f'(\zeta)f''(\zeta) \neq 0$ and that the x_v are already in the interval around ζ in which $f'(x)$ and $f''(x)$ do not change their signs. If we denote, therefore, the right-hand expression in (12.25) by $\delta + k_{v+1}$, we have

$$k_{v+1} = 3 \ln \frac{f'(\xi)}{f'(\zeta)} - \ln \frac{f'(\xi_1)}{f'(\zeta)} - \ln \frac{f'(\xi_2)}{f'(\zeta)} - \ln \frac{f''(\xi)}{f''(\zeta)}.$$

12. Denote the maximum moduli of $f''(x)$ and $f'''(x)$ in the considered interval by M_2, M_3 and the minimum moduli of $f'(x), f''(x)$ by m_1, m_2 . Then we have by the mean value theorem and by (12.26)

$$\begin{aligned} \max \left(\left| \ln \frac{f'(\xi)}{f'(\zeta)} \right|, \left| \ln \frac{f'(\xi_1)}{f'(\zeta)} \right|, \left| \ln \frac{f'(\xi_2)}{f'(\zeta)} \right| \right) &\leq \frac{M_2}{m_1} \max(\delta_{v-1}, \delta_v), \\ \left| \ln \frac{f''(\xi)}{f''(\zeta)} \right| &\leq \frac{M_3}{m_2} \max(\delta_{v-1}, \delta_v). \end{aligned}$$

It follows finally that

$$k_{v+1} = O(\delta_{v-1} + \delta_v).$$

† We have to apply here the mean value theorem to either $\ln f'(x)$ or $\ln [-f'(x)]$ and correspondingly in the case of $f''(x)$.

On the other hand, from the discussions in Sections 9 and 10 of Chapter 3 it follows that

$$\delta_v = O(d^{t_1^v + 1/\sqrt{5}}) \quad (12.28)$$

where $0 < d < 1$ and $t_1 > 1$. Then we have obviously $\sqrt{\delta_v} \rightarrow 0$, and we obtain finally from (12.25), putting $y_v = v_v - \delta_v$,

$$v_{v+1} - v_v - v_{v-1} = k_{v+1}, \quad (12.29)$$

where $\sum_{v=0}^{\infty} k_v x^v$ is an entire function. But then the conditions of Theorem 12.1 are satisfied, since both roots of $x^2 - x - 1$ are t_1, t_2 with $t_1 > 1 > t_2 > 0$. We obtain, therefore,

$$y_v + \delta - \alpha t_1^v = v_v - \alpha t_1^v = O(t_2^v) \quad (12.30)$$

where, since with $x_v \rightarrow \zeta$, $y_v \rightarrow \infty$, α is a positive constant. It follows further by (12.24), (12.27), and (12.29),

$$|x_v - \zeta| = \left| \frac{2f'(\zeta)}{f''(\zeta)} \right| \exp(-\alpha t_1^v) [1 + O(t_2^v)]. \quad (12.31)$$

We raise this formula to the power t_1 , rewrite (12.31) for $v+1$, and divide; then we obtain

$$\frac{|x_{v+1} - \zeta|}{|x_v - \zeta|^{t_1}} = \left| \frac{2f'(\zeta)}{f''(\zeta)} \right|^{t_1} + O(t_2^v), \quad (12.32)$$

which is (3.19).

Similarly, in the problem of Chapter 11 we have the difference equation (11.32) the characteristic polynomial of which, $x^2 - 2x - 1$, has the two roots, $1 + \sqrt{2} > 1$ and $1 - \sqrt{2}$ with the modulus $\sqrt{2} - 1 < 1$. We have therefore, by (12.13) and (12.14) with $s = 1$, for a convenient real α as $\mu \rightarrow \infty$:

$$y_\mu = \alpha(1 + \sqrt{2})^\mu + O(1),$$

where α is certainly > 0 since the O -term remains bounded while y_μ tends to ∞ . From this, Eq. (11.33) follows immediately.

A THEOREM ON ROOTS OF CERTAIN EQUATIONS

13. In the applications of the above theory the following theorem due to Cauchy is often useful:

If in the equation

$$x^n - b_1 x^{n-1} - \cdots - b_n = 0 \quad (12.33)$$

all b_v ($v = 1, 2, \dots, n$) are ≥ 0 , but not all vanish, (12.33) has a unique simple root $p > 0$, and all other roots of (12.33) have moduli $\leq p$.

As a matter of fact, a more precise statement can be made, which is of some importance in the applications.

Theorem 12.2. *If in Eq. (12.33) the b_v are ≥ 0 and the indices of the b_v , which are > 0 have the common greatest divisor 1, then (12.33) has a unique simple positive root p and the moduli of all other roots of (12.33) are $< p$.*

Proof. Let

$$b_{k_1}, b_{k_2}, \dots, b_{k_m}, \quad k_1 < k_2 < \dots < k_m \leq n,$$

be all coefficients of (12.33) which are *positive*. By the assumption about the k_v there exist m integers s_1, s_2, \dots, s_m such that

$$s_1 k_1 + s_2 k_2 + \dots + s_m k_m = 1. \quad (12.34)$$

Equation (12.33) can be written in the form

$$F(x) \equiv \frac{b_{k_1}}{x^{k_1}} + \frac{b_{k_2}}{x^{k_2}} + \dots + \frac{b_{k_m}}{x^{k_m}} - 1 = 0. \quad (12.35)$$

$F(x)$ is for positive x strictly monotonically decreasing from ∞ to -1 and vanishes therefore for exactly one value $x = p > 0$. The derivative of $F(x)$ at the point p is

$$F'(p) = -k_1 \frac{b_{k_1}}{p^{k_1+1}} - k_2 \frac{b_{k_2}}{p^{k_2+1}} - \dots - k_m \frac{b_{k_m}}{p^{k_m+1}} < 0,$$

and we see that p is a *simple* root of (12.33).

Let $x \neq p$ be another root of $F(x)$. Then we have, putting $|x| = q$,

$$1 = \frac{b_{k_1}}{x^{k_1}} + \dots + \frac{b_{k_m}}{x^{k_m}} \leq \frac{b_{k_1}}{q^{k_1}} + \frac{b_{k_2}}{q^{k_2}} + \dots + \frac{b_{k_m}}{q^{k_m}}, \quad (12.36)$$

and we see that $F(q) \geq 0$. If we now have $F(q) > 0$, it follows that $q < p$. If we have $F(q) = 0$, then we have the sign of equality in (12.36) and all quotients b_{k_μ}/x^{k_μ} must be positive. By (12.34)

$$\left(\frac{b_{k_1}}{x^{k_1}} \right)^{s_1} \left(\frac{b_{k_2}}{x^{k_2}} \right)^{s_2} \dots \left(\frac{b_{k_m}}{x^{k_m}} \right)^{s_m} = \frac{b_{k_1}^{s_1} b_{k_2}^{s_2} \dots b_{k_m}^{s_m}}{x}$$

is also positive, and therefore $x > 0$. We have then $x = p$, contrary to our assumption. Theorem 12.2 is now proved.[†]

[†] The condition imposed in Theorem 12.2 on the indices of the nonvanishing b_v is obviously essential. For if the left-hand polynomial in (12.33) is a polynomial in x^k , $k > 1$, then there are k roots of (12.33) with the modulus p .

14. We will need later the following theorem:

Theorem 12.3. Consider the equation

$$x^n - \sum_{\kappa=0}^{n-1} p_\kappa x^\kappa = 0, \quad p_\kappa \geq 0 \quad (\kappa = 0, \dots, n-1), \quad (12.37)$$

with the positive root σ . Consider an infinite sequence u_v ($v = 1, 2, \dots$) satisfying the difference inequality

$$u_{v+n} - \sum_{\kappa=0}^{n-1} p_\kappa u_{v+\kappa} \geq 0 \quad (v = 1, 2, \dots) \quad (12.38)$$

and such that u_1, \dots, u_n are positive. Then we have

$$u_v \geq \alpha \sigma^v \quad (v = 1, 2, \dots), \quad (12.39)$$

where $\alpha > 0$ is given by

$$\alpha = \min_{\kappa=1, \dots, n} u_\kappa / \sigma^\kappa. \quad (12.40)$$

Proof. Relation (12.39) is obviously true, by virtue of (12.40), for $v = 1, 2, \dots, n$. Assume that (12.39) is true for $v = 1, 2, \dots, n+N-1$ with an $N \geq 1$. Then we have from (12.38)

$$u_{n+N} \geq \sum_{\kappa=0}^{n-1} p_\kappa u_{N+\kappa} \geq \alpha \sum_{\kappa=0}^{n-1} p_\kappa \sigma^{N+\kappa} = \alpha \sigma^N \sum_{\kappa=0}^{n-1} p_\kappa \sigma^\kappa.$$

But the last right-hand sum has the value σ^n since σ satisfies the equation (12.37). We see that (12.39) is also satisfied for $v = n+N$ and therefore for all $v \geq 1$. Theorem 12.3 is proved.

13

n Distinct Points of Interpolation

ERROR ESTIMATES

1. Let $w = f(x)$ be defined in J_x and $f'(\zeta) = 0$, $\zeta \in J_x$, where $f'(x) \neq 0$ for all x in J_x . Assume that we are given n distinct interpolation points x_v ($v = 1, \dots, n$) in J_x in a sufficiently close neighborhood of ζ , $f(x_v) = y_v$ ($v = 1, \dots, n$). We write $x = \Phi(w)$ for the inverse function of $w = f(x)$ and approximate Φ by the Lagrangian polynomial of degree $n-1$, $T_{n-1}(w) = T(w)$ with $T(y_v) = x_v$ ($v = 1, \dots, n$).

Let $T(0) = x_{n+1}$ and assume that $f^{(n)}(x)$ is continuous in J_x . We then apply (1B.21) by replacing there t_v with $y_v, f(t_v)$ with x_v, x with 0, and $F(x)$ with

$$W(w) = \prod_{v=1}^n (w - y_v).$$

Then

$$x_{n+1} = T(0) = (-1)^{n-1} \prod_{v=1}^n y_v \sum_{v=1}^n \frac{x_v}{y_v} \frac{1}{W'(y_v)}. \quad (13.1)$$

We now apply (2.2) for $x = 0$ in replacing there φ by Φ and the x_v by y_v . We have then, putting

$$S = \frac{1}{n!} \Phi^{(n)}(\eta), \quad \eta \in (0, y_1, \dots, y_n), \quad (13.2)$$

$$\zeta - x_{n+1} = (-1)^n S \prod_{v=1}^n y_v. \quad (13.3)$$

Since

$$y_v = f(x_v) = (x_v - \zeta) f'(\xi_v), \quad \xi_v \in (x_v, \zeta),$$

we have from (13.3)

$$\begin{aligned} \zeta - x_{n+1} &= S \prod_{v=1}^n (\zeta - x_v) \prod_{v=1}^n f'(\xi_v), \\ \frac{\zeta - x_{n+1}}{\prod_{v=1}^n (\zeta - x_v)} &= S \prod_{v=1}^n f'(\xi_v), \end{aligned} \quad (13.4)$$

and therefore in particular, if the x_v ($v = 1, \dots, n$) tend to ζ ,

$$\frac{\zeta - x_{n+1}}{\prod_{v=1}^n (\zeta - x_v)} \rightarrow \frac{1}{n!} \Phi^{(n)}(0) f'(\zeta)^n \quad (x_v \rightarrow \zeta, \quad v = 1, \dots, n). \quad (13.5)$$

2. Let k be an upper bound of the modulus of the right-hand side of (13.4). Then

$$|\zeta - x_{n+1}| \leq k \prod_{v=1}^n |\zeta - x_v|.$$

Put $K = k^{1/(n-1)}$. Then we have further

$$K|\zeta - x_{n+1}| \leq \prod_{v=1}^n (K|\zeta - x_v|). \quad (13.6)$$

Taking all x_v ($v = 1, \dots, n$) sufficiently close to ζ , we make all factors on the right-hand side of (13.6) < 1 . We then introduce ε_v by

$$K|\zeta - x_v| = u^{\varepsilon_v} \quad (v \geq 1), \quad u = \text{Max}(K|\zeta - x_1|, K|\zeta - x_2|), \quad (13.7)$$

where $0 < u < 1$. From (13.6) and (13.7) we have

$$u^{\varepsilon_{n+1}} \leq u^{\varepsilon_1 + \dots + \varepsilon_n}. \quad (13.8)$$

ITERATION WITH n DISTINCT INTERPOLATION POINTS

3. Since in (13.7) $u < 1$, we have

$$\varepsilon_{n+1} \geq \varepsilon_1 + \dots + \varepsilon_n. \quad (13.9)$$

We can now proceed in principle in two different ways. We could begin with $n = 2$ and go on with $n = 3, 4, \dots$. Let

$$s_n = \sum_{v=1}^n \varepsilon_v \quad (n \geq 2).$$

Then we have from (13.9)

$$s_{n+1} = s_n + \varepsilon_{n+1} \geq 2s_n.$$

Now we have $s_2 = \varepsilon_1 + \varepsilon_2$ and $\text{Min}(\varepsilon_1, \varepsilon_2) = 1$. Therefore, it follows that

$$s_n \geq 2^{n-2} s_2 \geq 2^{n-1},$$

and we have at each step

$$K|\zeta - x_{n+1}| \leq u^{2^{n-1}}. \quad (13.10)$$

At each step our error will be squared. This is accomplished at the expense of one horner, i.e., by the calculation of f_{n+1} , and the efficiency index of this

procedure is 2. This procedure cannot, however, be recommended, since we have no check other than repeated calculation. We may, as an alternative, check a result by applying the Newton–Raphson formula, but this checking would be at the expense of two horners.[†]

4. Another possibility is to use the fixed number n of points at each step, i.e., to use x_1, \dots, x_n to compute x_{n+1} , and then to use x_2, \dots, x_{n+1} to compute x_{n+2} , etc. As before, at each step we use one horner. If we start from $x_\mu, x_{\mu+1}, \dots, x_{\mu+n-1}$, and compute $x_{\mu+n}$, then in the notation of (13.7), for a convenient value of k ,

$$u^{\varepsilon_{\mu+n}} \leq u^{\varepsilon_\mu + \varepsilon_{\mu+1} + \dots + \varepsilon_{\mu+n-1}}$$

and, since we assumed $u \leq 1$,

$$\varepsilon_{\mu+n} \geq \varepsilon_\mu + \varepsilon_{\mu+1} + \dots + \varepsilon_{\mu+n-1}. \quad (13.11)$$

5. Since the ε_μ are > 0 , it follows from (13.11) that ε_μ increase monotonically with μ . If we write $\lim_{\mu \rightarrow \infty} \varepsilon_\mu = A$, it follows from (13.11) that $A \geq nA$, and we see that $\varepsilon_\mu \rightarrow \infty$ ($\mu \rightarrow \infty$),

$$x_v - \zeta \rightarrow 0. \quad (13.12)$$

6. Let

$$\ln \frac{1}{|\zeta - x_v|} = y_v. \quad (13.13)$$

Then if we apply (13.4) to the sequence $x_\mu, x_{\mu+1}, \dots, x_{\mu+n-1}, x_{\mu+n}$ and take the moduli on both sides of (13.4), we have

$$y_{\mu+n} = y_\mu + \dots + y_{\mu+n-1} + k_\mu, \quad (13.14)$$

where k_μ tends to the finite limit

$$\kappa = -\ln \frac{1}{n!} |\Phi^{(n)}(0)| - n \ln |f'(\zeta)|.$$

7. Formula (13.14) is a linear difference equation of the type (12.1), where the k_μ form a bounded sequence. We will now show that the hypotheses of Theorem 12.1 are all satisfied. The characteristic polynomial of (13.14) is

$$f_n(x) \equiv x^n - x^{n-1} - \dots - x - 1 \quad (13.15)$$

and we now have to discuss it.

[†] This procedure appears to present another disadvantage, since it seems that the first f_ν must be calculated with arbitrarily great precision. However, as follows from Appendix H, this disadvantage is not very essential.

DISCUSSION OF THE ROOTS OF SOME SPECIAL EQUATIONS

8. We consider for an integer $n > 1$ and a p with $1 \leq p < n$ the equation

$$f_{n,p}(x) := px^n - x^{n-1} - \cdots - x - 1 = 0. \quad (13.16)$$

By the rule of Descartes, (13.16) has exactly one positive root $\mu_{n,p}$. We will write $\mu_{n,1} = \mu_n$. Since $f_{n,p}(1) = -(n-p) < 0$, we have $\mu_{n,p} > 1$.

Put

$$g_{n,p}(x) \equiv (x-1)f_{n,p}(x) = (px-p-1)x^n + 1.$$

We have

$$g_{n,p}\left(1 + \frac{1}{p}\right) = 1, \quad f_{n,p}\left(1 + \frac{1}{p}\right) = p > 0$$

and see that

$$1 < \mu_{n,p} < 1 + \frac{1}{p}. \quad (13.17)$$

From

$$f_{n+1,p}(x) = xf_{n,p}(x) - 1, \quad f_{n+1,p}(\mu_{n,p}) = -1$$

follows

$$\mu_{n,p} < \mu_{n+1,p}.$$

We verify immediately that

$$g'_{n,p}(x) = (n+1)p\left(x - \frac{p+1}{n+1} \cdot \frac{n}{p}\right)x^{n-1}$$

and we see that, by Rolle's theorem, $g'_{n,p}(x)$ has only one positive zero lying between 1 and $\mu_{n,p}$, while $\mu_{n,p} > (1+1/p)/(1+1/n)$ is a simple zero. We give in the following another proof for the lower bound of $\mu_{n,p}$.

9. We will have to use the inequality, valid for integer $m \geq n$:

$$U_m(n,p) := (n+1)^{m+1} - (p+1)\left(1 + \frac{1}{p}\right)^m n^m < 0 \quad (1 \leq p < n \leq m). \quad (13.18)$$

This follows from

$$\frac{\partial}{\partial p} U_m(n,p) = \left(\frac{m}{p} - 1\right)n^m \left(1 + \frac{1}{p}\right)^m > 0 \quad (13.19)$$

if we observe that

$$U_m(n,n) = 0 \quad (m \geq n).$$

Using (13.18) with $m = n$ we see that

$$g_{n,p} \left(\frac{n}{n+1} \left(1 + \frac{1}{p} \right) \right) = 1 - \left(\frac{p+1}{n+1} \right)^{n+1} \left(\frac{n}{p} \right)^n < 0$$

and therefore

$$\frac{n}{n+1} \left(1 + \frac{1}{p} \right) < \mu_{n,p} < 1 + \frac{1}{p}, \quad \mu_{n,p} \uparrow 1 + \frac{1}{p} \quad (n \rightarrow \infty). \quad (13.20)$$

10. We now prove a lemma about the roots of (13.16) distinct from $\mu_{n,p}$. We denote by $q_{n,p}$ the maximum modulus of all roots of (13.16) which are not equal to $\mu_{n,p}$ and further put $q_{n,1} = q_n$.

Lemma 13.1. *We have $q_{2,p} = \mu_{2,p} - 1/p < 1$ and*

$$q_{n,p} < \mu_{n,p} - \frac{1}{p} < 1 \quad (n > 2). \quad (13.21)$$

Proof. We write for the sake of simplicity μ for $\mu_{n,p}$ and put $\xi = 1/\mu$, Then we have

$$\frac{f_{n,p}(x)}{x-\mu} = c_0 x^{n-1} + \cdots + c_v x^{n-v+1} + \cdots + c_{n-1}, \quad (13.22)$$

where

$$c_0 = p, \quad c_v = p\mu^v - \mu^{v-1} - \cdots - \mu - 1 \quad (1 \leq v \leq n-1).$$

If we bring the last $n-v$ terms to the right in the equation

$$p\mu^n - \mu^{n-1} - \cdots - \mu - 1 = 0$$

and divide by μ^{n-v} , we obtain

$$c_v = \frac{1}{\mu} + \frac{1}{\mu^2} + \cdots + \frac{1}{\mu^{n-v}} = \xi \frac{1 - \xi^{n-v}}{1 - \xi} \quad (v = 1, \dots, n-1).$$

We see that all coefficients c_v of (13.22) are positive.

11. We now prove

$$\frac{c_{v+1}}{c_v} < \frac{c_v}{c_{v-1}} \quad (v = 1, \dots, n-2). \quad (13.23)$$

For $v = 1$ this is equivalent to

$$c_1^2 > pc_2, \quad (p\mu - 1)^2 > p(p\mu^2 - \mu - 1), \quad \mu < 1 + \frac{1}{p},$$

and this holds indeed by (13.17).

On the other hand, for $2 \leq v \leq n-2$, putting $k = n-v$, $2 \leq k \leq n-2$, (13.23) is equivalent to

$$\begin{aligned}\frac{1-\xi^{k+1}}{1-\xi^k} &< \frac{1-\xi^k}{1-\xi^{k-1}}, \\ (1-\xi^k)^2 &> (1-\xi^{k+1})(1-\xi^{k-1}), \\ 2\xi^k &< \xi^{k-1} + \xi^{k+1},\end{aligned}$$

and this is true by the inequality between the arithmetic and the geometric mean.

12. We now make use of the following theorem due (without the second part) to G. Eneström and S. Kakeya:

If in the equation

$$g(x) = a_0 x^n + \cdots + a_n = 0 \quad (13.24)$$

all coefficients a_v are positive, then we have for each root ξ of (13.24)

$$|\xi| \leq \gamma = \operatorname{Max}_{1 \leq v \leq n} \frac{a_v}{a_{v-1}}. \quad (13.25)$$

Let k_1, k_2, \dots, k_m be the indices k , for which $a_k/a_{k-1} < \gamma$. Then, if the greatest common divisor of $k_1, \dots, k_m, n+1$ is 1, we have in (13.25) the strict inequality.[†]

Applying this theorem to the zeros of the polynomial (13.22), we have by virtue of (13.23) for $n > 2$, since the greatest common divisor of $2, 3, \dots, n$ is 1,

$$q_n < \operatorname{Max} \left(\frac{c_1}{c_0}, \frac{c_2}{c_1}, \dots, \frac{c_{n-1}}{c_{n-2}} \right) = \frac{c_1}{c_0} = \frac{1}{p} c_1. \quad (13.26)$$

Since $c_1 = p\mu_{n,p} - 1$, it remains only to deal with $n = 2$. But then in the equation $px^2 - x - 1 = 0$ the sum of both roots $\mu_{2,p}$ and $-q_{2,p}$ is $1/p$ and we have $q_{2,p} = \mu_{2,p} - 1/p$. Our lemma is now completely proved.

13. For the sake of completeness, we give in what follows a proof of the above theorem, since only the mixed inequality is dealt with in the literature.

We have

$$(x-\gamma)g(x) = a_0 x^{n+1} - (\gamma a_0 - a_1)x^n - \cdots - (\gamma a_v - a_{v+1})x^{n-v} - \cdots - a_n \gamma.$$

Since by definition of γ all expressions $\gamma a_{v-1} - a_v$ ($v = 1, \dots, n$) are ≥ 0 , γ is a

[†] The example of the polynomial $x^3 + yx^2 + cx + yc = (x+y)(x^2 + c)$ with $0 < c < y^2$ shows that without an additional condition the strict inequality in (13.25) cannot be enforced.

simple zero and the only positive zero of $(x - y)g(x)$, and (13.25) follows from Theorem 12.2. If the conditions of the second part of the theorem are satisfied, the indices of the nonvanishing coefficients of $(x - y)g(x)$ have the greatest common divisor 1. Therefore, the one positive zero γ of this polynomial is greater than the moduli of all other zeros, that is, of all zeros of (13.24).

14. We can now apply Theorem 12.1 with $u_1 = \mu_n$ and $s = 1$ and obtain from (12.13) $y_v/\mu_n^v \rightarrow \alpha > 0$, and therefore

$$\ln |\zeta - x_v| \sim -\alpha \mu_n^v, \quad (13.27)$$

$$\frac{\ln |\zeta - x_{v+1}|}{\ln |\zeta - x_v|} \rightarrow \mu_n. \quad (13.28)$$

Hence, if we use several steps of the n -point interpolation, the error will at each step be raised (asymptotically) to the power μ_n , and the efficiency index of our iteration procedure is μ_n .

15. In order to apply (12.16) to the sequence y_v , we obtain from (13.2) and (13.4)

$$\frac{|\zeta - x_1| \cdots |\zeta - x_n|}{|\zeta - x_{n+1}|} = \frac{n!}{|\Phi^{(n)}(\eta)|} \prod_{v=1}^n \frac{1}{|f'(\xi_v)|}$$

and, therefore, assuming that

$$\Phi^{(n)}(0) \neq 0, \quad f'(\zeta) \neq 0, \quad (13.29)$$

$$y_{n+1} - y_1 - \cdots - y_n = \ln \left| n! \frac{f'(\zeta)^{-n}}{\Phi^{(n)}(0)} \right| - \ln \left| \frac{\Phi^{(n)}(\eta)}{\Phi^{(n)}(0)} \right| - \sum_{v=1}^n \ln \left| \frac{f'(\xi_v)}{f'(\zeta)} \right|.$$

If now, as $v \rightarrow \infty$,

$$x_v \rightarrow \zeta, \quad y_v \rightarrow \infty,$$

we can write

$$\begin{aligned} y_{\mu+n} - y_\mu - \cdots - y_{\mu+n-1} &= \ln \left| n! \frac{f'(\zeta)^{-n}}{\Phi^{(n)}(0)} \right| \\ &\quad - \ln \left| \frac{\Phi^{(n)}(\eta^{(\mu)})}{\Phi^{(n)}(0)} \right| - \sum_{v=1}^n \ln \left| \frac{f'(\xi_v^{(\mu)})}{f'(\zeta)} \right|. \end{aligned} \quad (13.30)$$

16. As to the $\xi_v^{(\mu)}$, they lie between the greatest and the smallest of the $n+2$ numbers

$$x_\mu, \dots, x_{\mu+n}, \quad \zeta$$

and tend to ζ as $\mu \rightarrow \infty$. And we easily see from (13.2) that $|\eta^{(\mu)}| \leq \text{Max}(|f(x_\mu)|, \dots, |f(x_{\mu+n-1})|)$. We have, therefore, since, from a certain index μ on, the $|x_\mu - \zeta|$ decrease,

$$\eta^{(\mu)} = O(x_\mu - \zeta), \quad \xi_v^{(\mu)} - \zeta = O(x_\mu - \zeta).$$

But then, if we assume that $\Phi^{(n+1)}$ is continuous in the neighborhood of the origin, that is to say, that $f^{(n+1)}$ is continuous in the neighborhood of ζ , we have

$$\begin{aligned} \frac{\Phi^{(n)}(\eta^{(\mu)})}{\Phi^{(n)}(0)} &= 1 + O(|x_\mu - \zeta|), & \frac{f'(\xi_v^{(\mu)})}{f'(\zeta)} &= 1 + O(|x_\mu - \zeta|), \\ k_{n+\mu} &\equiv -\ln \left| \frac{\Phi^{(n)}(\eta^{(\mu)})}{\Phi^{(n)}(0)} \right| - \sum_{v=1}^n \ln \left| \frac{f'(\xi_v^{(\mu)})}{f'(\zeta)} \right| = O(|x_\mu - \zeta|). \end{aligned} \quad (13.31)$$

If we further put

$$\beta = \frac{1}{1-n} \ln \left| \frac{n!}{\Phi^{(n)}(0) f'(\zeta)^n} \right|, \quad (13.32)$$

we can write (13.30) in the form

$$y_{\mu+n} - y_\mu - \dots - y_{\mu+n-1} = (1-n)\beta + k_{n+\mu}. \quad (13.33)$$

Finally, put

$$y_\mu = v_\mu + \beta. \quad (13.34)$$

Equation (13.33) becomes

$$v_{\mu+n} - v_\mu - \dots - v_{\mu+n-1} = k_{n+\mu}. \quad (13.35)$$

17. The polynomial (13.15) belongs to the difference equation (13.35) as its characteristic polynomial. But now all conditions of Theorem 12.1 are satisfied, $K(x)$ being an entire function, so that s can be taken arbitrarily small. We obtain from (12.16) for $u_1 = \mu_n$, $m \leq n-1$,

$$v_v = \alpha \mu_n^v + O(v^{n-1} q_n^v), \quad y_v = \beta + \alpha \mu_n^v + O(v^{n-1} q_n^v),$$

$$|x_v - \zeta| = \exp(-\beta) \exp(-\alpha \mu_n^v) [1 + O(v^{n-1} q_n^v)], \quad (13.36)$$

$$\frac{|x_{v+1} - \zeta|}{|x_v - \zeta|^{\mu_n}} = \exp[(\mu_n - 1)\beta] + O(v^{n-1} q_n^v), \quad (13.37)$$

where

$$0 < q_n \leq \mu_n - 1 < 1. \quad (13.38)$$

18. The following table[†] of the μ_n shows that it is usually not worthwhile to take $n > 3$ or $n > 4$:

n	μ_n	n	μ_n
2	1.61803	9	1.99803
3	1.83929	10	1.99902
4	1.92756	11	1.99952
5	1.96595	12	1.99976
6	1.98358	13	1.99988
7	1.99196	14	1.99994
8	1.99603	15	1.99997

[†] Computed by Mrs. Bertha H. Walter, Computation Laboratory, National Bureau of Standards, Washington, D.C.

14

$n+1$ Coincident Interpolation Points and Taylor Development of the Root

STATEMENT OF THE PROBLEM

1. Let $w = f(z)$ be defined in J_z . Assume that we are given $n+1$ coincident interpolation points, i.e., z_0 with given values

$$f(z_0) =: w_0, \quad f'(z_0), \quad f''(z_0), \dots, f^{(n)}(z_0).$$

We develop z around the point w_0 :

$$z = \varphi(w) = z_0 + \sum_{v=1}^{\infty} \frac{(w-w_0)^v}{v!} \varphi^{(v)}(w_0). \quad (14.1)$$

For $w = 0$ it follows that

$$\zeta - z_0 = \sum_{v=1}^{\infty} \frac{(-1)^v}{v!} w_0^v \varphi^{(v)}(w_0). \quad (14.2)$$

The n th section of this series gives an approximation using only the given data. The estimate of the error can then be obtained from formula (1B.13), if the expression of $\varphi^{(n+1)}$ is obtained from (2.5) and the table in Chapter 2, Section 8, or directly from formula (C.5) of Appendix C. But this very soon becomes too complicated for practical use, and the existence of the root ζ has to be discussed separately.

A THEOREM ON INVERSE FUNCTIONS AND CONFORMAL MAPPING

2. On the other hand, the use of the series (14.2) may be particularly advisable if the computation of the derivatives at z_0 is easier than for other values of z . But of course, then all $f^{(v)}(z_0)$ used have to be computed at once with a considerable number of decimals.

In order to discuss the convergence of (14.2) and the existence of ζ and to obtain a practical estimate for the remainder, we now use the theory of functions of a complex variable. We prove first:

Theorem 14.1. Assume that $w = f(z)$ is analytic in a circle K_z : $|z - z_0| \leq r$. Let $w_0 = f(z_0)$. If we have

$$M = M(r) = \max_{K_z} |f''(z)| < \frac{2|f'(z_0)|}{r}, \quad (14.3)$$

then the inverse series

$$z = \varphi(w) = z_0 + \sum_{v=1}^{\infty} \frac{(w-w_0)^v}{v!} \varphi^{(v)}(w_0) \quad (14.4)$$

converges in the circle K_w : $|w - w_0| < R$ with

$$R = R(r) = |f'(z_0)|r - \frac{Mr^2}{2} \quad (14.5)$$

and there satisfies the inequality

$$|\varphi(w) - z_0| < r. \quad (14.6)$$

3. Proof. Without loss of generality we can assume that $z_0 = w_0 = f(z_0) = 0$. Put $f'(0) = f'_0$ and write

$$f(z) = f'_0 z + T(z). \quad (14.7)$$

Consider the expression

$$z^2 \int_0^1 (1-t) f''(tz) dt. \quad (14.8)$$

We introduce here a new variable of integration $u = tz$ and integrate by parts; (14.8) becomes

$$\int_0^z (z-u) f''(u) du = (z-u) f'(u) \Big|_0^z + \int_0^z f'(u) du = f(z) - f'_0 z,$$

and hence by (14.7)

$$T(z) = z^2 \int_0^1 (1-t) f''(tz) dt. \quad (14.9)$$

4. We have now for all z on the boundary of K_z

$$|T(z)| \leq r^2 \left| \int_0^1 (1-t) f''(tz) dt \right| \leq r^2 M \int_0^1 (1-t) dt = \frac{Mr^2}{2}. \quad (14.10)$$

If, on the other hand, we define $L(z)$ by

$$L(z) = f'_0 z - w, \quad (14.11)$$

we have from (14.7)

$$f(z) - w = L(z) + T(z) \quad (14.12)$$

and from (14.11) and (14.5) for $|z| = r$, if $|w| < R$,

$$|L(z)| \geq |f'_0|r - |w| > \frac{Mr^2}{2},$$

and therefore by (14.10) for z on the boundary of K_z and w inside K_w :

$$|L(z)| > |T(z)| \quad (|w| < R). \quad (14.12a)$$

5. We now need the so-called theorem of *Rouché*:

Let the two functions $g(z)$ and $h(z)$ be analytic in the simply connected region G . Let the simple closed path C lie within G and let $|g(z)| > |h(z)|$ along C . Then the functions $g(z)$ and $g(z) + h(z)$ have the same number of zeros in the subregion of G enclosed by C .

We apply this theorem to $g(z) = L(z)$, $h(z) = T(z)$, the boundary of K_z for C , and $L(z) + T(z) = f(z) - w$. From (14.11) we see that the only zero of $L(z)$ is given by

$$z_0' := \frac{w}{f'_0}.$$

Since by (14.5)

$$|z_0'| = \frac{|w|}{|f'_0|} < r - \frac{Mr^2}{2|f'_0|} < r,$$

z_0' lies inside K_z . Hence we have in K_z exactly one root of $f(z) = w$.

We have therefore in $w = f(z)$ a conformal mapping of a certain subregion of K_z containing 0 into the whole interior of K_w . We see that the inverse function $z = \varphi(w)$ is analytic in K_w and there satisfies inequality (14.6). Theorem 14.1 is proved.

6. In applying Theorem 14.1, it may be important to choose r in such a way as to make R as large as possible. Although this is not essential for our immediate purpose, which is the deduction of Theorem 14.2 about the Taylor development of a root of $f(z) = 0$, we shall say a few words about this problem. It is well known that

$$M(r) = \max_{|z| \leq r} |f''(z)|$$

as a function of r is continuous and strictly monotonically increasing, unless $f''(z)$ is constant.

Now, if we have for the chosen value of r

$$M(r) > \frac{|f'(z_0)|}{r}, \quad (14.13)$$

then the value (14.5) can be improved. Indeed, there exists in this case a positive $\rho < r$, such that $\rho M(\rho) = |f'(z_0)|$. And, the hypotheses of Theorem 14.1 being satisfied for ρ instead of for r , we obtain for $R(\rho)$ the value

$$|f'(z_0)|\rho - \frac{1}{2}M(\rho)\rho^2 = \frac{1}{2}|f'(z_0)|\rho = \frac{|f'(z_0)|^2}{2M(\rho)}. \quad (14.14)$$

This value is greater than $|f'(z_0)|^2/(2M(r))$ and is by virtue of the identity

$$2M\left[\frac{1}{2}\frac{|f'_0|^2}{M} - \left(|f'_0|r - \frac{Mr^2}{2}\right)\right] = (|f'_0| - rM)^2, \quad (14.15)$$

where we wrote M for $M(r)$, greater than $R(r)$.

In practice it may be sufficient to use the value $r_1 = |f'(z_0)|^2/(2M(r))$ if this r_1 satisfies (14.13).

THEOREM ON THE ERROR OF THE TAYLOR APPROXIMATION TO THE ROOT

7. From Theorem 14.1 we now deduce a theorem giving a good practical solution of the problem concerning the numerical use of the series (14.2):

Theorem 14.2. *Let $f(z)$ be nonlinear and analytic at z_0 and denote by σ the radius of convergence of the Taylor development of $f(z)$ around z_0 . Put $f(z_0) =: w_0$, $|f'(z_0)| =: f'_0$, and for $0 < r < \sigma$*

$$M(r) := \max_{|z-z_0| \leq r} |f''(z)|;$$

further,

$$R(r) = f'_0 r - \frac{1}{2}M(r)r^2. \quad (14.16)$$

Then, if we have $|w_0| < R(r)$ for an $r < \sigma$, there exists one and only one root ζ of $f(z) = 0$ with $|\zeta - z_0| < r$, and the development (14.2) is convergent to the value $\zeta - z_0$ and is majorized by

$$r \sum_{v=1}^{\infty} \frac{|w_0|^v}{R(r)^v}.$$

We have in particular, putting

$$E_n = \zeta - z_0 - \sum_{v=1}^n \frac{(-1)^v}{v!} w_0^v \varphi^{(v)}(w_0), \quad (14.17)$$

$$|E_n| \leq \left[\frac{|w_0|}{R(r)} \right]^{n+1} \frac{r}{1 - |w_0|/R(r)}. \quad (14.18)$$

8. Proof. Since by Theorem 14.1 the radius of convergence of the series (14.1) is at least $R(r)$, the convergence of (14.2) is clear. Then it follows from (14.6) by Cauchy's inequalities that

$$\left| \frac{1}{v!} \varphi^{(v)}(w_0) \right| \leq \frac{r}{R^v(r)} \quad (v = 1, 2, \dots), \quad (14.19)$$

which gives us the majorant for (14.2) and in particular

$$\begin{aligned} \left| \zeta - z_0 - \sum_{v=1}^n \frac{(-1)^v}{v!} w_0^v \varphi^{(v)}(w_0) \right| &\leq r \sum_{v=n+1}^{\infty} \left[\frac{|w_0|}{R(r)} \right]^v \\ &= \left[\frac{|w_0|}{R(r)} \right]^{n+1} \frac{r}{1 - |w_0|/R(r)}, \quad \text{Q.E.D.} \end{aligned}$$

DISCUSSION OF THE CONDITIONS OF THE THEOREM

9. We shall now discuss the condition $|w_0| < R(r)$. If we divide this inequality by $f'_0/2$ (this is not = 0 if $R(r) > 0$) and use the notation

$$\rho = |h_0|, \quad h_0 = \frac{f(z_0)}{f'(z_0)} = \frac{w_0}{f'(z_0)}, \quad (14.20)$$

we obtain the condition

$$2\rho < r \left[2 - \frac{rM(r)}{f'_0} \right]; \quad (14.21)$$

it follows immediately that $r > \rho$. Solving (14.21) with respect to $rM(r)$, we get

$$rM(r) < 2 \left(1 - \frac{\rho}{r} \right) f'_0, \quad (14.22)$$

$$M(r) < 2 \left(\frac{1}{r} - \frac{\rho}{r^2} \right) f'_0. \quad (14.23)$$

Suppose now that (14.23) is satisfied for an $r > 2\rho$. If we replace this r by 2ρ , then the left-hand expression in (14.23) becomes smaller and the right-hand expression larger, since the right-hand expression decreases monotonically with increasing $r > 2\rho$. Therefore, (14.22) is then *a fortiori* satisfied for $r = 2\rho$. We obtain therefore in

$$r = 2\rho < \sigma, \quad 2\rho M(2\rho) < f'_0 \quad (14.24)$$

a very handy sufficient condition for the convergence of the development (14.2).

We shall now compare the convergence conditions (14.22) and (14.24) with the condition of Theorem 7.2,

$$2\rho M \leq f'_0, \quad (14.25)$$

where ρ is given by (14.20), while $M = \text{Max}|f''(z)|$ is taken in the circle K_0 , $|z - z_1| = |z - z_0 - h_0| \leq \rho$. Since $M(2\rho)$ is the $\text{Max}|f''(z)|$ taken in the circle $K^{(2\rho)}$, $|z - z_0| \leq 2\rho$, which contains K_0 (see Fig. 5), (14.24) does not follow necessarily from (14.25), although, in practice, (14.24) will usually be satisfied together with (14.25).

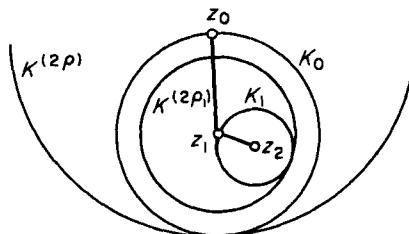


FIGURE 5

It can be deduced from Theorem 7.1, however, that condition (14.24) becomes satisfied as soon as we replace z_0 by the next Newton-Raphson approximation z_1 , except for a very special case. Indeed, from (7.17) it follows that if we put $\rho_1 = |h_1|$, the circle with the radius $2\rho_1$ around z_1 lies in the circle K_0 (see Fig. 5) and we can therefore use, indeed, M of Theorem 7.2 as a bound for $M(2\rho_1)$ computed in the circle around z_1 . And we have in these relations the strict inequality, unless $f''(z)$ is a constant, i.e., $f(z)$ is a quadratic polynomial. The same holds obviously for any z_v , $v > 1$. If we use further the last remark of Appendix F, we obtain:

Corollary to Theorem 14.2. *If the conditions of Theorem 7.2 are satisfied, Theorem 14.2 and its error estimates hold, replacing z_0 , $f(z_0)$ with z_v , $f(z_v)$, $v > 0$, unless $f(z)$ is a quadratic polynomial with a double root.*

If we use such a development, it would, however, not be very advantageous to take $r = 2|h_v|$, since then the error estimate (14.18) gives only a relatively large value. We can, though, usually take for r the distance of z_v from the circle $|z - z_1| = |h_0|$ and obtain in this way very satisfactory estimates, if v is sufficiently great.

10. The bound (14.18) is usually much greater than the actual value of the error, although it is of the right order in w_0 . On the other hand, the computer usually assumes—and very often rightly—that the actual error is of the

order of the next term of the series. This cannot be concluded directly from (14.18), since the estimate (14.19) is usually a very rough one. We can, however, proceed as follows:

We apply (14.18) for an $m > n$ and obtain then

$$|E_n| \leq \left[\frac{|w_0|}{R(r)} \right]^{m+1} \frac{r}{1 - |w_0|/R(r)} + \sum_{v=n+1}^m \frac{|w_0|^v}{v!} |\varphi^{(v)}(w_0)|. \quad (14.26)$$

Using this formula, it is, of course, not necessary to compute the values of the single terms in the second sum to the right with more than one or two significant figures; even rough estimates can be used if they are better than (14.19).

11. Example. The equation

$$f(x) = x^3 - 2x - 5 = 0$$

has exactly one positive root:

$$2.0945514815423 \dots$$

We have here $f'(x) = 3x^2 - 2$, $f''(x) = 6x$ and take

$$x_0 = 2, \quad w_0 = -1, \quad f'_0 = 10, \quad f''(x_0) = 12.$$

Obviously $M(r) = 12 + 6r$. If we take $r = \frac{1}{2}$, we obtain $R(r) = 3.125$.

The first three derivatives of $\varphi(w)$ in w_0 are

$$\varphi'(w_0) = 0.1, \quad \varphi''(w_0) = -0.012, \quad \varphi'''(w_0) = 0.00372,$$

and with $n = 3$ we have

$$\zeta = 2 + 0.1 - 0.006 + 0.00062 = 2.09462.$$

The true error E_3 is here 0.00007, while the estimated error from (14.18) is 0.0072.

If we take

$$x_0 = 2.1, \quad w_0 = 0.061, \quad f'_0 = 11.23, \quad f''_0 = 12.6, \quad r = \frac{1}{2}, \quad (14.27)$$

we see as before that $R = 3$ will satisfy the hypothesis of Theorem 14.1. Then we get

$$\begin{aligned} \varphi'_0 &= 0.089047195, & \varphi''_0 &= -0.00889674773, & \varphi'''_0 &= 0.002289382342, \\ \zeta &= 2.094551482097. \end{aligned} \quad (14.28)$$

Our true error is $E = 4.6 \cdot 10^{-10}$, while our estimated error is $9.33 \cdot 10^{-8}$.

If two more terms of our series in the case (14.27) had been used, we would have obtained the root correct to 15 decimal places.

Another method of approximation to the roots of $f(z) = 0$ will be discussed in Appendix J.

15

The Square Root Iteration

POLYNOMIALS WITH SIMPLE REAL ZEROS ONLY

1. Consider a polynomial

$$f(x) = \prod_{v=1}^n (x - \zeta_v) \quad (15.1)$$

of exact degree $n > 1$ with n real zeros ordered monotonically:

$$\zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_n. \quad (15.2)$$

Then $f'(x)$ has $n-1$ real zeros ζ'_v which can be ordered in such a way that we have, by Rolle's theorem,

$$\zeta_1 \leq \zeta'_1 \leq \zeta_2 \leq \zeta'_2 \leq \cdots \leq \zeta'_{n-1} \leq \zeta_n, \quad (15.3)$$

while, if $\zeta_v < \zeta_{v+1}$, we have even

$$\zeta_v < \zeta'_v < \zeta_{v+1}. \quad (15.4)$$

2. To any real x which is *distinct from all ζ_v and ζ'_v* we can assign in a unique way a certain zero of $f(x)$, the associated zero $\zeta(x)$. If $x < \zeta_1$, we put $\zeta(x) = \zeta_1$ and if $x > \zeta_n$, then $\zeta(x) = \zeta_n$. If, on the other hand, $\zeta_v < x < \zeta_{v+1}$ we take as $\zeta(x)$ the one of the two zeros ζ_v, ζ_{v+1} which is separated from ζ'_v by x . In this way in each interval between x and $\zeta(x)$, $f(x)/f'(x)$ and therefore $f'(x)/f(x)$ keeps constant sign. If in particular $x > \zeta(x)$ and if $\operatorname{sgn} f'(x) = +1$ then, since x lies to the right of $\zeta(x)$, we have also $\operatorname{sgn} f(x) = +1$, and if $\operatorname{sgn} f'(x) = -1$, we have in the same way $\operatorname{sgn} f(x) = -1$, that is, in any case $\operatorname{sgn}(f/f') = +1$.

In exactly the same way we see that if $x < \zeta(x)$, then $\operatorname{sgn} f(x) = -1$. We can write therefore generally

$$\operatorname{sgn}(f(x)/f'(x)) = \operatorname{sgn}(x - \zeta(x)). \quad (15.5)$$

3. Taking the logarithmic derivative of (15.1),

$$\frac{f'(x)}{f(x)} = \sum_{v=1}^n \frac{1}{x - \zeta_v}, \quad (15.6)$$

and differentiating this again, we obtain after multiplication by -1

$$H(x) := \frac{f'^2 - ff''}{f^2} = \sum_{v=1}^n \frac{1}{(x - \zeta_v)^2}. \quad (15.7)$$

Formulas (15.6) and (15.7) also hold, of course, if the ζ_v are not necessarily real. But since in our case all ζ_v are real, we have obviously from (15.7)

$$\frac{1}{(x - \zeta(x))^2} < H(x), \quad |x - \zeta(x)| > \frac{1}{\sqrt{H(x)}}. \quad (15.8)$$

We see in particular that $H(x)$ is always *positive* for real x .

4. Introduce now the expression

$$K(x) := \frac{f(x)/f'(x)}{\sqrt{1 - f(x)f''(x)/f'(x)^2}}, \quad (15.9)$$

where the root in the denominator must be taken, of course, as positive.

Obviously we have $K(x)^2 = 1/H(x)$.

We take now an x distinct from all ζ_v and ζ'_v and form, starting with $x_0 := x$, the sequence x_v by the iteration rule

$$x_{v+1} = x_v - K(x_v) \quad (v = 0, 1, \dots, \quad x_0 = x). \quad (15.10)$$

5. Theorem 15.1. *The x_v in (15.10) ($v = 1, 2, \dots$) all lie in the open interval between x and $\zeta(x)$ and converge monotonically to $\zeta(x)$.*

Proof. Assume that we have $x = x_0 < \zeta(x)$ so that

$$f'(x)/f(x) < 0, \quad K(x) < 0.$$

Then it follows, if we use (15.8), that

$$x_0 < x_1 < \zeta(x),$$

and by the definition of $\zeta(x)$, $\zeta(x_0) = \zeta(x_1)$. But then we can apply the same argument repeatedly to x_1, x_2, \dots and obtain

$$x_0 < x_1 < x_2 < \dots < \zeta(x).$$

It follows that the sequence x_v converges monotonically to a certain limit ζ :

$$x_v \uparrow \zeta \leq \zeta(x).$$

From this convergence it follows that $x_{v+1} - x_v \rightarrow 0$ and, by (15.10),

$$K(\zeta) = \frac{f(\zeta)/f'(\zeta)}{\sqrt{1 - f(\zeta)f''(\zeta)/f'(\zeta)^2}} = 0.$$

We see that $\zeta = \zeta(x)$.

If $x = x_0 > \zeta(x)$, the argument is completely symmetric. Theorem 15.1 is proved.

6. The convergence of the sequence x_v in (15.10) to $\zeta(x)$ is, if $\zeta(x)$ is a simple zero of $f(x)$, particularly good: it is *cubic*. On the other hand, if $\zeta(x)$ is a multiple zero of $f(x)$, the convergence of x_v to $\zeta(x)$ is only *linear*. More precisely:

Theorem 15.2. *If, in the hypotheses of Theorem 15.1, $\zeta(x)$ is a simple zero of $f(x)$, we have, writing ζ instead of $\zeta(x)$,*

$$\frac{\zeta - x_{v+1}}{(\zeta - x_v)^3} \rightarrow \frac{3f''(\zeta)^2 - 4f'(\zeta)f'''(\zeta)}{24f'(\zeta)^2} \quad (v \rightarrow \infty). \quad (15.11)$$

If, however, ζ is a zero of $f(x)$ of multiplicity $p > 1$, we have

$$\frac{\zeta - x_{v+1}}{\zeta - x_v} \rightarrow 1 - \frac{1}{\sqrt{p}} \quad (v \rightarrow \infty). \quad (15.12)$$

Proof. Assume first that ζ is a *simple zero* of $f(x)$. Letting x tend to $\zeta = \zeta(x)$ and putting $k = -f(x)/f'(x)$ we have, using the first three terms of the binomial development of the root in (15.9), with $k = -f(x)/f'(x) \rightarrow 0$,

$$x - x_1 = K(x) = -k + \frac{1}{2}k^2 \frac{f''(x)}{f'(x)} - \frac{3}{8}k^3 \frac{f''(x)^2}{f'(x)^2}(1 + O(k)).$$

On the other hand, from (2.20) we have

$$\zeta - x = k - \frac{1}{2}k^2 \frac{f''(x)}{f'(x)} + \frac{3f''(x)^2 - f'(x)f'''(x)}{6f'(x)^2}k^3(1 + O(k)).$$

From these two formulas we obtain by addition

$$\zeta - x_1 = k^3 \frac{3f''(x)^2 - 4f'(x)f'''(x)}{24f'(x)^2}(1 + O(k)). \quad (15.13)$$

On the other hand, it follows from (2.20) that $\zeta - x \sim k$ and therefore from (15.13)

$$\frac{\zeta - x_1}{(\zeta - x)^3} \rightarrow \frac{3f''(\zeta)^2 - 4f'(\zeta)f'''(\zeta)}{24f'(\zeta)^2} \quad (x \rightarrow \zeta = \zeta(x)). \quad (15.14)$$

If we replace x by x_v here, (15.11) follows immediately.

7. Assume now that ζ is a multiple zero of $f(x)$ of multiplicity $p > 1$. Then we have $f(x) = (x - \zeta)^p \varphi(x)$, $\varphi(\zeta) \neq 0$, and therefore for $x \rightarrow \zeta$

$$f(x) \sim (x - \zeta)^p \varphi(\zeta),$$

and further, by differentiation,

$$f'(x) \sim p(x-\zeta)^{p-1} \varphi(\zeta), \quad f''(x) \sim p(p-1)(x-\zeta)^{p-2} \varphi(\zeta).$$

From these formulas we get

$$\frac{f(x)f''(x)}{f'(x)^2} \rightarrow 1 - \frac{1}{p} < 1, \quad \frac{f(x)}{f'(x)} \sim \frac{1}{p}(x-\zeta),$$

and, introducing this into (15.9),

$$K(x) \sim (x-\zeta)/\sqrt{p}.$$

But then it follows from (15.10), with $v = 0$, $x_0 = x$, that

$$x_1 - \zeta = x - \zeta - K(x),$$

and, dividing on both sides by $x - \zeta$,

$$\frac{x_1 - \zeta}{x - \zeta} \rightarrow 1 - \frac{1}{\sqrt{p}}.$$

Replacing x by x_v here, we obtain (15.12), and Theorem 15.2 is proved.

MODIFICATION FOR MULTIPLE ZEROS

8. In the case where $\zeta = \zeta(x)$ is a multiple zero of multiplicity $p > 1$ and if p is known, we can still obtain cubic convergence to ζ by slightly modifying the rule (15.10). Put $x = x_0$ and

$$x_{v+1} = x_v - \sqrt{p} K(x_v) \quad (x_0 = x; \quad v = 0, 1, \dots). \quad (15.15)$$

Then we have:

Theorem 15.3. If under the conditions of Theorem 15.2 the exact multiplicity of $\zeta = \zeta(x)$ is p and the rule (15.10) is replaced by the rule (15.15), then the x_v converge monotonically to ζ and we have

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^3} \rightarrow \frac{(p+2)f^{(p+1)}(\zeta)^2 - 2(p+1)f^{(p)}(\zeta)f^{(p+2)}(\zeta)}{2p(p+1)^2(p+2)f^{(p)}(\zeta)^2}. \quad (15.16)$$

9. Proof. We have $f(x) = (x-\zeta)^p \varphi(x)$ and, denoting by $\varphi_0, \varphi_0', \varphi_0''$, respectively, the values of $\varphi(\zeta), \varphi'(\zeta), \varphi''(\zeta)$,

$$\varphi(x) = \varphi_0 + (x-\zeta)\varphi_0' + \frac{1}{2}(x-\zeta)^2\varphi_0'' + O((x-\zeta)^3),$$

$$\varphi'(x) = \varphi_0' + (x-\zeta)\varphi_0'' + O((x-\zeta)^2),$$

$$\varphi''(x) = \varphi_0'' + O((x-\zeta)),$$

and therefore, differentiating $(x - \zeta)^p \varphi(x)$,

$$\begin{aligned} f(x) &= (x - \zeta)^p \varphi_0 + (x - \zeta)^{p+1} \varphi_0' + \frac{1}{2}(x - \zeta)^{p+2} \varphi_0'' + \dots, \\ f'(x) &= p(x - \zeta)^{p-1} \varphi_0 + (p+1)(x - \zeta)^p \varphi_0' \\ &\quad + \frac{1}{2}(p+2)(x - \zeta)^{p+1} \varphi_0'' + \dots, \\ f''(x) &= (p-1)p(x - \zeta)^{p-2} \varphi_0 + p(p+1)(x - \zeta)^{p-1} \varphi_0' \\ &\quad + \frac{1}{2}(p+1)(p+2)(x - \zeta)^p \varphi_0'' + \dots. \end{aligned} \tag{15.17}$$

From these formulas we have

$$\begin{aligned} p \left(\frac{f(x)}{(x - \zeta)^p} \right)^2 &= p \varphi_0^2 + 2p \varphi_0 \varphi_0' (x - \zeta) + p(\varphi_0'^2 + \varphi_0 \varphi_0'') (x - \zeta)^2 + \dots, \\ \left(\frac{f'(x)}{(x - \zeta)^{p-1}} \right)^2 &= p^2 \varphi_0^2 + 2p(p+1) \varphi_0 \varphi_0' (x - \zeta) \\ &\quad + [(p+1)^2 \varphi_0'^2 + p(p+2) \varphi_0 \varphi_0''] (x - \zeta)^2 + \dots, \\ \frac{f(x)f''(x)}{(x - \zeta)^{2p-2}} &= p(p-1) \varphi_0^2 + 2p^2 \varphi_0 \varphi_0' (x - \zeta) \\ &\quad + [p(p+1) \varphi_0'^2 + (p^2+p+1) \varphi_0 \varphi_0''] (x - \zeta)^2 + \dots. \end{aligned}$$

Adding the first and the third of these formulas and subtracting the second one, we obtain on the right-hand side

$$(\varphi_0 \varphi_0'' - \varphi_0'^2) (x - \zeta)^2 + O((x - \zeta)^3),$$

and therefore, multiplying by $(x - \zeta)^{2p}$,

$$\begin{aligned} pf(x)^2 - (x - \zeta)^2(f'(x)^2 - f(x)f''(x)) \\ = (\varphi_0 \varphi_0'' - \varphi_0'^2) (x - \zeta)^{2p+2} + O((x - \zeta)^{2p+3}). \end{aligned}$$

10. It follows now from (15.7) that

$$\begin{aligned} 1 - \frac{(x - \zeta)^2 H(x)}{p} &= \frac{pf(x)^2 - (x - \zeta)^2(f'(x)^2 - f(x)f''(x))}{pf(x)^2} \\ &= \frac{\varphi_0 \varphi_0'' - \varphi_0'^2}{p \varphi_0^2} (x - \zeta)^2 + O((x - \zeta)^3), \end{aligned}$$

and further that

$$\frac{p}{(x - \zeta)^2 H(x)} = 1 + \frac{\varphi_0 \varphi_0'' - \varphi_0'^2}{p \varphi_0^2} (x - \zeta)^2 + O((x - \zeta)^3).$$

Therefore, since $K(x)^2 = 1/H(x)$,

$$K(x)^2 = \frac{(x - \zeta)^2}{p} \left(1 + \frac{\varphi_0 \varphi_0'' - \varphi_0'^2}{p \varphi_0^2} (x - \zeta)^2 + O((x - \zeta)^3) \right).$$

If we take the square root, it follows that

$$K(x) = \frac{x-\zeta}{\sqrt{p}} \left(1 + \frac{\varphi_0 \varphi_0'' - \varphi_0'^2}{2p\varphi_0^2} (x-\zeta)^2 \right) + O((x-\zeta)^4). \quad (15.18)$$

Indeed, in (15.9) the sign of $K(x)$ is that of $f(x)/f'(x)$ and therefore that of $(x-\zeta)$, so that the sign in (15.18) has been correctly chosen.

11. From (15.18) we have

$$x - \zeta - \sqrt{p} K(x) = \frac{\varphi_0'^2 - \varphi_0 \varphi_0''}{2p\varphi_0^2} (x-\zeta)^3 + O((x-\zeta)^4),$$

and therefore, by virtue of (15.15),

$$x_{v+1} - \zeta = \frac{\varphi'(x_v)^2 - \varphi(x_v) \varphi''(x_v)}{2p\varphi_0^2} (x_v - \zeta)^3 + O((x_v - \zeta)^4).$$

From this we have finally

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^3} \rightarrow \frac{\varphi_0'^2 - \varphi_0 \varphi_0''}{2p\varphi_0^2}. \quad (15.19)$$

If we observe that, by (15.17),

$$\varphi_0 = \frac{f^{(p)}(\zeta)}{p!}, \quad \varphi_0' = \frac{f^{(p+1)}(\zeta)}{(p+1)!}, \quad \varphi_0'' = 2 \frac{f^{(p+2)}(\zeta)}{(p+2)!},$$

(15.16) now follows immediately from (15.19). Theorem 15.3 is proved.

DIFFERENTIABLE FUNCTIONS AND COMPLEX ZEROS

12. It can be expected *a priori* that iteration (15.10) is also convergent in the case of a three times differentiable function $f(x)$ which need not be a polynomial, as soon as this iteration starts in a sufficiently closed neighborhood of a simple zero of $f(x)$. This amounts to saying that a simple zero of $f(x)$ is always a point of attraction for iteration (15.10).

We are going to prove even more, namely, that even in the complex case, if we start in a sufficiently close neighborhood of a (real or complex) zero of $f(x)$, iteration (15.10) converges to this zero. It will even turn out that it is not necessary to *assume* the existence of a zero in the neighborhood of the starting point x_0 , but this existence can be *proved*, as soon as $f(x_0)$ is small enough compared with some other quantities. This proof can be given, using the general theorems 4.4 and 4.5.

Theorem 15.4. *Consider a function $f(x)$ defined in a neighborhood of x_0 and three times continuously differentiable in this neighborhood with*

$f'(x) \neq 0$. Put

$$h(x) = \frac{f(x)}{f'(x)}, \quad q(x) = \frac{f(x)f''(x)}{f'(x)^2}, \quad |h(x_0)| = h_0, \quad |q(x_0)| = q_0. \quad (15.20)$$

Assume that the above assumptions hold in the neighborhood U of x_0 : $|x - x_0| \leq 6h_0$ and that we have

$$q_0 \leq \frac{1}{16}. \quad (15.21)$$

Denoting $\text{Max}|f^{(3)}(x)|$ in U by M_3 , assume further that we have throughout U

$$|q(x)| \leq \frac{1}{2}, \quad 2M_3|h(x)|^2 \leq |f'(x)| \quad (|x - x_0| \leq 6h_0). \quad (15.22)$$

Then $f(x)$ has exactly one zero ζ_0 in the neighborhood U_0 , $|x - x_0| \leq 2h_0$, of x_0 and iteration (15.10) starting with every x from U_0 converges to ζ_0 .

13. Proof. We will apply Theorem 4.4 or, in the complex case, Theorem 4.5. In what follows, the letters $f, f', f'', f''', h, h', q, q'$ denote functions of x in U , while the moduli of their values in x_0 will be denoted, respectively, by $f_0, f'_0, f''_0, f'''_0, h_0, h'_0, q_0, q'_0$. Then, by (15.9) and (15.10), our iteration function is

$$\psi(x) = x - h(1-q)^{-1/2}, \quad (15.23)$$

and we will have to discuss $\psi'(x)$.

Observe that we have the relations

$$h' = 1 - q, \quad q' = \frac{ff'f^{(3)} + f'^2f'' - 2ff''^2}{f'^3} = h \frac{f^{(3)}}{f'} + \frac{f'^2f'' - 2ff''^2}{f'^3}. \quad (15.24)$$

Differentiating (15.23), we have by (15.24)

$$\begin{aligned} \psi'(x) &= 1 - (1-q)(1-q)^{-1/2} - \frac{1}{2}h(1-q)^{-3/2}q', \\ \psi'(x) &= 1 - \frac{h^2}{2}(1-q)^{-3/2} \frac{f^{(3)}}{f'} - (1-q)^{-3/2} \left[(1-q)^2 + \frac{h}{2} \frac{f'^2f'' - 2ff''^2}{f'^3} \right]. \end{aligned}$$

Here the expression in the bracket is, by (15.20),

$$\frac{(f'^2 - ff'')^2}{f'^4} + \frac{f(f'^2f'' - 2ff''^2)}{2f'^4} = \frac{2f'^4 - 3ff'^2f''}{2f'^4} = 1 - \frac{3}{2}q$$

and therefore

$$\begin{aligned} \psi'(x) &= 1 - \frac{1 - \frac{3}{2}q}{(1-q)^{3/2}} - \frac{h^2}{2(1-q)^{3/2}} \frac{f^{(3)}}{f'}, \\ \psi'(x) &= Q(q) - \frac{h^2}{2(1-q)^{3/2}} \frac{f^{(3)}}{f'}, \quad Q(u) = 1 - \frac{1 - \frac{3}{2}u}{(1-u)^{3/2}}. \quad (15.25) \end{aligned}$$

14. Differentiating $Q(u)$ in (15.25), we obtain at once

$$Q'(u) = \frac{3}{4} \frac{u}{(1-u)^{5/2}},$$

and therefore, if $|u| < 1$, $|Q'(u)| \leq Q'(|u|)$. It follows, therefore, that

$$|Q(u)| = \left| \int_0^u Q'(t) dt \right| \leq \left| \int_0^u Q'(|t|) dt \right|,$$

where the integration is (in the complex case) along the straight line joining 0 and u . We have further, putting

$$u = \varepsilon|u|, \quad |\varepsilon| = 1, \quad t = \varepsilon v,$$

$$|Q(u)| \leq \int_0^{|u|} Q'(v) dv = Q(|u|),$$

and therefore, since $Q'(u)$ is positive for $1 > u > 0$,

$$|Q(u)| \leq Q(w) \quad (|u| \leq w < 1).$$

Applying this to $Q(u)$ in (15.25) and using (15.22), we have

$$|Q(q)| \leq Q\left(\frac{1}{5}\right) = 1 - \frac{\frac{7}{5}}{2\left(\frac{4}{5}\right)^{3/2}} = \frac{16 - 7\sqrt{5}}{16} = \frac{11}{16(16 + 7\sqrt{5})} < \frac{1}{40}.$$

As to the second term in formula (15.25) for $\psi'(x)$, we have, by virtue of (15.22),

$$\left| \frac{h^2}{2(1-q)^{3/2}} \frac{f^{(3)}}{f'} \right| \leq \frac{1}{4|1-q|^{3/2}} \leq \frac{1}{4} \left(\frac{5}{4}\right)^{3/2} = \frac{5\sqrt{5}}{32} < 0.35.$$

It follows now from (15.25), throughout U , that

$$|\psi'(x)| \leq \frac{1}{40} + 0.35 < 0.4 \quad (|x-x_0| \leq 6h_0). \quad (15.26)$$

15. As to the value of $|\psi(x_0) - x_0|$, we have by (15.23) and (15.21)

$$|\psi(x_0) - x_0| = \frac{h_0}{(1-q_0)^{1/2}} \leq \sqrt{\frac{5}{4}} h_0.$$

We will now apply Theorems 4.4 and 4.5 to U_0 , that is, with $\eta = 2h_0$, and we have to verify that the condition $|\psi(x_0) - x_0| \leq \eta m$ or the corresponding condition of Theorem 4.5 is satisfied. Since our η here is $2h_0$, and m is 0.6, we have to verify that the condition

$$\sqrt{\frac{5}{4}} h_0 \leq 0.6 \cdot 2h_0,$$

that is to say, $\sqrt{5} = 2.236 \dots \leq 2.4$, holds, and Theorems 4.4 and 4.5 can be applied. There exists therefore in U_0 exactly one fixed point ζ of $\psi(x)$.

16. Observe now that the neighborhood of ζ , $|x - \zeta| \leq 4h_0$, is contained in U and contains U_0 . By Theorem 4.3 and (15.26) we have the convergence to ζ for every starting point from this neighborhood and therefore, *a fortiori*, from every starting point of U_0 . Theorem 15.4 is proved.

The conditions of this theorem differ from the conditions of the analogous existence theorems 7.1 and 7.2 in so far as the hypotheses (15.22) are assumed valid in the whole neighborhood and these hypotheses depend on the values of f, f', f'' in U . From this theorem we can, though, deduce also an “initial value theorem” depending on the values of f, f', f'' in x_0 and the bound M_3 of the modulus of the highest occurring derivative, f''' , throughout U . However, the constants must be considerably increased if we proceed in this way. We give, therefore, in the following chapter a direct derivation of an “initial value theorem” concerning the iteration by (15.23) which proceeds more along the lines of the proof in Chapter 7.

16

Further Discussion of Square Root Iteration

LOCAL FORMULATION OF THE EXISTENCE AND CONVERGENCE THEOREM

1. Theorem 16.1. Assume that $f(z)$ has in the point z_0 the continuous third derivative $f'''(z_0)$ and that $f(z_0)f'(z_0) \neq 0$. Form

$$h_0 = \frac{f(z_0)}{f'(z_0)}$$

and consider the neighborhood U_0 of z_0 :

$$U_0 \quad (|z - z_0| \leq 2|h_0|).$$

Assume that $f'''(z)$ is continuous in U_0 and put

$$M = \max_{z \in U_0} |f'''(z)|.$$

Further, put

$$f_0 = |f(z_0)|, \quad f'_0 = |f'(z_0)|, \quad f''_0 = \max(1, |f''(z_0)|),$$

$$q_0 = h_0 \frac{f''(z_0)}{f'(z_0)} = \frac{f(z_0)f''(z_0)}{f'(z_0)^2}, \quad p_0 = \frac{M|h_0|^2}{f'_0},$$

$$Q_0 = \frac{f_0 f''_0}{f'^2_0}, \quad k_0 = \frac{h_0}{\sqrt{1-q_0}},$$

where the value of the root is fixed, developing the root in powers of q_0 , as long as $|q_0| < 1$, and taking 1 as the first term.

$$Q_0 \leq \frac{1}{2}, \tag{16.1}$$

$$p_0 \leq \frac{1}{2}, \tag{16.2}$$

$$|h_0|M \leq \frac{1}{2}f''_0. \tag{16.3}$$

Then, if we form, starting with z_0 , the sequence z_v by the iteration formula

$z_{v+1} = \psi(z_v)$, where $\psi(x)$ is given by (15.23) and (15.20), all z_v remain in U_0 and converge to a zero ζ of $f(z)$ for which $|\zeta - z_0| < 1.6|h_0|$.

2. Proof. Put

$$w = \sqrt{1-q_0};$$

here the root is uniquely determined by its binomial development in powers of q_0 :

$$w = 1 - \sum_{v=1}^{\infty} \pi_v q_0^v,$$

where, as is well known, the coefficients π_v are *positive*. Since by (16.1)

$$|q_0| \leq Q_0 \leq \frac{1}{4}, \quad (16.4)$$

we have from our development

$$|1-w| \leq \sum_{v=1}^{\infty} \pi_v \left(\frac{1}{4}\right)^v = 1 - \sqrt{1-\frac{1}{4}} = \frac{2-\sqrt{3}}{2}.$$

From this we have further

$$|1+w| = |2-(1-w)| \geq 2 - \frac{2-\sqrt{3}}{2} = \frac{2+\sqrt{3}}{2},$$

$$\frac{1}{|1+w|} \leq \frac{2}{2+\sqrt{3}} = 2(2-\sqrt{3}),$$

and further,

$$\left| \frac{1}{(1+w)^2} \right| \leq 4(2-\sqrt{3})^2 = 4(7-4\sqrt{3}) \quad (16.5)$$

and, since

$$\frac{1}{|w|} = \frac{1}{|\sqrt{1-q_0}|} \leq \frac{1}{\sqrt{\frac{1}{4}}} = \frac{2\sqrt{3}}{3}, \quad (16.6)$$

$$\frac{1}{|w(1+w)|} \leq \frac{1}{3\sqrt{3}} \cdot 2(2-\sqrt{3}) = \frac{4}{3}(2\sqrt{3}-3) < 0.7. \quad (16.7)$$

Further, from (16.5) and (16.6)

$$\left| \frac{w+2}{2w(w+1)^2} \right| \leq 4 \left(\frac{1}{2} + \frac{2\sqrt{3}}{3} \right) (7-4\sqrt{3}) = \frac{2}{3} (16\sqrt{3}-27) < 0.5. \quad (16.8)$$

3. We now put, by (15.23),

$$k_0 = z_0 - z_1 = \frac{h_0}{w}, \quad \eta_0 := k_0 - h_0 = \frac{1-w}{w} h_0 = \frac{h_0 q_0}{w(w+1)}. \quad (16.9)$$

Then, by (16.7) and (16.4),

$$|\eta_0| = \frac{|h_0 q_0|}{|w(1+w)|} < 0.7|h_0 q_0| \leq (\frac{7}{40})|h_0| < \frac{1}{5}|h_0|. \quad (16.10)$$

Therefore, by (16.9),

$$|k_0| \leq 1.2|h_0|. \quad (16.11)$$

Now, putting

$$\delta_0 = \eta_0 - \frac{h_0 q_0}{2}, \quad (16.12)$$

we have further, by (16.9) and (16.8),

$$\begin{aligned} \delta_0 &= \frac{h_0 q_0}{w(1+w)} - \frac{h_0 q_0}{2} = \frac{(w+2)(1-w)}{2w(1+w)} h_0 q_0 = \frac{w+2}{2w(1+w)^2} h_0 q_0^2, \\ |\delta_0| &< \frac{1}{2}|h_0 q_0|^2, \end{aligned}$$

and further, since $q_0 = h_0 f''(z_0)/f'(z_0)$,

$$|\delta_0| < \frac{1}{2}|h_0|^3 \frac{f_0''^2}{f_0'^2}. \quad (16.13)$$

4. Now we put

$$f_1 = |f'(z_1)|, \quad f_1' = |f'(z_1)|, \quad f_1'' = \text{Max}(1, |f''(z_1)|), \quad (16.14)$$

$$h_1 = \frac{f(z_1)}{f'(z_1)}, \quad Q_1 = \left| \frac{f_1 f_1''}{f_1'^2} \right|, \quad p_1 = h_1^2 \frac{M}{f_1'}.$$

Since, by (16.9) and (16.11), z_1 lies in U_0 , we have, developing $f''(z_1)$ around z_0 and using the definitions of f_0'' and f_1'' in (16.14),

$$f''(z_1) - f''(z_0) = \theta k_0 M, \quad |\theta| \leq 1,$$

$$|f''(z_1) - f''(z_0)| \leq 1.2|h_0|M = 1.2f_0''\left(|h_0|\frac{M}{f_0''}\right),$$

and by (16.3)

$$|f''(z_1) - f''(z_0)| \leq 0.15f_0''.$$

Thence $|f''(z_1)| \leq 1.15f_0''$ and since in any case $1.15f_0'' > 1$,

$$f_1'' \leq 1.15f_0''.$$

On the other hand, $|f''(z_1)| \geq |f''(z_0)| - 1.5f_0'', f_1' \geq |f''(z_0)| - 0.15f_0''$, and from this it follows now that $f_1' \geq 0.85f_0''$, whether $|f''(z_0)|$ is ≤ 1 or > 1 . We have

$$\frac{4}{5}f_0'' < f_1'' < \frac{6}{5}f_0''. \quad (16.15)$$

Further, developing $f'(z_1)$, we have by (16.11),

$$f'(z_1) = f'(z_0) - k_0 f''(z_0) + \theta \frac{k_0^2}{2} M, \quad |\theta| \leq 1,$$

$$f_1' \geq f_0' - |k_0| f_0'' - \frac{|k_0|^2}{2} M \geq f_0' \left(1 - 1.2|h_0| \frac{f_0''}{f_0'} - 0.72 \frac{|h_0|^2 M}{f_0'} \right).$$

Here, however, $|h_0| f_0'' / f_0' = Q_0 \leq \frac{1}{4}$. Using this and (16.2), we obtain

$$f_1' \geq f_0' (1 - 0.3 - 0.09) > \frac{1}{2} f_0'. \quad (16.16)$$

5. Finally, developing $f(z_1)$, we have

$$f(z_1) = f(z_0) - k_0 f''(z_0) + \frac{k_0^2}{2} f''(z_0) + \theta \frac{k_0^3}{6} M, \quad |\theta| \leq 1.$$

We replace k_0 by $h_0 + \eta_0$ and k_0^2 by $(h_0 + \eta_0)^2$ here; then we have

$$f(z_1) = (f(z_0) - h_0 f'(z_0)) - \eta_0 f'(z_0) + \frac{(h_0 + \eta_0)^2}{2} f''(z_0) + \theta \frac{1.2^3}{6} |h_0|^3 M.$$

Here, the first right-hand bracket vanishes by the definition of h_0 . In the term $-\eta_0 f'(z_0)$ we replace η_0 by

$$\delta_0 + \frac{h_0 q_0}{2} = \delta_0 + \frac{h_0^2}{2} \frac{f''(z_0)}{f'(z_0)}.$$

Then we get

$$f(z_1) = -\delta_0 f'(z_0) - \frac{h_0^2}{2} f''(z_0) + \frac{(h_0 + \eta_0)^2}{2} f''(z_0) + \theta \frac{1.2^3}{6} |h_0|^3 M.$$

Combining the terms containing $f''(z_0)$, we obtain further

$$f_1 \leq |\delta_0| f_0' + |\eta_0| \frac{|2h_0 + \eta_0|}{2} f_0'' + 0.288 |h_0|^3 M.$$

The first right-hand term here is, by (16.13), $\leq |h_0|^3 \frac{1}{2} (f_0''^2 / f_0')$.

As to the second right-hand term, we have, by (16.10),

$$|\eta_0| < 0.7 |h_0 q_0| \leq 0.7 |h_0|^2 \frac{f_0''}{f_0'}, \quad \frac{|2h_0 + \eta_0|}{2} \leq 1.1 |h_0|,$$

and therefore

$$f_1 \leq |h_0|^3 \left(0.5 \frac{f_0''^2}{f_0'} + 0.77 \frac{f_0''^2}{f_0'} + 0.288M \right) < |h_0|^3 \left(1.3 \frac{f_0''^2}{f_0'} + 0.3M \right). \quad (16.17)$$

6. Dividing (16.17) on both sides by f_1' , we have, by (16.14), (16.16), (16.1), and (16.2),

$$\begin{aligned} |h_1| &= \frac{f_1}{f_1'} < |h_0| \left(2.6 \left(|h_0| \frac{f_0''}{f_0'} \right)^2 + 0.6 |h_0|^2 \frac{M}{f_0'} \right) \\ &< |h_0| \left(2.6 Q_0^2 + \frac{0.6}{8} \right) \leq \frac{19}{80} |h_0|, \\ |h_1| &< \frac{1}{4} |h_0|. \end{aligned} \quad (16.18)$$

On the other hand, multiplying (16.17) on both sides by $f_1''/f_1'^2$ and observing that this is, by (16.15) and (16.16), $\leq 4 \cdot 1.2 f_0''/f_0'^2 < 5(f_0''/f_0'^2)$, we have from the definitions of Q_1 , Q_0 , and p_0

$$Q_1 < 5 |h_0|^3 \left(1.3 \frac{f_0''^3}{f_0'^3} + 0.3 \frac{M f_0''}{f_0'^2} \right) = 6.5 Q_0^3 + 1.5 p_0 Q_0.$$

By (16.1) and (16.2) this is

$$\leq \frac{6.5}{64} + \frac{1.5}{8} \cdot \frac{1}{4} = \frac{9.5}{64} < \frac{1}{4},$$

and we have

$$Q_1 < \frac{1}{4}. \quad (16.19)$$

Further, by (16.18) and (16.16),

$$p_1 = |h_1|^2 \frac{M}{f_1'} \leq \frac{1}{16} |h_0|^2 \frac{M}{f_0'} = \frac{1}{8} p_0 < \frac{1}{8}. \quad (16.20)$$

Finally, from (16.3), (16.15), and (16.18)

$$\frac{|h_1|M}{f_1''} \leq \frac{1}{4} \cdot \frac{5}{4} \frac{|h_0|M}{f_0''} < \frac{1}{8}. \quad (16.21)$$

On the other hand, the neighborhood U_1 ($|z - z_1| \leq 2|h_1|$) is contained in U_0 since, by (16.11) and (16.18),

$$|z_1 - z_0| \leq 1.2 |h_0| < 2|h_0| - 2|h_1|.$$

7. We see that, if the conditions of our theorem are satisfied in the point

z_0 , the exactly corresponding conditions are satisfied in the point z_1 . Therefore we can indeed form the whole infinite sequence z_v by the iteration function (15.23) and have in particular

$$|z_{v+1} - z_v| \leq 1.2 \left| \frac{f(z_v)}{f'(z_v)} \right|, \quad \left| \frac{f(z_{v+1})}{f'(z_{v+1})} \right| < \frac{1}{4} \left| \frac{f(z_v)}{f'(z_v)} \right|.$$

From these inequalities we have now

$$\left| \frac{f(z_v)}{f'(z_v)} \right| \leq \frac{|h_0|}{4^v}, \quad |z_{v+1} - z_v| \leq \frac{1.2}{4^v} |h_0| \quad (v = 0, 1, 2, \dots),$$

and we see that $\zeta = \lim_{v \rightarrow \infty} z_v$ exists and that we have

$$|\zeta - z_0| \leq 1.2 \frac{|h_0|}{1 - \frac{1}{4}} = 1.6 |h_0|. \quad (16.22)$$

Now it follows from

$$0 = \lim_{v \rightarrow \infty} \frac{f(z_v)}{f'(z_v)} = \frac{f(\zeta)}{f'(\zeta)}$$

that in any case $f(\zeta) = 0$. Theorem 16.1 is proved.

EXTENSION TO ENTIRE FUNCTIONS

8. Theorem 15.1 has been formulated and proved for polynomials. Since the degree of the polynomial does not enter into the iteration formula (15.10), it can be expected that this theorem can be generalized to certain entire functions. This is indeed the case for a class of entire functions of order ≤ 2 .

We will say that an *entire function* $f(z)$ is of the class P if it is given by the formula

$$f(z) = z^m \exp(-\gamma z^2 + \alpha z + \beta) \prod_v \left[\left(1 - \frac{z}{a_v} \right) \exp \left(\frac{z}{a_v} \right) \right], \quad \gamma \geq 0, \quad (16.23)$$

where m is a nonnegative integer, α, β, γ are real constants, and $\gamma \geq 0$, while the a_v are real numbers $\neq 0$ for which

$$\sum_v \frac{1}{a_v^2} < \infty, \quad (16.24)$$

if the number of the a_v is infinite. We require further that there be at least one a_v and, if $m = 0$, even at least two.

It is well known that the product in (16.23) is uniformly convergent in any bounded region if it contains an infinite number of factors.

The most important example is given by

$$\sin z = z \prod_{v=1}^{\infty} \left(1 - \frac{z^2}{(\pi v)^2}\right) = z \prod_{\substack{v=-\infty \\ v \neq 0}}^{\infty} \left[\left(1 - \frac{z}{\pi v}\right) \exp\left(\frac{z}{\pi v}\right)\right]. \quad (16.25)$$

9. Taking the logarithmic derivative of (16.23) and differentiating it, we obtain

$$\frac{f'(z)}{f(z)} = -2\gamma z + \alpha + \frac{m}{z} + \sum_v \left(\frac{1}{z-a_v} + \frac{1}{a_v} \right), \quad (16.26)$$

$$H(z) \equiv -\left(\frac{f'(z)}{f(z)}\right)' = \frac{f'(z)^2 - f(z)f''(z)}{f(z)^2} = \frac{m}{z^2} + \sum_v \frac{1}{(z-a_v)^2} + 2\gamma. \quad (16.27)$$

From (16.27) we see that $f'(z)/f(z)$ is monotonically decreasing in any continuity interval.

If now $\zeta_0, \zeta_1, \zeta_0 < \zeta_1$, are two distinct ones among the zeros of $f(z)$ such that there are no further zeros in (ζ_0, ζ_1) it follows from (16.26) that

$$\lim_{z \downarrow \zeta_0} \frac{f'(z)}{f(z)} = \infty, \quad \lim_{z \uparrow \zeta_1} \frac{f'(z)}{f(z)} = -\infty.$$

We see that $f'(z)$ has then exactly one zero between ζ_0 and ζ_1 . Now, exactly as in the Section 2 of Chapter 15, we can associate with any z from the interval (ζ_0, ζ_1) , for which $f'(z) \neq 0$, its associated zero $\zeta(z)$ which has the value ζ_0 or ζ_1 and for which $f'(z)$ does not vanish between z and $\zeta(z)$.

10. From now on we can repeat the discussion of Sections 4 and 5 of Chapter 15, replacing there x by z , if this z is neither less nor greater than all zeros of $f(z)$. We obtain:

Theorem 16.2. *Assume that for a function (16.23) of the class P, $K(z)$ is defined for real z , with x replaced with z , by (15.9).*

Then, if the real z_0 is such that $f(z_0)f'(z_0) \neq 0$ and is neither greater nor less than all a_v , the iteration rule

$$z_{v+1} = z_v - K(z_v), \quad v = 0, 1, \dots, \quad (16.28)$$

gives a sequence z_v converging monotonically to $\zeta(z_0)$.[†]

[†] The result of Theorem 16.2 is far less general than that of Theorem 15.1 insofar as those values of z_0 are forbidden which are greater than all zeros or smaller than all zeros of $f(z)$. This is due to the fact that in our case the corresponding regions could still contain zeros of $f'(z)$ while in the polynomial case, discussed in Theorem 15.1, no zeros of $f'(z)$ can lie outside an interval containing all zeros of $f(x)$.

As to Theorems 15.2 and 15.3, they are generalized immediately to our case, assuming that the starting value satisfies the condition of Theorem 16.2.

The square root iteration discussed in Chapters 15 and 16 can be considered as the limiting case of an iteration formula given first by Laguerre in the case of real polynomials with only real zeros. We give the corresponding developments in Appendix O.

11. Applying Theorem 16.2 to $\sin z$ we obtain

$$K(z) = \frac{\tan z}{\sqrt{1+\tan^2 z}} = \frac{|\cos z| \sin z}{\cos z} = (\operatorname{sgn} \cos z) \sin z.$$

Assuming, for instance, z_0 between $\pi/2$ and π , $\operatorname{sgn} \cos z$ is -1 and our iteration formula becomes

$$z_{v+1} = z_v + \sin z_v.$$

Replacing $\sin z_v$ by $\sin(\pi - z_v)$ here and developing this in powers of $(\pi - z_v)$, we obtain for our iteration

$$\pi - z_{v+1} = \pi - z_v - [(\pi - z_v) - \frac{1}{6}(\pi - z_v)^3 + \dots],$$

and we see that we have indeed

$$\frac{\pi - z_{v+1}}{(\pi - z_v)^3} \rightarrow \frac{1}{6},$$

which agrees with (15.11).

17

A General Theorem on Zeros of Interpolating Polynomials

1. Theorem 17.1. *Let $f(z)$ be analytic on the disk*

$$|z - \zeta| \leq r \quad (17.1)$$

and have a zero in ζ of the exact order p . Take a natural $n > p$. Then there exists a positive $\varepsilon_0 < r/3$ with the following property: For a positive $\varepsilon \leq \varepsilon_0$ take n numbers z_1, \dots, z_n (not necessarily all distinct) on the disk

$$|z - \zeta| \leq \varepsilon \quad (17.2)$$

and form the interpolating polynomial of $f(z)$, $L(z)$, of order $\leq n-1$, corresponding to the interpolation abscissas z_1, \dots, z_n . Then $L(z)$ has in the interior of the disk (17.2) exactly p zeros, while all other zeros of $L(z)$ lie outside of the disk

$$|z - \zeta| \leq 3\varepsilon. \quad (17.3)$$

2. Put, under the conditions of Theorem 17.1,

$$K = \max_{|z| \leq r} |f^{(n)}(z)|. \quad (17.4)$$

Developing $f(z)$ in powers of $z - \zeta$, put

$$f(z) = T(z) + \varphi(z), \quad T(z) = \sum_{v=p}^{n-1} a_v (z - \zeta)^v, \quad a_p = \frac{f^{(p)}(\zeta)}{p!} \neq 0, \quad (17.5)$$

where $\varphi(z)$ has in ζ a zero of order $\geq n$. Put

$$A = |a_{p+1}| + \cdots + |a_{n-1}|. \quad (17.6)$$

In these notations, Theorem 17.1 will follow from:

Theorem 17.2. *The assertions of Theorem 17.1 are true for every $\varepsilon_0 > 0$ satisfying the conditions[†]*

$$\varepsilon_0 < \frac{r}{3}, \quad \varepsilon_0 \leq \frac{1}{3}, \quad \varepsilon_0 \leq \frac{|a_p|}{6A}, \quad \varepsilon_0^{n-p} \frac{K}{n!} \leq \frac{|a_p|}{7^n}. \quad (17.7)$$

[†] If $A = 0$, the third of the conditions (17.7) is to be disregarded.

3. Proof of Theorem 17.2. Without loss of generality assume $\zeta = 0$. $L(z)$ is obtained from $f(z)$, applying to $f(z)$ the operator $L_{n-1}(f, z)$ corresponding to the interpolation abscissas z_v , and given by formula (1B.10). We have in our notations

$$\begin{aligned} L_{n-1}(f, z) &= f(z_1) + (z - z_1)[z_1, z_2]f + (z - z_1)(z - z_2)[z_1, z_2, z_3]f \\ &\quad + \cdots + (z - z_1) \cdots (z - z_{n-1})[z_1, \dots, z_n]f. \end{aligned} \quad (17.8)$$

This is a linear operator applied to $f(z)$ and we have therefore from (17.5)

$$L(z) = L_{n-1}(T, z) + L_{n-1}(\varphi, z).$$

But $L_{n-1}(T, z)$ is a polynomial of degree $\leq n-1$ and since $T(z)$ itself is of degree $\leq n-1$, we have obviously $L_{n-1}(T, z) = T(z)$ and therefore

$$L(z) = T(z) + L^*(z), \quad L^*(z) = L_{n-1}(\varphi, z). \quad (17.9)$$

4. We have from (17.5)

$$\frac{T(z)}{z^p} = \sum_{v=1}^{v=p} a_v z^{v-p}, \quad \left| \frac{T(z)}{z^p} \right| \geq |a_p| - \sum_{v=p+1}^{n-1} |a_v| \cdot |z|^{v-p}.$$

Assuming now z lying in the disk (17.3), we have by (17.6), since $3\varepsilon_0 \leq 1$,

$$\left| \frac{T(z)}{z^p} \right| \geq |a_p| - 3\varepsilon A \geq |a_p| - \frac{|a_p|}{2} \geq \frac{|a_p|}{2} \quad (|z| \leq 3\varepsilon). \quad (17.10)$$

5. In order to apply to the decomposition (17.5) Rouché's theorem from Chapter 14, Section 5, we now have to obtain convenient upper bound for $L^*(z)$, for both $|z| = \varepsilon$ and $|z| = 3\varepsilon$.

Put, for $v = 0, 1, \dots$,

$$\max_{|z|=\varepsilon} |\varphi^{(v)}(z)| = M_v, \quad \max_{|z|=3\varepsilon} |\varphi^{(v)}(z)| = M'_v. \quad (17.11)$$

Then we have, replacing $f(z)$ by $\varphi(z)$ in (17.8) and estimating the divided differences by (1A.16),

$$|L^*(z)| \leq \sum_{v=0}^{n-1} \frac{(2\varepsilon)^v}{v!} M_v \quad (|z| \leq \varepsilon), \quad (17.12)$$

$$|L^*(z)| \leq \sum_{v=0}^{n-1} \frac{(4\varepsilon)^v}{v!} M'_v \quad (|z| \leq 3\varepsilon),$$

and we now have to obtain the estimates of M_v and M'_v .

6. Differentiating (17.5) v times, we have

$$f^{(v)}(z) = T^{(v)}(z) + \varphi^{(v)}(z),$$

where $\varphi^{(v)}(z)$ is the remainder of the Maclaurin development of $f^{(v)}(z)$ if we stop at z^{n-1-v} . But then we have for a positive $\delta < r$ and $|z| \leq \delta$

$$|\varphi^{(v)}(z)| \leq \frac{\delta^{n-v}}{(n-v)!} \max_{|z|=\delta} |(f^{(v)}(z))^{(n-v)}|.$$

From (17.4) we have, therefore, putting $\delta = \varepsilon$ and $\delta = 3\varepsilon$,

$$M_v \leq \frac{\varepsilon^{n-v}}{(n-v)!} K, \quad M_v' \leq \frac{(3\varepsilon)^{n-v}}{(n-v)!} K \quad (v = 0, 1, \dots, n-1). \quad (17.13)$$

7. Putting this into (17.12), we have for $|z| \leq \varepsilon$

$$|L^*(z)| \leq \sum_{v=0}^{n-1} \frac{(2\varepsilon)^v}{v!} \frac{\varepsilon^{n-v}}{(n-v)!} K = \frac{K}{n!} \varepsilon^n \sum_{v=0}^{n-1} \binom{n}{v} 2^v < \frac{(3\varepsilon)^n}{n!} K, \quad (17.14)$$

and for $|z| \leq 3\varepsilon$

$$|L^*(z)| \leq \sum_{v=0}^{n-1} \frac{(4\varepsilon)^v}{v!} \frac{(3\varepsilon)^{n-v}}{(n-v)!} K = \frac{K\varepsilon^n}{n!} \sum_{v=0}^{n-1} \binom{n}{v} 4^v 3^{n-v} < \frac{(7\varepsilon)^n}{n!} K. \quad (17.15)$$

If we use the last inequality (17.7), relations (17.14) and (17.15) become, since $n > p \geq 1$,

$$|L^*(z)| \leq 3^n \varepsilon^p \frac{|a_p|}{7^n} < \frac{1}{2^{n-1}} \left(\frac{|a_p|}{2} \varepsilon^p \right) \leq \frac{|a_p|}{2} \varepsilon^p \quad (|z| = \varepsilon), \quad (17.16)$$

$$|L^*(z)| \leq 7^n \varepsilon^p \frac{|a_p|}{7^n} = 2 \left(\frac{1}{3} \right)^p \left(\frac{|a_p|}{2} (3\varepsilon)^p \right) < \frac{|a_p|}{2} (3\varepsilon)^p \quad (|z| = 3\varepsilon). \quad (17.17)$$

8. For $|z| = \varepsilon$ it follows now from (17.10) and (17.16) that

$$|T(z)| \geq \frac{|a_p|}{2} \varepsilon^p > |L^*(z)|,$$

and similarly for $|z| = 3\varepsilon$ from (17.10) and (17.17)

$$|T(z)| \geq \frac{|a_p|}{2} (3\varepsilon)^p > |L^*(z)|.$$

Rouché's theorem can therefore be applied to the decomposition (17.9) on both contours $|z| = \varepsilon$ and $|z| = 3\varepsilon$. We see that $L(z)$ has in the interior of (17.2) and in (17.3) the same number of zeros as $T(z)$; but by (17.10) $T(z)$ has on the whole circle (17.3) exactly p zeros which are all concentrated in $z = 0$. The assertions of Theorems 17.1 and 17.2 now follow immediately.

9. Put, under the hypotheses of Theorem 17.1,

$$K_{p+1} = \max_{|z-\zeta|=r} |f^{(p+1)}(z)|, \quad K_{n+1} = \max_{|z-\zeta|=r} |f^{(n+1)}(z)|, \quad (17.18)$$

$$\gamma = \frac{p! f^{(n)}(\zeta)}{n! f^{(p)}(\zeta)}, \quad (17.19)$$

$$C_1 = \frac{K_{p+1}}{(p+1)|f^{(p)}(\zeta)|}, \quad C_2 = \frac{2K_{n+1}}{|f^{(n)}(\zeta)|}, \quad (17.20)$$

where we assume that $f^{(n)}(\zeta) \neq 0$.

Theorem 17.3. Assume in the hypotheses of Theorems 17.1 and 17.2 and in the notations of (17.18)–(17.20) that $f^{(n)}(\zeta) \neq 0$ and

$$\varepsilon C_1 < 1. \quad (17.21)$$

Then we have, denoting any zero of $L(z)$ situated in (17.2) by z_{n+1} ,

$$(z_{n+1} - \zeta)^p = \gamma \prod_{v=1}^n (z_{n+1} - z_v) \frac{1 + \theta_2 \varepsilon C_2}{1 + \theta_1 \varepsilon C_1}, \quad |\theta_1| \leq 1, \quad |\theta_2| \leq 1. \quad (17.22)$$

10. Proof. Without loss of generality we can and will assume that $\zeta = 0$. Developing $f(z_{n+1})$ in powers of z_{n+1} and stopping at the first non-vanishing term, we have

$$f(z_{n+1}) = \frac{f^{(p)}(0)}{p!} z_{n+1}^p + \theta \frac{z_{n+1}^{p+1}}{(p+1)!} f^{(p+1)}(\xi), \quad |\theta| \leq 1, \quad |\xi - \zeta| \leq \varepsilon.$$

Using the definitions (17.18) of K_{p+1} and (17.20) of C_1 , we have then

$$f(z_{n+1}) = a_p z_{n+1}^p (1 + \theta_1 \varepsilon C_1), \quad z_{n+1}^p = \frac{f(z_{n+1})}{a_p} \frac{1}{1 + \theta_1 \varepsilon C_1}, \quad |\theta_1| \leq 1, \quad (17.23)$$

where a_p is given by (17.5).

11. On the other hand, we have $f(z) = L(z) + R(z)$, where $R(z)$ is the remainder term of the interpolation formula, which can be obtained from (1B.14).

Since $L(z_{n+1}) = 0$, we have, using (1B.14),

$$f(z_{n+1}) = R(z_{n+1}) = \prod_{v=1}^n (z_{n+1} - z_v) \left(\frac{f^{(n)}(0)}{n!} + \theta_2 2\varepsilon \frac{K_{n+1}}{n!} \right), \quad |\theta_2| \leq 1.$$

Using the definition of C_2 in (17.20), we have

$$\frac{f(z_{n+1})}{\prod_{v=1}^n (z_{n+1} - z_v)} = \frac{f^{(n)}(0)}{n!} (1 + \theta_2 \varepsilon C_2). \quad (17.24)$$

From (17.23) and (17.24) the assertion (17.22) follows immediately, using (17.19) and (17.5).

18

Approximation of Equations by Algebraic Equations of a Given Degree. Asymptotic Errors for Simple Roots

CONVERGENCE OF ZEROS OF INTERPOLATING POLYNOMIALS

1. Under the conditions of Theorems 17.1 and 17.2, denote by z_{n+1} that root of $L(z) = 0$ which lies in (17.2) and is the nearest (or one of the nearest) to z_n . In the same way, starting from z_2, \dots, z_{n+1} , we can obtain a further number z_{n+2} of the sequence, and proceeding in the same way indefinitely, obtain for any $v = 0, 1, \dots$, from z_{v+1}, \dots, z_{v+n} the z_{v+n+1} . Using the notation (17.19) and (17.20) put

$$C := [1 + 2^{n+1}|\gamma|(1+rC_2)]^{1/(n-p)} > 1. \quad (18.1)$$

We will now prove:

Theorem 18.1. *Under the conditions of Theorems 17.1–17.3, and if beyond them we assume that*

$$\delta_0 := C\varepsilon < 1, \quad C_1\varepsilon \leq \frac{1}{2}, \quad (18.2)$$

the sequence z_v can be formed indefinitely and is convergent to ζ in such a way that

$$\sqrt[n]{|z_v - \zeta|} \rightarrow 0 \quad (v \rightarrow \infty). \quad (18.3)$$

2. Proof. Put

$$\xi_v = C|z_v - \zeta| \quad (v = 1, 2, \dots). \quad (18.4)$$

Without loss of generality we can and will assume $\zeta = 0$.

From (17.22) and (17.7) we have then

$$|z_{n+1}|^p < |\gamma|(2\varepsilon)^n \frac{1+rC_2}{\frac{1}{2}} < C^{n-p} \cdot \varepsilon^n = C^{-p}\delta_0^n,$$

and by (18.4)

$$\xi_{n+1} \leq \delta_0^{n/p} < \delta_0.$$

Applying the same argument to $\xi_{n+2}, \dots, \xi_{2n}$, we obtain

$$\xi_{n+1}, \quad \xi_{n+2}, \dots, \xi_{2n} \leq \delta_0^{n/p}. \quad (18.5)$$

3. The corresponding z_{n+1}, \dots, z_{2n} have the distances from the origin

$$\leq C^{-1} \delta_0^{n/p} = \varepsilon \delta_0^{(n-p)/p} =: \varepsilon_1. \quad (18.6)$$

Starting from z_{n+1}, \dots, z_{2n} , we can therefore replace ε by ε_1 and obtain from (18.6), putting

$$\delta_1 := C\varepsilon_1 = CC^{-1}\delta_0^{n/p} = \delta_0^{n/p}; \quad (18.7)$$

$$\xi_{2n+1}, \dots, \xi_{3n} \leq \delta_1^{n/p} = \delta_0^{(n/p)^2}. \quad (18.8)$$

Proceeding in the same way we obtain for every $k = 0, 1, \dots$

$$\xi_{kn+1}, \dots, \xi_{(k+1)n} \leq \delta_0^{(n/p)^k} \quad (k = 0, 1, \dots). \quad (18.9)$$

Since $\delta_0 < 1$ and $n/p > 1$, we see that indeed $\xi_v \rightarrow 0$, $z_v \rightarrow 0$. Further, for $v \rightarrow \infty$, $v = kn + \kappa$, $\kappa = 1, 2, \dots, n$,

$$\frac{1}{v} \ln \xi_v \leq \frac{1}{(k+1)n} \left(\frac{n}{p}\right)^k \ln \delta_0 \rightarrow -\infty, \quad \sqrt[p]{Cz_v} \rightarrow 0,$$

and (18.3) is proved.

ASYMPTOTIC ERRORS FOR SIMPLE ZEROS

4. To obtain a basic formula for the asymptotic error analysis, we return to (17.22) and take the moduli on both sides. Then we get

$$|z_{n+1} - \zeta|^p \leq |\gamma| \prod_{v=1}^n |z_{v+1} - z_v| \frac{1+rC_2}{\frac{1}{2}} < 2^{-n} C^{n-p} \prod_{v=1}^n |z_{v+1} - z_v|.$$

Multiplying this on both sides by C^p and using (18.4), we get

$$\xi_{n+1}^p \leq 2^{-n} \prod_{v=1}^n (\xi_{v+1} + \xi_v). \quad (18.10)$$

We will assume from now on that none of the ξ_v is 0.

5. For $p > 1$ further discussion presents singular difficulties. We deal with this case in Appendix P and assume in this chapter from now on that $p = 1$, that is, that ζ is a simple zero of $f(z)$.

Put $\text{Min}(\xi_1, \dots, \xi_n) = u$. If we had $\xi_{n+1} \geq u$, then one of the factors in the product in (18.10) would be $\leq 2\xi_{n+1}$ and we would have from (18.10) for $p = 1$

$$\xi_{n+1} \leq 2^{-n} 2\xi_{n+1} (2\varepsilon)^{n-1} = \xi_{n+1} \varepsilon^{n-1}.$$

But then it would follow that $\varepsilon \geq 1$, contrary to (17.7). We have therefore $\xi_{n+1} < u$ and (18.10) gives

$$\xi_{n+1} \leq 2^{-n} \prod_{v=1}^n (2\xi_v) = \prod_{v=1}^n \xi_v. \quad (18.11)$$

Since the ξ_v tend to 0 and are $< \delta_0 < 1$ we see that, for $v \geq 1$,

$$\xi_{v+n} \leq \xi_{v+n-1} \xi_{v+n-2} \cdots \xi_v \quad (v \geq 1), \quad (18.12)$$

$$\xi_{v+1} < \xi_v \quad (v \geq n), \quad \frac{\xi_{v+1}}{\xi_v} \rightarrow 0 \quad (v \rightarrow \infty). \quad (18.13)$$

6. Assume now $\zeta = 0$ and apply (17.22) with $p = 1$, $\zeta_i = 0$ to $z_v, z_{v+1}, \dots, z_{v+n}$, $v \geq n$. Since we can, by (18.13), replace here ε by $|z_v|$, we have, for sufficiently large v ,

$$\begin{aligned} z_{v+n} &= \gamma \prod_{\kappa=0}^{n-1} (z_{v+\kappa} - z_{v+\kappa}) \frac{1 + \theta_2 |z_v| C_2}{1 + \theta_1 |z_v| C_1}, \\ \frac{z_{v+n}}{\prod_{\kappa=0}^{n-1} z_{v+\kappa}} &= (-1)^n \gamma \prod_{\kappa=0}^{n-1} \left(1 - \frac{z_{v+\kappa}}{z_{v+\kappa}} \right) \frac{1 + \theta_2 |z_v| C_2}{1 + \theta_1 |z_v| C_1}. \end{aligned} \quad (18.14)$$

By (18.12) and (18.13) we have for the single factors in the right-hand product in (18.14)

$$\left| \frac{z_{v+\kappa}}{z_{v+\kappa}} \right| < |z_v|, \quad 1 - \frac{z_{v+\kappa}}{z_{v+\kappa}} = 1 + O(z_v).$$

We have therefore from (18.14)

$$z_{v+n} = (-1)^n (\gamma + O(z_v)) \prod_{\kappa=0}^{n-1} z_{v+\kappa}. \quad (18.15)$$

7. Take the moduli on both sides of (18.15) and put

$$\ln \frac{1}{|z_v|} =: y_v \quad (v \geq 1), \quad \ln |\gamma| =: (n-1)\beta. \quad (18.16)$$

Then we obtain

$$y_{v+n} - \sum_{\kappa=0}^{n-1} y_{v+\kappa} = (1-n)\beta + k_{v+n}, \quad k_{v+n} = O(z_v). \quad (18.17)$$

This is, however, the difference equation (13.33) and the condition (12.12) is satisfied for every $s > 0$ by virtue of (18.3).

We have, therefore, exactly as in Chapter 13,

$$y_v = \beta + \alpha \mu_n^v + O(v^{n-1} q_n^v)$$

and, going back to (18.16) and replacing z_v by $z_v - \zeta$, similarly to (13.36) and (13.37),

$$|z_v - \zeta| = \exp(-\beta) \exp(-\alpha \mu_n v) (1 + O(v^{n-1} q_n^v)), \quad (18.18)$$

$$\frac{z_{v+1} - \zeta}{|z_v - \zeta|^{\mu_n}} = \exp[(\mu_n - 1)\beta] + O(v^{n-1} q_n^v), \quad (18.19)$$

where μ_n and q_n are the numbers defined in Chapter 13 and satisfy (13.38) and β is given by (18.16) and (17.19).

8. These asymptotic results show that the convergence of the z_v in our case is of the same type as that of the x_v in Chapter 13. But here we have to solve an algebraic equation of degree n at every step, while in Chapter 13 each new approximation is obtained by rational operations. On the other hand, numerical experience appears to show that the procedure dealt with in this chapter is much less sensitive with respect to the choice of initial values z_1, \dots, z_n .

19

Norms of Vectors and Matrices

VECTOR NORMS

1. Let ξ be a *row vector* or a *point*[†] where for real or complex x_v ,

$$\xi = (x_1, \dots, x_n). \quad (19.1)$$

We define the “ p norm” of ξ as

$$|\xi|_p := (|x_1|^p + \dots + |x_n|^p)^{1/p} \quad (p \geq 1). \quad (19.2)$$

For $p \rightarrow \infty$, the largest term in parentheses in (19.2) will dominate. Now suppose $|x_1| \geq |x_2| \geq \dots \geq |x_n|$. Then

$$(n|x_1|^p)^{1/p} \geq |\xi|_p \geq (|x_1|^p)^{1/p}, \quad \lim_{p \rightarrow \infty} (n|x_1|^p)^{1/p} = |x_1|.$$

Hence we have as the convenient definition of $|\xi|_\infty$:

$$|\xi|_\infty := \max_v |x_v|. \quad (19.3)$$

The values of $p = 1$ and $p = \infty$ are particularly important in numerical analysis. It turns out on the average to be more convenient to use $p = \infty$ in the theory of *convergence*, while $p = 1$ is apparently the best norm in the study of *divergence*.

To any $p \geq 1$ there corresponds a q , so that

$$\frac{1}{p} + \frac{1}{q} = 1, \quad q = \frac{p}{p-1}, \quad \frac{p}{q} = p-1, \quad p \geq 1, \quad q \geq 1. \quad (19.4)$$

However, for $p = 1$ the corresponding q is assumed as ∞ and we also consider the case $p = \infty$, $q = 1$.

Then the so-called Hölder's inequality can be written as an inequality between the components of the vectors (19.1) and $\eta := (y_1, \dots, y_n)$,

$$\left| \sum_{v=1}^n x_v y_v \right| \leq |\xi|_p |\eta|_q. \quad (19.5)$$

[†] Both names will be used in the following discussion.

This inequality is proved in the calculus for $1 < p < \infty$; it is immediately verified, however, if one of the numbers p, q becomes 1.

2. Clearly $|c\xi|_p = |c||\xi|_p$ for any constant c . The so-called “triangle inequality” for the p norms,

$$|\xi + \eta|_p \leq |\xi|_p + |\eta|_p \quad (p \geq 1), \quad (19.6)$$

has been proved by Minkowski and is often called the “Minkowski inequality.”

To prove (19.6), observe first that if we replace the components of $|\xi|$ and $|\eta|$ by their moduli, the right-hand sum in (19.6) is not changed, while the left-hand expression is not decreased. Therefore it is sufficient to prove (19.6) under the assumption that all x_v, y_v are ≥ 0 . Further, the inequality (19.6) follows immediately for $p = 1$, and from (19.3) for $p = \infty$. We can therefore assume that $1 < p < \infty$; but then we have by (19.5) applied to p and q in (19.4):

$$\sum_{v=1}^n x_v(x_v + y_v)^{p-1} \leq |\xi|_p \left(\sum_{v=1}^n (x_v + y_v)^{(p-1)q} \right)^{1/q} = |\xi|_p |\xi + \eta|_p^{p-1}.$$

By symmetry we also have

$$\sum_{v=1}^n y_v(x_v + y_v)^{p-1} \leq |\eta|_p |\xi + \eta|_p^{p-1}.$$

Adding the two inequalities, we obtain

$$\sum_{v=1}^n (x_v + y_v)(x_v + y_v)^{p-1} = |\xi + \eta|_p^p \leq (|\xi|_p + |\eta|_p)|\xi + \eta|_p^{p-1},$$

and (19.6) follows immediately.

From (19.6) it follows further, decomposing ξ as $\xi = (\xi + \eta) + (-\eta)$, that

$$|\xi|_p \leq |\xi + \eta|_p + |\eta|_p, \quad |\xi + \eta|_p \geq |\xi|_p - |\eta|_p.$$

Interchanging here ξ and η , we have

$$|\eta + \xi|_p = |\xi + \eta|_p \geq |\eta|_p - |\xi|_p;$$

therefore

$$|\xi + \eta|_p \geq ||\xi|_p - |\eta|_p|.$$

If $\xi = (x_v)$ and $\eta = (y_v)$ are two n -dimensional vectors, we call the expression

$$\bar{x}_1 y_1 + \cdots + \bar{x}_n y_n = (\xi, \eta),$$

the inner product of ξ and η (in this order) and denote it by the symbol (ξ, η) or also by the symbol $\xi_l \eta_c$, where ξ_l is the row vector corresponding to ξ and η_c

the *column* vector corresponding to η . Then the so-called *Cauchy-Schwarz inequality* can be written as

$$|(\xi, \eta)| \leq |\xi|_2 |\eta|_2. \quad (19.7)$$

MATRIX NORMS $|A|_1$ AND $|A|_\infty$

3. Let $A = (a_{ij})$ denote an $n \times n$ matrix. We use $\det A$ to denote the determinant of A and ξ' to denote the *transpose* of ξ , i.e., the corresponding column vector. The matrix obtained from A by interchanging the rows with the columns is called the *transpose of A* and will be denoted A' . We write

$$A\xi' = \eta', \quad (19.8)$$

where the components of $\eta = (y_1, \dots, y_n)$ are given by

$$y_i = \sum_{j=1}^n a_{ij} x_j \quad (i = 1, \dots, n). \quad (19.9)$$

From (19.9) we have

$$|y_i| \leq \sum_{j=1}^n |a_{ij}| |x_j| \leq \sum_{j=1}^n |a_{ij}| |\xi|_\infty \quad (i = 1, \dots, n). \quad (19.10)$$

If we introduce a measure of the “size of A ” by

$$|A|_\infty := \max_i \sum_{j=1}^n |a_{ij}|, \quad (19.11)$$

we have from (19.10)

$$|y_i| \leq |A|_\infty |\xi|_\infty \quad (i = 1, \dots, n),$$

and therefore

$$|\eta|_\infty \leq |A|_\infty |\xi|_\infty. \quad (19.12)$$

4. We now show that in (19.12) $|A|_\infty$ cannot be replaced by a smaller constant; i.e., for each A it is possible to construct a ξ so that the equality in (19.12) holds. Without loss of generality, let $A \neq 0$. Let

$$|A|_\infty = |a_{m1}| + |a_{m2}| + \cdots + |a_{mn}|.$$

We then define ξ by

$$x_v = \begin{cases} |a_{mv}|/a_{mv}, & a_{mv} \neq 0 \\ 0, & a_{mv} = 0 \end{cases} \quad (v = 1, \dots, n). \quad (19.13)$$

Then from (19.9)

$$y_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = \sum_{v=1}^n |a_{mv}| = |A|_\infty$$

and hence

$$|\eta|_\infty \geq |A|_\infty. \quad (19.14)$$

Now, not all x_v ($v = 1, \dots, n$) are zero, for otherwise $A = 0$. Hence $|\xi|_\infty = 1$ and from (19.12) and (19.14) we have $|\eta|_\infty = |A|_\infty |\xi|_\infty$.

5. In order to obtain the corresponding relations for $p = 1$, observe that from (19.9) it follows that

$$|y_i| \leq \sum_{j=1}^n |a_{ij}| |x_j| \quad (i = 1, \dots, n),$$

and by summation

$$|\eta|_1 \leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}| |x_j| = \sum_{j=1}^n \left(\sum_{i=1}^n |a_{ij}| \right) |x_j|. \quad (19.15)$$

Putting

$$t_j = \sum_{i=1}^n |a_{ij}| \quad (j = 1, \dots, n),$$

we can write (19.15) in the form

$$|\eta|_1 \leq \sum_{j=1}^n t_j |x_j|. \quad (19.16)$$

If we now introduce another measure of the “size of A ” by

$$|A|_1 := \max_j \sum_{i=1}^n |a_{ij}| = \max_j t_j, \quad (19.17)$$

we have

$$\begin{aligned} |\eta|_1 &\leq |A|_1 \sum_{j=1}^n |x_j|, \\ |\eta|_1 &\leq |A|_1 |\xi|_1. \end{aligned} \quad (19.18)$$

As before, we show that in (19.18) $|A|_1$ cannot be replaced by a smaller bound. Indeed, assume $A \neq 0$ and suppose that a maximum of the t_j is assumed for $j = m$, i.e.,

$$|A|_1 = t_m.$$

We take a ξ with

$$x_m = 1, \quad x_i = 0 \quad (i \neq m).$$

Then

$$y_j = \sum_{i=1}^n a_{ij} x_i = a_{im} \quad (i = 1, \dots, n),$$

$$|y_i| = |a_{im}|,$$

$$|\eta|_1 = \sum_{i=1}^n |y_i| = \sum_{i=1}^n |a_{im}| = t_m = |A|_1,$$

and since $|\xi|_1 = 1$, we have indeed

$$|\eta|_1 = |A|_1 |\xi|_1.$$

We see in particular that

$$|A|_\infty = |A'|_1.$$

$|A|_p$ ($p = 1, \infty$) is called the *norm of the matrix A, induced by the vector norm $|\xi|_p$* .

6. We now deduce some properties of $|A|_p$ ($p = 1, \infty$). Clearly, for any constant c ,

$$|cA|_p = |c||A|_p \quad (p = 1, \infty). \quad (19.19)$$

Let $B = (b_{ij})$ be another $n \times n$ matrix; then $A + B = (a_{ij} + b_{ij})$. Put $\eta_1' = A\xi'$, $\eta_2' = B\xi'$. From (19.12) and (19.18) we have

$$\begin{aligned} |\eta_1|_p &\leq |A|_p |\xi|_p, & |\eta_2|_p &\leq |B|_p |\xi|_p \quad (p = 1, \infty), \\ \eta_1' + \eta_2' &= (A + B)\xi', \\ |\eta_1 + \eta_2|_p &\leq |A + B|_p |\xi|_p \quad (p = 1, \infty). \end{aligned} \quad (19.20)$$

We choose $\xi \neq 0$ so that equality holds in (19.20). Then from (19.6) we have

$$|A + B|_p |\xi|_p = |\eta_1 + \eta_2|_p \leq |\eta_1|_p + |\eta_2|_p \leq (|A|_p + |B|_p) |\xi|_p$$

and since $|\xi|_p \neq 0$, we have

$$|A + B|_p \leq |A|_p + |B|_p \quad (p = 1, \infty). \quad (19.21)$$

Let $\eta' = AB\xi'$. Then $|\eta|_p \leq |A|_p |B\xi'|_p \leq |A|_p |B|_p |\xi|_p$ and therefore, if we choose ξ such that $|\eta|_p = |AB|_p |\xi|_p$,

$$|AB|_p \leq |A|_p |B|_p \quad (p = 1, \infty). \quad (19.22)$$

EIGENVALUES OF A

7. We use I to denote the unity matrix, which consists of *ones* down the main diagonal and zeros elsewhere. Clearly $AI = IA$. The roots λ_v ($v = 1, \dots, n$) of

$$\det(\lambda I - A) = 0 \quad (19.23)$$

are called the *fundamental* (or *characteristic*) *roots* or *eigenvalues* of A and correspondingly (19.23) is called the *fundamental* (or *characteristic*) *equation* of A . We introduce the following notation:

$$\lambda_A = \max_v |\lambda_v| \quad (v = 1, \dots, n). \quad (19.24)$$

From (19.23) follows for $\lambda = 0$:

$$\det A = \lambda_1 \lambda_2 \cdots \lambda_n. \quad (19.25)$$

Notice that if λ_v is a root of (19.23), then we can find a vector $\xi_v \neq 0$ which is a solution of the system

$$(\lambda_v I - A) \xi'_v = 0, \quad \text{i.e., } A \xi'_v = \lambda_v \xi'_v. \quad (19.26)$$

ξ_v is called a *fundamental* (or *characteristic* or *eigen-*) *vector* corresponding to λ_v . If A is replaced by cA , each characteristic root of A is multiplied by c , while the corresponding characteristic vectors remain the same.

Theorem 19.1. *For any $n \times n$ matrix A we have*

$$\lambda_A \leq |A|_p \quad (p = 1, \infty). \quad (19.27)$$

Proof. Let m be such that $|\lambda_m| = \lambda_A$ and let ξ_m be a fundamental vector corresponding to λ_m . Then

$$A \xi'_m = \lambda_m \xi'_m, \quad |\lambda_m| |\xi_m|_p \leq |A|_p |\xi_m|_p,$$

and, since $|\xi_m|_p \neq 0$, $\lambda_A = |\lambda_m| \leq |A|_p$, Q.E.D.

8. Let S be a *regular* $n \times n$ matrix, i.e., one such that $\det S \neq 0$. The *transform* of a matrix A by S is defined as SAS^{-1} .

Theorem 19.2. *A matrix A and its transforms have the same fundamental roots.*

Indeed, we have $(\lambda I - SAS^{-1}) = S(\lambda I - A)S^{-1}$,

$$\det(\lambda I - SAS^{-1}) = \det(S(\lambda I - A)S^{-1}) = \det(\lambda I - A).$$

Remark. If we transform A by S , λ_A does not change, but $|A|_p$ may change.

Corollary. *If A, B are two $n \times n$ matrices, AB and BA have the same fundamental equations.*

This follows, if $\det A \neq 0$, immediately from the identity $BA = A^{-1}(AB)A$. If $\det A = 0$, A goes by arbitrary small variation of its elements into a non-singular matrix, and the assertion follows by continuity, since the left-hand side of (19.23) is a polynomial in λ and in the elements of A .

9. We now state a well-known result of C. Jordan (Jordan's canonical form): *Given an $n \times n$ matrix A , there exists an S such that*

$$SAS^{-1} = D + J, \quad (19.28)$$

where D is a diagonal matrix whose elements are the n fundamental roots of A , and J is a matrix with zeros and ones along the first superdiagonal (the diagonal parallel to the principal diagonal and above it) and zeros everywhere else. More precisely, (19.28) can be written as

$$SAS^{-1} = \begin{pmatrix} U_1 & & & 0 \\ & U_2 & & \\ & & \ddots & \\ 0 & & & U_m \end{pmatrix}, \quad (19.29)$$

$$U_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}. \quad (19.30)$$

If λ_i is a simple root, then the matrix U_i is of the order *one*. If the fundamental roots are all distinct, then all U_i are of order 1, J vanishes, and A in (19.28) is reduced to the "diagonal form." This may also happen if the fundamental roots are multiple.

10. We have obviously

$$|J|_p \leq 1 \quad (p = 1, \infty). \quad (19.31)$$

For any $\varepsilon \neq 0$ it follows from (19.28), applied to $(1/\varepsilon) A$, that

$$\frac{1}{\varepsilon} AS^{-1} = \frac{1}{\varepsilon} D + J, \quad SAS^{-1} = D + \varepsilon J. \quad (19.32)$$

Hence we see that in Jordan's canonical form the value *one* of the non-vanishing elements of J is not essential; it can be replaced by any number $\varepsilon \neq 0$. Now by (19.31)

$$|\varepsilon J|_p = |\varepsilon| |J|_p \leq |\varepsilon|. \quad (19.33)$$

Theorem 19.3. *Given an $\varepsilon > 0$, there exists an S such that*

$$\lambda_A \leq |SAS^{-1}|_p \leq \lambda_A + \varepsilon \quad (p = 1, \infty). \quad (19.34)$$

Indeed, from Theorem 19.1 and (19.32) we have

$$\lambda_A \leq |SAS^{-1}|_p \leq |D|_p + |\epsilon J|_p \leq \lambda_A + \epsilon.$$

As the fundamental roots of the matrix A are roots of the algebraic equation (19.23), the results of Appendices A and B can be applied to them. However, the direct computation of the coefficients of the fundamental equation presents considerable numerical difficulties. It is therefore important to have estimates for the variation of the roots of (19.23) in terms of the variations of the elements of A . We give such estimates in Appendix K.

11. If Eq. (19.23) is developed in powers of λ , we see easily that the coefficient of λ^{n-1} is $-(a_{11} + a_{22} + \dots + a_{nn})$. It follows therefore that

$$\sum_{v=1}^n \lambda_v = \sum_{v=1}^n a_{vv}. \quad (19.35)$$

The expression on the right is called the *trace* of the matrix A and is sometimes denoted by $\text{tr}(A)$.

From the corollary to Theorem 19.2 follows that under its hypotheses

$$\text{tr}(AB) = \text{tr}(BA). \quad (19.36)$$

As we can add and multiply $(n \times n)$ matrices, we can form powers of a given matrix A, A^2, A^3, \dots and more generally, if $P(z) = \sum_{v=1}^k \alpha_v z^v$ is a polynomial of the degree k , we can form the matrix

$$P(A) = \sum_{v=1}^k \alpha_v A^v.$$

More generally, assume that $R(z) = P_1(z)/P_2(z)$ is a rational function represented as the quotient of two polynomials $P_1(z), P_2(z)$. We have by definition

$$R(A) := P_1(A) P_2(A)^{-1},$$

provided the matrix $P_2(A)$ is not singular. It is easily seen that $R(A)$ is independent of the particular representation of $R(z)$ as the quotient of two polynomials.

Later we will have to use the following well-known theorem:

If $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A , then for any rational function $\varphi(z)$, the eigenvalues of $\varphi(A)$ are given by the expressions

$$\varphi(\lambda_1), \varphi(\lambda_2), \dots, \varphi(\lambda_n),$$

provided all these numbers are finite, that is, that the denominator of $\varphi(z)$ does not vanish in any of the points $\lambda_1, \dots, \lambda_n$.

In particular the eigenvalues of A^2 are λ_v^2 , and those of $A^{-1}, 1/\lambda_v$.

20

Two Theorems on Convergence of Products of Matrices

1. Theorem 20.1. Let A be an $n \times n$ matrix and $\varepsilon > 0$. There exist two positive constants $\eta_1 > 0$ and $\sigma > 0$ depending only on A and ε , such that if for a sequence of $n \times n$ matrices U_μ with

$$|U_\mu|_\infty \leq \eta_1 \quad (\mu = 1, 2, \dots) \quad (20.1)$$

we form

$$\prod_m := \prod_{\mu=1}^m (A + U_\mu), \quad (20.2)$$

then we have, irrespective of the order of the factors in (20.2)

$$\left| \prod_m \right|_\infty \leq \sigma (\lambda_A + \varepsilon)^m \quad (m = 1, 2, \dots). \quad (20.3)$$

If in particular $\lambda_A + \varepsilon < 1$, we have

$$\prod_m \rightarrow 0 \quad (m \rightarrow \infty). \quad (20.4)$$

2. Proof. Find an $n \times n$ matrix S for which by Theorem 19.3 the following relation holds:

$$|SAS^{-1}|_\infty \equiv s < \lambda_A + \frac{\varepsilon}{2}. \quad (20.5)$$

Define σ and η_1 by

$$\sigma = |S|_\infty |S^{-1}|_\infty, \quad \eta_1 = \varepsilon/2\sigma, \quad (20.6)$$

and put

$$B = SAS^{-1}, \quad V_\mu = SU_\mu S^{-1} \quad (\mu = 1, 2, \dots). \quad (20.7)$$

Then from (20.2) we have

$$\begin{aligned} S \prod_m S^{-1} &= [S(A+U_1)S^{-1}] \cdots [S(A+U_m)S^{-1}] \\ &= [SAS^{-1} + SU_1 S^{-1}][SAS^{-1} + SU_2 S^{-1}] \cdots [SAS^{-1} + SU_m S^{-1}], \end{aligned}$$

$$S \prod_m S^{-1} = \prod_{\mu=1}^m (B + V_\mu). \quad (20.8)$$

From (20.1) and (20.5)–(20.8) it follows by (19.22) that

$$\begin{aligned} |V_\mu|_\infty &\leq \sigma\eta_1 = \frac{\varepsilon}{2}, & |B+V_\mu|_\infty &\leq s + \sigma\eta_1 < \lambda_A + \varepsilon, \\ \left| S \prod_m S^{-1} \right|_\infty &\leq (\lambda_A + \varepsilon)^m. \end{aligned} \quad (20.9)$$

Hence, since $|\prod_m|_\infty = |S^{-1}(S \prod_m S^{-1})S|_\infty$, it follows from (20.9) and (20.6) that

$$\left| \prod_m \right|_\infty \leq \sigma(\lambda_A + \varepsilon)^m,$$

and this is (20.3). If $\lambda_A < 1$, then taking $\varepsilon > 0$ such that $\lambda_A + \varepsilon < 1$, (20.4) follows from (20.3).

3. Theorem 20.2. *Let A be an $n \times n$ matrix with $\lambda_A < 1$. Form recursively the matrices A_μ as follows:*

$$A_1 = A, \quad A_{\mu+1} = A_\mu A + W_\mu, \quad (20.10)$$

with $n \times n$ matrices W_μ . Let $\varepsilon > 0$ be such that $\lambda_A + \varepsilon < 1$. There exist two positive constants $\eta_2 > 0$, $\sigma > 0$ such that, if

$$|W_\mu|_\infty \leq \eta_2 |A_\mu|_\infty \quad (\mu = 1, 2, \dots), \quad (20.11)$$

then

$$|A_m|_\infty < \sigma(\lambda_A + \varepsilon)^m \quad (m = 1, 2, \dots), \quad A_m \rightarrow 0 \quad (m \rightarrow \infty). \quad (20.12)$$

4. Proof. The symbols S , s , B , and σ are defined as in the proof of Theorem 20.1. Then we define η_2 by

$$\eta_2 := \varepsilon/2\sigma^2. \quad (20.13)$$

If we introduce T_μ and B_μ by

$$T_\mu = SW_\mu S^{-1}, \quad B_\mu = SA_\mu S^{-1} \quad (\mu = 1, 2, \dots), \quad B_1 = B, \quad (20.14)$$

we have from (20.10)

$$SA_{\mu+1}S^{-1} = SA_\mu S^{-1}SAS^{-1} + SW_\mu S^{-1},$$

and consequently

$$B_{\mu+1} = B_\mu B + T_\mu \quad (\mu = 1, 2, \dots). \quad (20.15)$$

On the other hand, from (20.14)

$$|T_\mu|_\infty \leq |S|_\infty |S^{-1}|_\infty |W_\mu|_\infty$$

and by (20.6) and (20.11)

$$|T_\mu|_\infty \leq \sigma\eta_2 |A_\mu|_\infty \quad (\mu = 1, 2, \dots). \quad (20.16)$$

5. But from (20.14) we have also $A_\mu = S^{-1}B_\mu S$,

$$|A_\mu|_\infty \leq \sigma |B_\mu|_\infty \quad (\mu = 1, 2, \dots), \quad (20.17)$$

and since $|B|_\infty = s$, we have from (20.15)–(20.17)

$$|B_{\mu+1}|_\infty \leq s|B_\mu|_\infty + \sigma^2 \eta_2 |B_\mu|_\infty,$$

$$|B_{\mu+1}|_\infty \leq |B_\mu|_\infty \left(s + \frac{\varepsilon}{2} \right) \quad (\mu = 1, 2, \dots),$$

$$|B_{\mu+1}|_\infty \leq (\lambda_A + \varepsilon)|B_\mu|_\infty \leq (\lambda_A + \varepsilon)^\mu |B_1|_\infty \rightarrow 0 \quad (\mu \rightarrow \infty),$$

and by (20.5)–(20.7) and (20.17), finally (20.12).

6. Theorems 20.1 and 20.2 generalize the result that the μ th power of a matrix A goes to zero if $\lambda_A < 1$. In particular, the second theorem assures the theoretical *stability of the convergence* of A^μ to 0 with respect to rounding off.

21

A Theorem on Divergence of Products of Matrices

1. We now give a theorem corresponding to Theorem 20.1 for $\lambda_A > 1$. In its proof we use the norms with $p = 1$.

Theorem 21.1. *Let A be an $n \times n$ matrix with $\lambda_A > 1$ and $\varepsilon > 0$ such that $\lambda_A - \varepsilon > 1$. Form for a sequence of $n \times n$ matrices U_v ($v = 1, 2, \dots$)*

$$\prod_m := \prod_{v=1}^m (A + U_v). \quad (21.1)$$

There exists a $\delta = \delta(A, \varepsilon) > 0$ such that if $|U_v|_1 \leq \delta$ ($v = 1, 2, \dots$), then (21.1) diverges as $m \rightarrow \infty$; more precisely, there exists in this case a solid angle L with its vertex at the origin such that for any $\zeta \neq 0$ in L

$$(\lambda_A - \varepsilon)^{-m} \left| \prod_m \zeta \right|_1 \rightarrow \infty \quad (m \rightarrow \infty), \quad (21.2)$$

where the factors in the product \prod_m may be multiplied in any order.

2. Proof. We denote the n fundamental roots of A by $\lambda_1, \lambda_2, \dots, \lambda_n$ in such a way that

$$\lambda_A = |\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_k| = s > 1 \geq |\lambda_{k+1}| \geq \dots \geq |\lambda_n|.^\dagger \quad (21.3)$$

Put

$$\eta = \frac{s-1}{2n+1} \quad (21.4)$$

and take a τ with $0 < \tau < \eta$. Transform A by S into Jordan's canonical form as in (19.32):

$$SAS^{-1} =: B = D + \tau J. \quad (21.5)$$

If we introduce

$$\sigma := |S|_1 |S^{-1}|_1, \quad \delta := \frac{\eta - \tau}{\sigma} \quad (21.6)$$

† If $k = n$, the expression σ_m defined by (21.18) is = 0, but the whole discussion remains valid.

and put

$$V_v = SU_v S^{-1}, \quad W_v = V_v + \tau J, \quad (21.7)$$

we have from (21.5)

$$B + V_v = D + W_v. \quad (21.8)$$

3. But then it obviously follows that

$$\begin{aligned} S(A + U_v)S^{-1} &= SAS^{-1} + SU_v S^{-1} = B + V_v, \\ S\left[\prod_{v=1}^m (A + U_v)\right]S^{-1} &= \prod_{v=1}^m (D + W_v). \end{aligned} \quad (21.9)$$

Hence we have replaced A by the simpler matrix D . Now from (21.6) and the hypothesis $|U_v|_1 \leq \delta$ follows

$$|V_v|_1 \leq |S|_1 |U_v|_1 |S^{-1}|_1 \leq \sigma \delta = \eta - \tau$$

and from (21.7), as $|J|_1 = 1$,

$$|W_v|_1 \leq \eta. \quad (21.10)$$

On the other hand, for any vector ζ ,

$$S\left[\prod_{v=1}^m (A + U_v)\right]S^{-1}(S\zeta') = \left[\prod_{v=1}^m (D + W_v)\right]S\zeta', \quad (21.11)$$

and we shall prove that for each vector $\xi := \xi_0 := (x_1, \dots, x_n) \neq 0$ with

$$|x_1| + |x_2| + \dots + |x_k| \geq |x_{k+1}| + \dots + |x_n| \quad (21.12)$$

we get

$$\left|\prod_{v=1}^m (D + W_v) \xi_0'\right|_1 \rightarrow \infty \quad (m \rightarrow \infty). \quad (21.13)$$

4. Put

$$\xi_m' = \prod_{v=1}^m (D + W_v) \xi_0' \quad (m = 1, 2, \dots), \quad (21.14)$$

$$\xi_m = (x_1^{(m)}, x_2^{(m)}, \dots, x_n^{(m)}) \quad (m = 0, 1, \dots). \quad (21.15)$$

We have then

$$\xi_{m+1}' = (D + W_{m+1}) \xi_m' \quad (m = 0, 1, \dots). \quad (21.16)$$

Put now for $m = 0, 1, \dots$

$$\gamma_m = |x_1^{(m)}| + \dots + |x_k^{(m)}|, \quad (21.17)$$

$$\sigma_m = |x_{k+1}^{(m)}| + \dots + |x_n^{(m)}|, \quad (21.18)$$

$$|\xi_v|_1 = \gamma_v + \sigma_v \quad (v = 0, 1, \dots). \quad (21.19)$$

5. Now we have by (21.12)

$$\gamma_0 \geq \sigma_0. \quad (21.20)$$

Suppose that we have for an $m \geq 0$

$$\gamma_m \geq \sigma_m. \quad (21.21)$$

We are going to show that this implies $\gamma_{m+1} \geq \sigma_{m+1}$. Indeed, if we put $W_{m+1} = (w_{\mu\kappa})$, we have from (21.16)

$$x_\mu^{(m+1)} = \lambda_\mu x_\mu^{(m)} + \sum_{\kappa=1}^n w_{\mu\kappa} x_\kappa^{(m)}. \quad (21.22)$$

But from (21.10) it follows in particular that

$$|w_{\mu\kappa}| \leq \eta \quad (\mu, \kappa = 1, \dots, n)$$

and therefore by (21.17) and (21.18)

$$\left| \sum_{\kappa=1}^n w_{\mu\kappa} x_\kappa^{(m)} \right| \leq \eta(\gamma_m + \sigma_m);$$

i.e., if we introduce $\theta_{\mu, m}$ by

$$\theta_{\mu, m} = \frac{\sum_{\kappa=1}^n w_{\mu\kappa} x_\kappa^{(m)}}{\eta(\gamma_m + \sigma_m)},$$

$$x_\mu^{(m+1)} = \lambda_\mu x_\mu^{(m)} + \theta_{\mu, m} \eta(\gamma_m + \sigma_m), \quad |\theta_{\mu, m}| \leq 1, \quad (21.23)$$

$$|x_\mu^{(m+1)}| \geq |\lambda_\mu| |x_\mu^{(m)}| - \eta(\gamma_m + \sigma_m) \quad (\mu = 1, \dots, n). \quad (21.24)$$

6. From (21.3) and (21.21) we have now for $\mu = 1, \dots, k$

$$|x_\mu^{(m+1)}| \geq s |x_\mu^{(m)}| - 2\eta\gamma_m$$

and, if we sum for $\mu = 1, \dots, k$,

$$\gamma_{m+1} \geq s\gamma_m - 2k\eta\gamma_m = (s - 2k\eta)\gamma_m;$$

i.e., since, by (21.4), $s = 1 + (2n+1)\eta$,

$$\gamma_{m+1} \geq [1 + \eta + 2(n-k)\eta]\gamma_m. \quad (21.25)$$

On the other hand, it follows from (21.3), (21.23), and (21.21) for $\mu = k+1, \dots, n$ that

$$\begin{aligned} |x_\mu^{(m+1)}| &\leq |x_\mu^{(m)}| + 2\eta\gamma_m, \\ \sigma_{m+1} &\leq [1 + 2(n-k)\eta]\gamma_m. \end{aligned} \quad (21.26)$$

From (21.25) and (21.26) we have now $\gamma_{m+1} \geq \sigma_{m+1}$, and we see that (21.21) holds for all $m = 0, 1, \dots$.

7. Therefore, (21.25) holds also for all $m = 0, 1, \dots$. But from (21.25) it now follows that

$$\gamma_{m+1} \geq (1 + \eta)\gamma_m \quad (m = 0, 1, \dots),$$

and hence $\gamma_m \rightarrow \infty$ ($m \rightarrow \infty$),

$$|\xi_m|_1 \rightarrow \infty \quad (m \rightarrow \infty). \quad (21.27)$$

For our vector ξ the vector $\zeta' = S^{-1}\xi'$ satisfies the relation

$$\left| \prod_m \zeta' \right|_1 \rightarrow \infty$$

and ζ lies in the solid angle obtained from (2.12) by the transformation S^{-1} .

In order to prove the complete assertion (21.2), consider the matrices

$$C = \frac{1}{\lambda_A - \varepsilon} A, \quad X_v = \frac{1}{\lambda_A - \varepsilon} U_v.$$

Then

$$\lambda_C = \frac{\lambda_A}{\lambda_A - \varepsilon} > 1,$$

and we see that there exist a positive number δ_ε and a solid angle L_ε with its vertex in the origin such that as soon as $|X_v|_1 \leq \delta_\varepsilon$, we have for any vector ζ from L_ε

$$(\lambda_A - \varepsilon)^{-m} \left| \prod_m \zeta' \right|_1 = \left| \prod_{v=1}^m (C + X_v) \zeta' \right|_1 \rightarrow \infty,$$

and this holds as long as

$$|U_v|_1 \leq (\lambda_A - \varepsilon) \delta_\varepsilon = \delta(A, \varepsilon) \quad (v = 1, 2, \dots).$$

Our theorem is proved.

8. Corollary. *Under the conditions of Theorem 21.1, we have obviously*

$$(\lambda_A - \varepsilon)^{-m} \left| \prod_m \zeta' \right|_1 \rightarrow \infty \quad (m \rightarrow \infty). \quad (21.28)$$

22

Characterization of Points of Attraction and Repulsion for Iterations with Several Variables

POINTS OF ATTRACTION AND REPULSION

1. Let $\xi = (x_1, \dots, x_n)$ be a point in the n -dimensional real or complex space S and put

$$y_i = f_i(x_1, \dots, x_n) =: f_i(\xi) \quad (i = 1, \dots, n). \quad (22.1)$$

Using the vector notation we write (22.1) as

$$\eta = \Phi(\xi), \quad (22.2)$$

where

$$\eta = (y_1, \dots, y_n) = [f_1(\xi), \dots, f_n(\xi)]. \quad (22.3)$$

A point ζ is called a *fixed point* or a *center* of the transformation (22.2) if we have

$$\zeta = \Phi(\zeta). \quad (22.4)$$

Now let ξ_0 be an “initial approximation” to a fixed point ζ of (22.2); we obtain a sequence (ξ_k) of approximations to ζ by the iteration

$$\xi_1 = \Phi(\xi_0), \dots, \xi_{k+1} = \Phi(\xi_k), \dots \quad (22.5)$$

If there exists a neighborhood of ζ in S such that for any ξ_0 in this neighborhood the sequence (22.5) converges to ζ , ζ is called a *point of attraction*, otherwise a *point of repulsion*.

A function $f(\xi) = f(x_1, \dots, x_n)$ is called *totally differentiable* at the point $\zeta(z_1, \dots, z_n)$ if we have for n constants a_j depending on ζ

$$f(\xi) - f(\zeta) = \sum_{j=1}^n (a_j + u_j)(x_j - z_j), \quad (22.6)$$

where the u_j tend to 0 with $\xi \rightarrow \zeta$. The a_j are then the partial derivatives of f at ζ . For the total differentiability of $f(\xi)$ at ζ it is sufficient (but not necessary) that the first partial derivatives of f exist in a neighborhood of ζ and are continuous at ζ .

We denote by $J(\xi)$ the Jacobian matrix of (22.1) at ξ and in particular put

$$A := J(\xi) = \left(\frac{\partial(f_i(\xi))}{\partial(z_j)} \right). \quad (22.7)$$

2. Theorem 22.1. *Assume that $f_i(\xi)$ ($i = 1, \dots, n$) are totally differentiable at the center ξ . Further, assume that for the matrix (22.7) we have (cf. (19.24)).*

$$\lambda_A < 1. \quad (22.8)$$

Then ξ is a point of attraction, i.e., there exist two neighborhoods V, V_0 of ξ such that if $\xi_0 \in V$ and the sequence ξ_κ is obtained by (22.5), we have $\xi_\kappa \in V_0$ ($\kappa = 1, 2, \dots$), and

$$\xi_\kappa \rightarrow \xi \quad (\kappa \rightarrow \infty). \quad (22.9)$$

3. Proof. Without loss of generality, we can assume that ξ is the origin. Since then $f_i(0) = 0$ ($i = 1, \dots, n$), we have from (22.6) and (22.7)

$$y_i = f_i(\xi) = \sum_{j=1}^n (a_{ij} + u_{ij}(\xi)) x_j \quad (i = 1, 2, \dots, n), \quad (22.10)$$

where the $u_{ij}(\xi)$ tend to 0 with $\xi \rightarrow 0$. Introducing the matrices

$$U(\xi) = (u_{ij}(\xi)), \quad (22.11)$$

$$A(\xi) = A + U(\xi), \quad (22.12)$$

we can write (22.3)

$$\eta' = A(\xi) \xi' \quad (22.13)$$

and have

$$U(\xi) \rightarrow 0 \quad (\xi \rightarrow 0). \quad (22.14)$$

Choose $\varepsilon > 0$ such that $\lambda_A + \varepsilon < 1$. Define η_1 and σ corresponding to A according to Theorem 20.1 and choose a convex neighborhood V_0 of the origin defined by $|\xi|_\infty \leq \tau$ for a convenient $\tau > 0$, such that throughout V_0

$$|u_{ij}| < \frac{\eta_1}{n+1} \quad (i, j = 1, \dots, n). \quad (22.15)$$

We introduce a new neighborhood V of 0 by

$$V = \frac{1}{1+\sigma} V_0 \quad (22.16)$$

and assume that $\xi_0 \in V$, $\xi_1, \dots, \xi_k \in V_0$. Then we have by (22.5), (22.12), and (22.13)

$$\xi'_{k+1} = [A + U(\xi_k)] \xi'_k$$

and therefore

$$\xi'_{k+1} = \prod_{v=0}^k [A + U(\xi_v)] \xi_0'. \quad (22.17)$$

4. But from (22.15) it follows that

$$|U(\xi_v)|_\infty < \eta_1 \quad (v = 0, 1, \dots, k),$$

and Theorem 20.1 can be applied. Then we have from (20.3)

$$\left| \prod_{v=0}^k [A + U(\xi_v)] \right|_\infty \leq \sigma (\lambda_A + \varepsilon)^{k+1} < \sigma,$$

and consequently by (22.17)

$$|\xi'_{k+1}|_\infty \leq \sigma |\xi_0|_\infty \leq \sigma \tau. \quad (22.18)$$

Hence we have by (22.16)

$$\xi'_{k+1} \in \frac{\sigma}{1+\sigma} V_0 \in V_0;$$

our assumptions hold for ξ'_{k+1} too, and consequently for all ξ_v . It follows now that for all κ , if $\xi_0 \in V$,

$$\xi'_{\kappa+1} = \prod_{v=0}^\kappa [A + U(\xi_v)] \xi_0' \quad (\kappa = 0, 1, \dots),$$

and since $\lambda_A + \varepsilon < 1$, we have from (20.4)

$$\xi'_{\kappa+1} \rightarrow 0 \quad (\kappa \rightarrow \infty), \quad \text{Q.E.D.} \quad (22.19)$$

5. Theorem 22.2. Assume that $f_i(\xi)$ ($i = 1, 2, \dots, n$) are totally differentiable at the center ζ . Further assume that for the matrix (22.7) we have

$$\lambda_A > 1. \quad (22.20)$$

Then ζ is a point of repulsion; more precisely, there exists a neighborhood V of ζ and a solid angle L with its vertex at ζ such that for any starting point $\xi_0 \in L \cap V$ (region common to L and V) the sequence ξ_v defined by (22.5) has one of its elements either at ζ itself or outside V .

6. Proof. Without loss of generality, we can assume that ζ is the origin, Take $\varepsilon > 0$ with $\lambda_A - \varepsilon > 1$, $\delta > 0$ and the solid angle L with its vertex at the origin such that the assertion of Theorem 21.1 holds, and define a neighborhood V of the origin such that for any point $\xi \in V$ we have in notations of Section 3:

$$|u_{ij}(\xi)| < \frac{\delta}{n+1} \quad (i, j = 1, \dots, n) \quad (22.21)$$

Suppose now that for a point ξ_0 from $L \cap V$ all points ξ_v in (22.5) stay in V and that none of them lies in the origin. Using the notations (22.11) and (22.12) we have by (22.21)

$$|U(\xi_v)|_1 < \delta \quad (v = 0, 1, \dots),$$

and the impossibility is proved by Theorem 21.1, Q.E.D.

7. As in the case of one variable, since we do not know the point ζ , we are faced with the problem of verifying whether $\lambda_A < 1$. In practice, we usually have to prove that the inequality

$$\lambda_{J(\xi)} < 1 \quad (22.22)$$

holds throughout the considered region. This verification may present difficulties, since we have to deal with the roots of a polynomial of n th degree and these polynomials would have to be considered for every point of the region. However, many bounds have been derived for fundamental roots of matrices and can be used in this connection. The simplest are the bounds of Theorem 19.1. If we have, for example, for a positive constant $q < 1$,

$$Q_i := \sum_{j=1}^n \left| \frac{\partial f_i(\xi)}{\partial x_j} \right| \leq q \quad (i = 1, \dots, n) \quad (22.23)$$

for any ξ in a neighborhood of ζ , then it follows by Theorem 19.1 that for all ξ from this neighborhood

$$\lambda_{J(\xi)} \leq |J(\xi)|_\infty \leq q < 1. \quad (22.24)$$

AN EXAMPLE

8. Observe that in Theorem 22.2 the “repulsion effect” is only asserted for the *intersection* of the neighborhood V with a solid angle L . It could very well happen that if the sequence ξ_μ remains outside of L , it converges to ζ . We will illustrate this by an example.

Consider, z being the complex variable $x + iy$, the iteration in the complex plane by means of

$$z' = z - \frac{1}{2}\bar{z} - \frac{1}{2}|z|^2\bar{z}. \quad (22.25)$$

This is equivalent, writing $z' = x' + iy'$, to the system of two relations:

$$x' = \frac{x}{2}(1 - (x^2 + y^2)), \quad y' = \frac{y}{2}(3 + (x^2 + y^2)). \quad (22.26)$$

The Jacobian matrix of the right-hand expressions is

$$\begin{pmatrix} \frac{1}{2} - \frac{3}{2}x^2 - \frac{1}{2}y^2 & -xy \\ xy & \frac{3}{2} + \frac{1}{2}x^2 + \frac{3}{2}y^2 \end{pmatrix}. \quad (22.27)$$

In the point $\zeta = (0, 0)$ this becomes a diagonal matrix with the maximum characteristic root $\frac{3}{2} > 1$. The conditions of Theorem 22.2 are satisfied. We are going to determine the angles L for which, taking V as the inside of the unit circle, $x^2 + y^2 < 1$, we have divergence for every starting point ξ_0 from $L \cap V$, and convergence for every starting point ξ_0 from V outside of L . From (22.25) we have, putting $z = re^{i\theta}$, and $z' = r'e^{i\theta'}$,

$$\begin{aligned}|z'|^2 &= z'\bar{z}' = r^2(1 + \frac{1}{4}(1+r^2)^2 - (1+r^2)\cos 2\theta), \\ \frac{|z'|^2}{|z|^2} &= 1 + \frac{1}{4}(1+r^2)^2 - (1+r^2)\cos 2\theta,\end{aligned}\quad (22.28)$$

and further

$$\tan \theta' = \frac{y'}{x'} = \frac{y(3+r^2)}{x(1-r^2)} = \frac{3+r^2}{1-r^2} \tan \theta. \quad (22.29)$$

9. We are going to prove now that we have *convergence* if z_0 is taken in V on the real axis and *divergence* if z_0 is taken in V outside of the real axis.

Indeed, if z is real and $|z| < 1$, then we see from (22.25) that z' has the sign of z and

$$\frac{z'}{z} = \frac{1-r^2}{2} < \frac{1}{2}.$$

We see that if we start with a real z_0 , $|z_0| < 1$, the sequence z_μ goes monotonically to zero.

10. Assume now that z_0 is in V , but not on the real axis, and that all z_μ remain in V . Then it follows from (22.29), denoting by θ_μ the argument of z_μ , that $t_\mu = \tan \theta_\mu$ tends either to $+\infty$ or to $-\infty$. But then we have

$$\cos(2 \arg z_\mu) = \frac{1-t_\mu^2}{1+t_\mu^2} \rightarrow -1.$$

Now it follows from (22.28), since $\cos 2\theta \rightarrow -1$, that

$$\underline{\lim} \frac{|z_{\mu+1}|^2}{|z_\mu|^2} = \underline{\lim} (1 + \frac{1}{4}(1 + |z_\mu|^2)^2) \geq \frac{3}{2},$$

and we see that $|z_\mu| \rightarrow \infty$, so that all z_μ cannot stay in V . This proves our assertion. We see that L is determined by

$$0 < \arg z < \pi, \quad \pi < \arg z < 2\pi.$$

23

Euclidean Norms

EUCLIDEAN LENGTH AND FROBENIUS NORM

1. For $p = 2$ we obtain from (19.2) the “Euclidean length” $|\xi|_2$ of the vector ξ ,

$$|\xi|_2 := \sqrt{|x_1|^2 + \cdots + |x_n|^2} = \sqrt{(\xi, \xi)}. \quad (23.1)$$

From (19.10) we have, applying (19.7),

$$|y_i|^2 \leq \sum_{j=1}^n |a_{ij}|^2 \sum_{j=1}^n |x_j|^2 = \sum_{j=1}^n |a_{ij}|^2 |\xi|_2^2$$

and, summing over i from 1 to n ,

$$|\eta|_2^2 \leq \sum_{i,j=1}^n |a_{ij}|^2 |\xi|_2^2.$$

If we therefore put

$$|A|_F := \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}, \quad (23.2)$$

we obtain

$$|\eta|_2 \leq |A|_F |\xi|_2. \quad (23.3)$$

2. $|A|_F$ is usually called the “Frobenius norm” of A . $|A|_F$ is easy to compute and in some cases gives a convenient measure for the size of A . However, often $|A|_F$ is too large and the equality sign in (23.7) cannot even be assured for the identity transformation, since $|1|_F = \sqrt{n} \neq 1$ ($n > 1$).

Still $|A|_F$ enjoys the multiplication property

$$|AB|_F \leq |A|_F |B|_F. \quad (23.4)$$

Indeed, putting $AB = C = (c_{ij})$, we have

$$\begin{aligned} c_{ij} &= \sum_{\mu=1}^n a_{i\mu} b_{\mu j}, \quad |c_{ij}|^2 \leq \sum_{\mu=1}^n |a_{i\mu}|^2 \sum_{v=1}^n |b_{vj}|^2, \\ &\sum_{i,j=1}^n |c_{ij}|^2 \leq \sum_{\mu,i=1}^n |a_{i\mu}|^2 \sum_{v,j=1}^n |b_{vj}|^2, \end{aligned}$$

which gives (23.4) immediately.

HERMITIAN MATRICES

3. If $A = (a_{ij})$ is an $n \times n$ matrix with real or complex entries, the transpose matrix of $\bar{A} = (\bar{a}_{ij})$ is often denoted as $A^* \equiv \bar{A}'$. If A has the property that $A^* = A$, that is, that $\bar{a}_{ij} = a_{ji}$ ($i = 1, \dots, n$; $j = 1, \dots, n$), then the matrix A is called a “Hermitian matrix.” If $\xi = (x_1, \dots, x_n)$ is a row vector with real or complex components x_μ , the expression

$$H_A(\xi) = \xi A \xi' = \sum_{i,j=1}^n a_{ij} x_i \bar{x}_j \quad (23.5)$$

is called a “Hermitian form” of $\{x_i\}$, $i = 1, \dots, n$, or of ξ . If a Hermitian matrix is real, then it is a real symmetric matrix. If both A and ξ are real, expression (23.5) is a *real quadratic form*.

The value of a Hermitian form is always real. Indeed we have from the definition

$$\overline{H_A(\xi)} = \sum_{i,j=1}^n \bar{a}_{ij} \bar{x}_i x_j = \sum_{i,j=1}^n a_{ji} x_j \bar{x}_i$$

and this is obviously $H_A(\xi)$, since we can replace the summation indices i, j with j, i . On the other hand, in general $H_A(\xi)$ is different from $H_A(\bar{\xi})$. If we put, e.g., for $n = 2$: $x_1 = 1$, $x_2 = i$, $a_{12} = -a_{21} = i$, $a_{11} = a_{22} = 0$, we have $H_A(\xi) = -H_A(\bar{\xi}) = 2$.

In particular we have

$$H_I(\xi) = \sum_{i=1}^n |x_i|^2 = |\xi|_2^2. \quad (23.6)$$

Observe that (23.5) is not a *polynomial* in the components of ξ ; it is not even an analytic function of these components. On the other hand, Hermitian forms have many properties precisely analogous to those of real quadratic forms, and the introduction of Hermitian forms often allows us to deal adequately with the mapping of a vector into its conjugate vector or into its Euclidean length.

4. In particular, it is proved in linear algebra that the eigenvalues of a Hermitian matrix are all real. If the Hermitian form (23.5) is always positive, unless $\xi = 0$, it is called *positive definite* and all its eigenvalues are *positive*. If (23.5) cannot assume negative values, but can vanish for some vector $\xi \neq 0$, it is called *positive semidefinite* and its eigenvalues are ≥ 0 , but at least one of them is $= 0$.

In both cases we denote the eigenvalues of A , ordered decreasingly, by

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0, \quad (23.7)$$

and put

$$\lambda_n =: \Lambda(A), \quad \lambda_1 =: \lambda(A) = \lambda_A. \quad (23.8)$$

Then it is well known that, for any $\xi \neq 0$,

$$\lambda(A) \leq H_A(\xi)/|\xi|_2^2 \leq \Lambda(A), \quad (23.9)$$

where on the right as well as on the left, for convenient choices of ξ , we can have the equality sign.

EUCLIDEAN NORM OF A MATRIX

5. If A is an arbitrary $n \times n$ matrix and $\eta = A\xi$, $\xi = (x_1, \dots, x_n)$, $\eta = (y_1, \dots, y_n)$, then

$$|\eta|_2^2 = \sum_{v=1}^n \bar{y}_v y_v = \sum_{v=1}^n \left(\sum_{i=1}^n \bar{a}_{vi} \bar{x}_i \right) \left(\sum_{j=1}^n a_{vj} x_j \right) = \sum_{i,j=1}^n \left(\sum_{v=1}^n \bar{a}_{vi} a_{vj} \right) \bar{x}_i x_j.$$

If we put now

$$s_{ij} = \sum_{v=1}^n a_{vi} \bar{a}_{vj}, \quad S = (s_{ij}), \quad (23.10)$$

it follows at once that

$$S = A^*A \quad (23.11)$$

and further that $S^* = A^*A^{**} = A^*A = S$, so that S is a Hermitian matrix and we can write $|\eta|_2^2 = H_S(\xi)$.

It then follows from (23.9), as $|\xi|_2 = |\xi|_2$, that

$$\lambda(S) \leq |\eta|_2/|\xi|_2^2 \leq \Lambda(S),$$

where for any A the equality signs on the right and on the left can be attained if we choose suitably the corresponding vectors ξ .

As A^*A and AA^* , by Section 8 of Chapter 19, have the same characteristic roots, we have $\lambda(A^*A) = \lambda(AA^*)$ and $\Lambda(A^*A) = \Lambda(AA^*)$.

6. Observe now that if A is Hermitian and positive definite or semi-definite, we have $S = A^2$ and by what has been said in Section 11 of Chapter 19,

$$\Lambda(S) = \Lambda(A)^2, \quad \lambda(S) = \lambda(A)^2.$$

We can therefore define for a general $n \times n$ matrix A :

$$\lambda(A) := \sqrt{\lambda(S)}, \quad \Lambda(A) := \sqrt{\Lambda(S)} \quad (23.12)$$

and it follows now that

$$\lambda(A) \leq |\eta|_2/|\xi|_2 \leq \Lambda(A), \quad (23.13)$$

where the bounds $\lambda(A)$ and $\Lambda(A)$ cannot be improved for any A . In this context $\Lambda(A) =: |A|_2$ is usually called *the Euclidean norm of A, induced by the Euclidean vector norm*. $\Lambda(A)$ and $\lambda(A)$ are also called the *upper resp. lower Euclidean bounds of A*.

If $|A| \neq 0$, we have $(A^{-1})^* = (A^*)^{-1}$ and therefore

$$(A^{-1})^* A^{-1} = A^{*-1} A^{-1} = (AA^*)^{-1}.$$

It follows that

$$\Lambda(A^{-1}) = 1/\lambda(A). \quad (23.14)$$

7. The exact computation of $|A|_2$ requires the formation and the solution of an equation of n th degree and is therefore in most cases impracticable. Therefore, the following estimates of $|A|_2$ and $\lambda(A)$ are often useful:

$$|A|_2 \leq F_A, \quad (23.15)$$

$$\lambda(A) > (n-1)^{(n-1)/2} |\det A| / |A|_F^{(n-1)/2}. \quad (23.16)$$

Indeed if we denote the eigenvalue of A^*A ordered decreasingly by α_v ,

$$\Lambda(A^*A) = \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n = \lambda(A^*A),$$

we have by (19.35) and (23.10)

$$\alpha_1 + \cdots + \alpha_n = \text{tr}(A^*A) = |A|_F^2, \quad |\det(A^*A)| = |\det A|^2 = \alpha_1 \cdots \alpha_n.$$

Formula (23.15) then follows immediately, since all α_v are ≥ 0 . On the other hand, we have, by the inequality between the geometric and arithmetic means, if $|A| \neq 0$, $\alpha_n > 0$:

$$\begin{aligned} \frac{|\det A|^2}{\alpha_n} &= \alpha_1 \alpha_2 \cdots \alpha_{n-1} \leq \left(\frac{\alpha_1 + \cdots + \alpha_{n-1}}{n-1} \right)^{n-1} = \left(\frac{|A|_F^2 - \alpha_n}{n-1} \right)^{n-1}, \\ \frac{|\det A|^2}{\alpha_n} &< \left(\frac{|A|_F^2}{n-1} \right)^{n-1}, \end{aligned}$$

and (23.16) follows immediately.

8. It may finally be observed that the estimates (23.15) and (23.16) cannot be improved by adding to the right-hand expression a convenient factor γ , depending only on n , < 1 for (23.15) and > 1 for (23.16). This is easily seen by choosing for A a diagonal matrix having along the diagonal $n-1$ times a number β and as the n th element λ , where β and λ are positive and in the first case $\beta/\lambda \rightarrow 0$, in the second case $\lambda/\beta \rightarrow 0$.

24

Minkowski Norms, $\Delta_p(A)$, $\Delta_{p,p'}(A)$

MINKOWSKI NORMS

1. We consider now the p norm of the vector ξ , defined by (19.2), and prove first that it is monotonically decreasing with the increasing p ,

$$|\xi|_\alpha \leq |\xi|_\beta \quad (\alpha > \beta \geq 1). \quad (24.1)$$

Both sides of (24.1) are homogeneous of dimension 1 with respect to a positive scalar multiplier of ξ . Multiplying, if necessary, with a convenient positive factor, we can assume without loss of generality that

$$|\xi|_\beta = 1, \quad \sum_{v=1}^n |x_v|^\beta = 1.$$

But then it follows, as each $|x_v| \leq 1$, that

$$|\xi|_\alpha = \left(\sum_{v=1}^n |x_v|^\alpha \right)^{1/\alpha} \leq \left(\sum_{v=1}^n |x_v|^\beta \right)^{1/\alpha} = 1 = |\xi|_\beta.$$

$|A|_p$ AND $|A|_{p,p'}$

2. The vector norm $|\xi|_p$ “induces” a corresponding matrix norm $|A|_p$ defined by

$$|A|_p := \text{Sup}(|A\xi'|_p / |\xi|_p) = \text{Sup}_{|\xi|_p=1} |A\xi'|_p. \quad (24.2)$$

Here the last supremum is at the same time a maximum, since the n -dimensional domain $|\xi|_p = 1$ is closed.

However, $|A|_p$ is known explicitly only in the cases $p = 1, 2, \infty$, so that in the general case we have to use convenient bounds of $|A|_p$. On the other hand, the definition above can be generalized, and this generalization is important even if we should prefer using only $|A|_p$. We can consider to this purpose *two indices* p and p' , both ≥ 1 , and use in the ξ -space the norm

$|\xi|_p$, and in the η -space given by $\eta' = A\xi'$, the norm $|\eta|_{p'}$. Then we define

$$|A|_{p,p'} = \sup_{\xi} (|A\xi'|_{p'}/|\xi|_p) = \max_{|\xi|_p=1} |A\xi'|_{p'}. \quad (24.3)$$

The norms $|A|_p, |A|_{p,p'}$ are called *Minkowski norms*. We have obviously $|A|_p = |A|_{p,p'}$. Further, it follows from the definitions that

$$|I|_p = 1. \quad (24.4)$$

3. The “triangle inequality” for the norms $|A|_{p,p'}$,

$$|A+B|_{p,p'} \leq |A|_{p,p'} + |B|_{p,p'} \quad (24.5)$$

follows easily. If we put

$$\eta' = A\xi', \quad \eta_1' = B\xi', \quad \eta' + \eta_1' = (A+B)\xi',$$

we obtain, assuming $|\xi|_p = 1$,

$$|\eta + \eta'|_{p'} \leq |\eta|_{p'} + |\eta'|_{p'} \leq |A|_{p,p'} + |B|_{p,p'},$$

and (24.5) follows immediately.

Further, if $p \wedge p' \wedge p''$, are ≥ 1 , we obtain, putting

$$\eta' = A\xi', \quad \zeta' = B\eta', \quad \zeta' = BA\xi'$$

and assuming $|\xi|_p = 1$,

$$|\zeta|_{p''} \leq |B|_{p',p''} |\eta|_{p'} \leq |B|_{p',p''} |A|_{p,p'},$$

and it follows that

$$|BA|_{p,p''} \leq |A|_{p,p'} |B|_{p',p''}, \quad (24.6)$$

and in particular for $p = p'' = p'$:

$$|AB|_p \leq |A|_p |B|_p. \quad (24.7)$$

As a corollary, we obtain, assuming $|A| \neq 0$ and taking $B = A^{-1}, p'' = p$:

$$|A^{-1}|_{p,p'} |A|_{p',p} \geq 1. \quad (24.8)$$

Finally, it follows from the definition (24.3) of $|A|_{p,p'}$, using the monotonicity property of $|\xi|_p$, that $|A|_{p,p'}$ is monotonically increasing with increasing p and monotonically decreasing with increasing p' .

It follows in particular that

$$\begin{aligned} |A|_{p,p'} &\geq |A|_2 & (p \geq 2 \geq p'), \\ |A|_{p,p'} &\leq |A|_2 & (p \leq 2 \leq p'). \end{aligned} \quad (24.9)$$

$\Delta_{p,p'}(A)$ AND $\Delta_p(A)$

4. In order to obtain an estimate of $|A|_{p,p'}$ by algebraic expressions, consider the number q correlated to p by (19.4) and apply, putting $\eta' = (y_1, \dots, y_n)' = A\xi'$, the Hölder inequality to $y_\mu = \sum_{v=1}^n a_{\mu v} x_v$ ($\mu = 1, \dots, n$). We obtain

$$\begin{aligned} |y_\mu| &\leq \left(\sum_{v=1}^n |a_{\mu v}|^q \right)^{1/q} |\xi|_p, \\ |y_\mu|^{p'} &\leq \left(\sum_{v=1}^n |a_{\mu v}|^q \right)^{p'/q} |\xi|_p^{p'}, \\ |\eta|_{p'}/|\xi|_p &\leq \left[\sum_{\mu=1}^n \left(\sum_{v=1}^n |a_{\mu v}|^q \right)^{p'/q} \right]^{1/p'}, \end{aligned}$$

and therefore

$$|A|_{p,p'} \leq \Delta_{p,p'}(A) := \left(\sum_{\mu=1}^n \left[\left(\sum_{v=1}^n |a_{\mu v}|^q \right)^{1/q} \right]^{p'} \right)^{1/p'}. \quad (24.10)$$

It follows in particular that

$$|A|_\infty = \max_{\mu} \sum_{v=1}^n |a_{\mu v}| =: \Delta_{\infty,\infty}(A), \quad |A|_1 = |A'|_\infty = \Delta_{\infty,\infty}(A'). \quad (24.10a)$$

Since the expression of $\Delta_{p,p'}(A)$ is still rather complicated, we consider now

$$\Delta_p(A) := \left(\sum_{\mu,v=1}^n |a_{\mu v}|^p \right)^{1/p} = \Delta_p(A'). \quad (24.11)$$

(This is obviously the p norm of A , if A is considered as a vector in the n^2 -dimensional space.)

From the definition (24.10) it follows, again applying the monotonicity of the p norm, that if $p' \geq q$, the right-hand expression for $\Delta_{p,p'}$ is increased if we replace p' with q , and decreased if we replace q with p' , while the inequalities go just the opposite way, if $p' \leq q$. It follows therefore that

$$\Delta_{\max(p',q)}(A) \leq \Delta_{p,p'}(A) \leq \Delta_{\min(p',q)}(A). \quad (24.12)$$

In particular, we now obtain, taking $p' = q$,

$$\Delta_q(A) = \Delta_{p,q}(A). \quad (24.13)$$

5. We will now prove a multiplication formula for $\Delta_{p,p'}$. In this formula q is correlated to p by (19.4), while r,s are an arbitrary couple satisfying (19.4), if p and q there are replaced with r,s :

$$\Delta_{p,p'}(AB) \leq \Delta_{s,p'}(A) \Delta_{r,q}(B'). \quad (24.14)$$

Indeed, we have, with $AB =: C = (c_{\mu\nu})$:

$$\begin{aligned} |c_{\mu\nu}| &= \left| \sum_{\kappa} a_{\mu\kappa} b_{\kappa\nu} \right|, \\ \sum_{v=1}^n |c_{\mu v}|^q &\leq \left(\sum_{\kappa=1}^n |a_{\mu\kappa}|^r \right)^{q/r} \sum_{v=1}^n \left(\sum_{\kappa=1}^n |b_{\kappa v}|^s \right)^{q/s} = \left(\sum_{\kappa=1}^n |a_{\mu\kappa}|^r \right)^{q/r} \Delta_{r,q}^q(B'), \\ \left(\sum_{v=1}^n |c_{\mu v}|^q \right)^{p'/q} &\leq \Delta_{r,q}^{p'}(B') \left(\sum_{\kappa=1}^n |a_{\mu\kappa}|^r \right)^{p'/r}. \end{aligned}$$

Summing this over μ and raising both sides into the power with the exponent $1/p'$, formula (24.14) follows immediately.

6. On the other hand, it follows from the monotonicity properties of $|A|_{p,p'}$ that

$$|A|_{1,\infty} \leq |A|_{p,p'} \leq |A|_{\infty,1}. \quad (24.15)$$

This inequality is of interest since $|A|_{1,\infty}$ (and, in the case of nonnegative $a_{\mu\nu}$, $|A|_{\infty,1}$) is easily computed.

As to $|A|_{1,\infty}$ we have, by virtue of (24.13),

$$|A|_{1,\infty} \leq \Delta_{1,\infty}(A) = \Delta_\infty(A) = \text{Max}_{\mu,\nu} |a_{\mu\nu}|.$$

Here we have, however, the equality sign. Indeed, if $\text{Max}_{\mu,\nu} |a_{\mu\nu}| = |a_{ik}|$, we have, putting $x_\kappa = 0$ if $\kappa \neq k$ and $= 1$ if $\kappa = k$, $y_\mu = a_{\mu k}$,

$$|\eta|_\infty = \text{Max}_\mu |a_{\mu k}| = |a_{ik}|,$$

and therefore

$$|A|_{1,\infty} = \Delta_{1,\infty}(A) = \Delta_\infty(A) = \text{Max}_{\mu,\nu} |a_{\mu\nu}|. \quad (24.16)$$

On the other hand, we have

$$|A|_{\infty,1} \leq \Delta_{\infty,1}(A) = \Delta_1(A) = \sum_{\mu,\nu} |a_{\mu\nu}| \quad (24.17)$$

and here the equality sign holds if all $a_{\mu\nu}$ are nonnegative, taking all $x_v = 1$.[†]

[†] Observe that the inequality in (24.17) can indeed be a strict inequality, as is easily seen, e.g., for the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -\sqrt{2} \end{pmatrix}$$

for which $\Delta_1(A) = 3 + \sqrt{2} = 4.414 \dots$ and $|A|_{\infty,1} = 3.154 \dots$

INEQUALITIES FOR $\Delta_{p,p'}(A)$

7. The expression $\Delta_p(A)$ becomes for $p = 2$ the Frobenius norm F_A and can be used for all p with $1 \leq p \leq 2$ as a convenient “norm” of the matrix A . Indeed, we have, from the inequality in (24.10), replacing p, p' there with q, p , $|\eta|_p \leq \Delta_{q,p}(A)|\zeta|_q$, and therefore by (24.13),

$$|A|_{q,p} \leq \Delta_p(A), \quad (24.18)$$

$$|\eta|_p \leq \Delta_p(A)|\zeta|_q \quad (\eta' = A\zeta'). \quad (24.19)$$

If we now have $p \leq 2 \leq q$, it follows that $|A|_p = |A|_{p,p} \leq |A|_{q,p}$ and we see that

$$|A|_p \leq \Delta_p(A) \quad (p \leq 2). \quad (24.20)$$

8. If, in (24.14), we replace p with q , p' with p , s with q , and r with p , we obtain

$$\Delta_{q,p}(AB) \leq \Delta_{q,p}(A)\Delta_{p,p}(B'),$$

and, using (24.13),

$$\Delta_p(AB) \leq \Delta_p(A)\Delta_{p,p}(B'). \quad (24.21)$$

On the other hand, replacing p' in (24.14) with p , s with q , and r with p and writing C instead of A , we obtain

$$\Delta_{p,p}(CB) \leq \Delta_{q,p}(C)\Delta_{p,q}(B') \leq \Delta_p(C)\Delta_q(B'). \quad (24.22)$$

If we now assume that $p \leq 2 \leq q$, it follows that

$$\Delta_{p,p}(B') \leq \Delta_{q,p}(B') = \Delta_p(B') = \Delta_p(B)$$

and from (24.21) we obtain the multiplication formula

$$\Delta_p(AB) \leq \Delta_p(A)\Delta_p(B) \quad (1 \leq p \leq 2).^{\dagger} \quad (24.23)$$

9. There exists, however, a simple multiplication formula for the product of three matrices:

$$\Delta_p(ABC) \leq \Delta_p(A)\Delta_q(B)\Delta_p(C) \quad (1 \leq p \leq \infty). \quad (24.24)$$

[†] Observe that this inequality need not be true for $p > 2$. Indeed, if

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad A^2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},$$

we have

$$\Delta_p(A) = \sqrt[2]{4}, \quad \Delta_p(A^2) = 2\sqrt[2]{4}$$

and therefore

$$\Delta_p(A^2) > (\Delta_p(A))^2 \quad (p > 2).$$

Indeed, replacing B in (24.21) with BC , it follows that

$$\Delta_p(ABC) \leq \Delta_p(A)\Delta_{p,p'}(C'B').$$

But here the second right-hand factor is, by virtue of (24.22), $\leq \Delta_q(B)\Delta_p(C)$ and (24.24) follows.

VARIATION OF THE INVERSE MATRIX

10. Relation (24.24) has an important corollary. Assume that A and B are $n \times n$ regular matrices. From the identity

$$A(B^{-1} - A^{-1})B = A - B$$

it follows that

$$B^{-1} - A^{-1} = A^{-1}(A - B)B^{-1}. \quad (24.25)$$

Applying (24.24) we obtain

$$\Delta_p(B^{-1} - A^{-1}) \leq \Delta_q(B - A)\Delta_p(A^{-1})\Delta_p(B^{-1}). \quad (24.25a)$$

Since Δ_p is the p norm in n^2 -dimensional space, it obeys the triangle inequality and we can write

$$|\Delta_p(B^{-1}) - \Delta_p(A^{-1})| \leq \Delta_p(B^{-1} - A^{-1}),$$

dividing by $\Delta_p(A^{-1})\Delta_p(B^{-1})$, we obtain

$$\left| \frac{1}{\Delta_p(A^{-1})} - \frac{1}{\Delta_p(B^{-1})} \right| \leq \Delta_q(A - B) \quad (\exists A^{-1} \wedge B^{-1}) \quad (24.26)$$

and

$$\frac{\Delta_p(A^{-1})}{1 + \Delta_p(A^{-1})\Delta_q(B - A)} \leq \Delta_p(B^{-1}) \leq \frac{\Delta_p(A^{-1})}{1 - \Delta_p(A^{-1})\Delta_q(B - A)}, \quad (24.27)$$

where for the validity of the right-hand inequality we must assume that $\Delta_p(A^{-1})\Delta_q(B - A) < 1$.

11. Formula (24.27) lets it appear intuitively clear that if we have $\Delta_p(A^{-1})\cdot\Delta_q(B - A) < 1$, then B^{-1} exists. Of course the argument is not conclusive since in deriving our formulas we assumed that B^{-1} existed. It is easy, however, to prove the corresponding assertion rigorously:

Theorem 24.1. *If we have two ($n \times n$) matrices A , B , with A assumed as regular, and*

$$\Delta_p(A^{-1})\Delta_q(B - A) < 1, \quad (24.28)$$

then B^{-1} exists and we have

$$\Delta_p(B^{-1} - A^{-1}) \leq \frac{\Delta_p(A^{-1})^2 \Delta_q(B-A)}{1 - \Delta_p(A^{-1}) \Delta_q(B-A)}. \quad (24.29)$$

Proof. Put

$$B_t = A + t(B-A) \quad (0 \leq t \leq 1).$$

For sufficiently small positive t , B_t is certainly regular. If $B = B_1$ is singular, there exists the smallest positive t_0 such that B_{t_0} is singular, while all B_t for $0 \leq t < t_0$ are regular. But then we can apply (24.27) and obtain

$$\Delta_p(B_t^{-1}) \leq \frac{\Delta_p(A^{-1})}{1 - \Delta_p(A^{-1}) \Delta_q(B-A)} \quad (0 \leq t < t_0);$$

we see that if t increases to t_0 , $\Delta_p(B_t^{-1})$ remains bounded. The same holds therefore for all elements of B_t^{-1} and therefore also for the determinant of B_t^{-1} , while this determinant cannot remain bounded in the neighborhood of t_0 . Formula (24.29) follows now from (24.27) and (24.25a).

For $p \leq 2$ an analogous result can be derived, using in (24.25) the norm $|A|_p$ instead of $\Delta_p(A)$. However, this result is a special case of a general theorem, Theorem 34.2, from the theory of Banach spaces, which will be derived in Chapter 34.

25

Method of Steepest Descent. Convergence of the Procedure

1. The method of solution of systems of equations discussed in Chapter 22 is characteristically a *local* method whose convergence is assured only if the starting approximation is already sufficiently near to the solution in question. A method of quite another character is that proposed by Cauchy in 1847, which assures “*global convergence*” under very general conditions. This is the so-called *method of steepest descent* or *gradient method*.

IDEA OF THE METHOD

2. Consider a bounded open portion Ω of the n -dimensional space. We suppose first, to explain Cauchy’s idea, that the complete boundary of Ω consists of a certain “regular” surface S .

Consider a function $f(\xi)$ defined and continuous on $\Omega \cup S$ which has on S a fixed value C and is in the interior of Ω everywhere $< C$. Assume that $f(\xi)$ has continuous first derivatives on $\Omega \cup S$.

3. Then Cauchy’s idea is that if we start from a point ξ_0 on S and go along the normal to S into the interior of Ω a segment of the length r_0 with the end point ξ_1 , the value of $f(\xi)$ will diminish from $C = C_0$ to $C_1 < C_0$. In the point ξ_1 there usually exists the nonvanishing gradient vector of f , and the opposite direction of this vector is the direction of steepest descent at ξ_1 . Going from ξ_1 along this direction a segment of the length r_1 , we come to a point ξ_2 where $f(\xi_2) = C_2 < C_1$. Continuing this process indefinitely and choosing the r_μ conveniently, we can then hope that the sequence of the ξ_μ converges to a point ξ^* in which $f(\xi)$ has a minimum and which satisfies the set of the equations

$$\frac{\partial f}{\partial x_v}(\xi^*) = 0 \quad (v = 1, \dots, n). \quad (25.1)$$

4. It appears first that the aim of the method described is the solution of a rather special system of equations (25.1).

However, if we have the general system of equations which can be written in the form

$$f_v(\xi) = 0 \quad (v = 1, \dots, n), \quad (25.2)$$

we can form $f(\xi)$ as

$$f(\xi) := f_1(\xi)^2 + \dots + f_n(\xi)^2. \quad (25.3)$$

This $f(\xi)$ has in the solution point of (25.2) the absolute minimum 0 and we can hope to obtain this point using Cauchy's method.

5. Analytically the vector $\text{grad } f(\xi_\mu)$ has the components

$$f'_{x_1}(\xi_\mu), \dots, f'_{x_n}(\xi_\mu).$$

We will denote this gradient in the point ξ generally by the symbol $f'(\xi)$. If we assume that $f'(\xi_\mu)$ does not already vanish, we can write, denoting generally by $|\xi|$ the Euclidean length of ξ , i.e., $|\xi|_2$,

$$|f'(\xi_\mu)| =: \kappa_\mu, \quad f'(\xi_\mu) =: \kappa_\mu \varphi_\mu. \quad (25.4)$$

Then Cauchy's rule can be represented by the formula

$$\xi_{\mu+1} := \xi_\mu - r_\mu \varphi_\mu. \quad (25.5)$$

6. As to the choice of r_μ , Cauchy's recommendation was to choose r_μ in such a way that the function $f(\xi_\mu - t\varphi_\mu)$ of the positive variable t has for $t = r_\mu$ its *minimum*. However, this is in praxis only possible without too extensive computations, if $f(\xi)$ is a quadratic polynomial. Therefore we can try instead to take r_μ as $c\kappa_\mu$, choosing c in such a way that we are certain not to overshoot the minimum point. On the other hand, computational convenience makes it desirable not to keep too rigidly to the formula $r_\mu = c\kappa_\mu$. We will therefore introduce a sequence of real numbers ε_μ satisfying for a positive $\varepsilon < 1$ the condition

$$|\varepsilon_\mu| \leq \varepsilon, \quad \varepsilon < 1 \quad (\mu = 0, 1, \dots)$$

and put

$$r_\mu := (1 + \varepsilon_\mu) c \kappa_\mu. \quad (25.6)$$

7. On the other hand, it is also not necessary to keep the direction rigidly fixed as that of φ_μ , i.e., of the gradient at ξ_μ . We can take instead any direction whose angle with the direction φ_μ remains essentially under $\pi/2$. We therefore introduce for each μ a vector ψ_μ of the length 1 such that we have for the cosine of the angle between φ_μ and ψ_μ , i.e., for the inner product (φ_μ, ψ_μ) ,

$$\delta_\mu := (\varphi_\mu, \psi_\mu) \geq \delta, \quad 0 < \delta < 1, \quad (25.7)$$

with a fixed δ .

8. The system of the $\binom{n}{2}$ second derivatives of f , which are now assumed to exist on Ω , can be ordered as an $(n \times n)$ symmetric matrix which we will denote by

$$f''(\xi) := \left(\frac{\partial^2 f(\xi)}{\partial x_i \partial x_j} \right). \quad (25.8)$$

Together with (25.8), we will also have to consider the more general matrix which is obtained from $f''(\xi)$, if we take in each line of this matrix an argument depending on this line, i.e.,

$$f''(\xi_1, \dots, \xi_n) := \left(\frac{\partial^2 f(\xi_i)}{\partial x_i \partial x_k} \right) \quad (i, k = 1, \dots, n). \quad (25.9)$$

We now make the assumption, which is fundamental for our discussion, that there exists a fixed *positive* Λ^* such that (cf. (23.8))

$$\Lambda(f''(\xi)) \leq \Lambda^* \quad (\xi \in \Omega). \quad (25.10)$$

9. In terms of δ and Λ^* the convenient value of c in (25.6) can now be taken as δ/Λ^* , so that we have

$$r_\mu = (1 + \varepsilon_\mu)(\delta/\Lambda^*)\kappa_\mu, \quad (25.11)$$

$$\xi_{\mu+1} = \xi_\mu - r_\mu \psi_\mu. \quad (25.12)$$

10. Denote by Ω^* the set of all points ξ from Ω in which $f'(\xi) = 0$:

$$\frac{\partial f}{\partial x_i}(\xi) = 0 \quad (i = 1, \dots, n). \quad (25.13)$$

We denote by the symbol $|\xi, \Omega^*|$ the distance of ξ from Ω^* , i.e.,

$$|\xi, \Omega^*| = \inf_{\xi' \in \Omega^*} |\xi - \xi'|, \quad (25.14)$$

and we say that *a sequence of points ξ_μ tends to Ω^** , $\xi_\mu \rightarrow \Omega^*$, if

$$|\xi_\mu, \Omega^*| \rightarrow 0 \quad (\mu \rightarrow \infty). \quad (25.15)$$

Later we will discuss in what cases it follows from (25.15) that ξ_μ is convergent in the usual sense.

CONVERGENCE OF THE PROCEDURE

11. We can now formulate our first theorem, generalizing the geometric configuration beyond the assumptions of Section 2.

Theorem 25.1. *Take a bounded, open, n -dimensional set Ω with the boundary S and assume that a function $f(\xi)$ is continuous on $\Omega \cup S$, has a con-*

stant value C on S , and is $< C$ everywhere on Ω . Take a boundary point ξ_0 of Ω and assume that $f(\xi)$ has continuous second derivatives everywhere on the set $\Omega \cup \xi_0$, that (25.10) is satisfied, and that $f'(\xi_0) \neq 0$. Then we can, starting from ξ_0 , use the rule (25.12) indefinitely for $\mu = 0, 1, \dots$; all ξ_μ lie in Ω , tend to Ω^* in the sense of (25.15), and we have

$$\xi_{\mu+1} - \xi_\mu \rightarrow 0 \quad (\mu \rightarrow \infty). \quad (25.16)$$

12. Proof. Putting

$$\xi^{(t)} = \xi_0 - tr_0 \psi_0, \quad 0 \leq t \leq 1,$$

develop $f(\xi^{(t)})$ in powers of t with the remainder term of second order. We obtain, using (25.4), for a θ with $0 \leq \theta \leq 1$,

$$f(\xi^{(t)}) - f(\xi_0) = -tr_0 \kappa_0 (\psi_0, \varphi_0) + \frac{t^2}{2} r_0^2 (\psi_0, f''(\xi_0 - \theta tr_0 \psi_0) \psi_0'). \quad (25.17)$$

Since the quotient of the right-hand expression by t has a negative limit as $t \downarrow 0$, we see that the right-hand expression for sufficiently small positive t is negative, so that $f(\xi^{(t)})$ remains $< C$. But then for sufficiently small positive t , $\xi^{(t)}$ remains in Ω and the same holds for the argument of f'' in the remainder term. We can, therefore, for sufficiently small positive t , write, using (25.7), (25.11), (25.10), and (23.9),

$$\begin{aligned} f(\xi^{(t)}) - f(\xi_0) &\leq -(1 + \varepsilon_0) t \frac{\delta}{\Lambda^*} \kappa_0^2 \delta + \frac{t^2}{2} (1 + \varepsilon_0)^2 \frac{\delta^2}{\Lambda^{*2}} \kappa_0^2 \Lambda^* \\ &= -(1 + \varepsilon_0) \frac{\delta^2}{\Lambda^*} \kappa_0^2 \frac{t}{2} (2 - t(1 + \varepsilon_0)) \\ &\leq -\frac{1 - \varepsilon_0^2}{2\Lambda^*} \delta^2 \kappa_0^2 t. \end{aligned}$$

Since $\xi^{(t)}$ and the whole segment from ξ_0 to $\xi^{(t)}$, save ξ_0 , lie in Ω , we can go up to $t = 1$ and remain always in Ω since $f(\xi^{(t)})$ stays $< C$ and therefore we cannot meet a point of S . We see that ξ_1 lies in Ω and that we have

$$\Delta_0 := f(\xi_0) - f(\xi_1) \geq \frac{1 - \varepsilon_0^2}{2\Lambda^*} \delta^2 \kappa_0^2 \geq \frac{1 - \varepsilon^2}{2\Lambda^*} \delta^2 \kappa_0^2.$$

13. Starting from ξ_1 , we can proceed in the same way further and obtain a sequence ξ_μ of points in Ω such that

$$\Delta_\mu := f(\xi_\mu) - f(\xi_{\mu+1}) \geq \frac{1 - \varepsilon^2}{2\Lambda^*} \delta^2 \kappa_\mu^2 \quad (\mu = 0, 1, \dots). \quad (25.18)$$

The sequence $f(\xi_\mu)$ is thus monotonically decreasing and convergent, and we have $\Delta_\mu \rightarrow 0$ and from (25.18)

$$\kappa_\mu \rightarrow 0, \quad f'(\xi_\mu) \rightarrow 0, \quad r_\mu \rightarrow 0, \quad \xi_{\mu+1} - \xi_\mu \rightarrow 0.$$

Suppose now that the sequence ξ_μ does not tend to Ω^* , i.e., the sequence of positive numbers $|\xi_\mu, \Omega^*|$ does not tend to 0. Then we can find a partial sequence μ_k for which

$$|\xi_{\mu_k}, \Omega^*| \rightarrow p > 0.$$

On the other hand, the sequence ξ_{μ_k} , which lies in a bounded portion of R_n , contains a subsequence converging in the usual sense to a point ζ contained in S_1 , i.e., in the interior of S . But then we have

$$|\zeta, \Omega^*| = p > 0,$$

while on the other hand, since $f'(\xi_\mu) \rightarrow 0$, we must have $f'(\zeta) = 0$, so that ζ belongs to Ω^* . With this contradiction our theorem is proved.

APPLICATION TO $|f(x+iy)|^2$

14. We show the working of our method on a general example derived from the theory of analytic functions of one variable. Consider a function of the complex variable $z = x + iy$:

$$f(z) = u(x, y) + iv(x, y), \quad F(z) = |f(z)|^2 = u^2(x, y) + v^2(x, y), \quad (25.19)$$

where by the Cauchy–Riemann equations

$$u_x' = v_y', \quad u_y' = -v_x'. \quad (25.20)$$

Denoting by $f' = f'(z)$ the derivative with respect to z , we have

$$\frac{\partial f(z)}{\partial x} = f'(z), \quad \frac{\partial f(z)}{\partial y} = if'(z), \quad (25.21)$$

$$\overline{\frac{\partial f(z)}{\partial x}} = u_x' - iv_x' = \overline{f'(z)}, \quad \overline{\frac{\partial f(z)}{\partial y}} = u_y' - iv_y' = -v_x' - iu_x' = -i\overline{f'(z)}.$$

It now follows that

$$F_x' = (ff)_x' = f_x'f + ff_x' = f'f + ff' = 2R(f\bar{f}'),$$

$$F_y' = f_y'f + ff_y' = if'f - iff' = \frac{1}{i}(ff' - f'f) = 2I(f\bar{f}'),$$

$$Z(z) = Z(x, y) := F_x' + iF_y' = 2f(z)\overline{f'(z)}. \quad (25.22)$$

15. From (25.20) we have further

$$\frac{\partial(u, v)}{\partial(x, y)} = u_x'^2 + v_x'^2 = |f'(z)|^2. \quad (25.23)$$

On the other hand, from (25.19) it follows, using $u_{xx}'' + u_{yy}'' = v_{xx}'' + v_{yy}'' = 0$ and (25.20), that

$$\begin{aligned} \frac{1}{2}F_{xx}'' &= u_x'^2 + v_x'^2 + uu_{xx}'' + vv_{xx}'' = |f'(z)|^2 + R(\bar{f}\bar{f}''), \\ \frac{1}{2}F_{xy}'' &= u_y' u_x' + v_y' v_x' + uu_{xy}'' + vv_{xy}'' \\ &= (-v_x') u_x' + u_x' v_x' + u(-v_x')_x + v(u_x')_x = I(\bar{f}\bar{f}''), \\ \frac{1}{2}F_{yy}'' &= (-uv_x' + vu_x')_y = -u_y' v_x' + v_y' u_x' - uv_{yx}'' + vu_{yx}'' \\ &= v_x'^2 + u_x'^2 - (uu_{xx}'' + vv_{xx}'') = |f'(z)|^2 - R(\bar{f}\bar{f}''); \end{aligned}$$

collecting these results we have

$$\begin{aligned} F_{xx}'' &= 2|f'|^2 + 2R(\bar{f}\bar{f}''), \\ F_{xy}'' &= 2I(\bar{f}\bar{f}''), \\ F_{yy}'' &= 2|f'|^2 - 2R(\bar{f}\bar{f}''). \end{aligned} \quad (25.24)$$

16. The Hessian matrix of F ,

$$H := \begin{pmatrix} F_{xx}'' & F_{xy}'' \\ F_{xy}'' & F_{yy}'' \end{pmatrix},$$

therefore has as its determinant

$$\begin{vmatrix} F_{xx}'' & F_{xy}'' \\ F_{xy}'' & F_{yy}'' \end{vmatrix} = 4|f'|^4 - 4((R(\bar{f}\bar{f}''))^2 + (I(\bar{f}\bar{f}''))^2) = 4|f'|^4 - 4|f|^2|f''|^2. \quad (25.25)$$

As the trace of this matrix is $F_{xx}'' + F_{yy}'' = 4|f'|^2$, we obtain as the fundamental equation of H

$$\lambda^2 - 4|f'|^2\lambda + 4|f'|^4 - 4|f|^2|f''|^2 = 0, \quad (25.26)$$

and since the roots of this equation are $2|f'|^2 \pm 2|f''|$, we have

$$\lambda(H) = 2|f'|^2 - 2|f''|, \quad \Lambda(H) = 2|f'|^2 + 2|f''|. \quad (25.27)$$

17. In order to apply the above theory to $F(z)$ given by (25.19), assume that $f'(z)$ is regular on a closed curve C_0 and in a domain Ω obtained from the interior of C_0 removing from this interior the interior of a finite number of closed curves C_1, C_2, \dots, C_m lying inside C_0 and having no points in common (m could be $= 0$). Assume now that $|f(z)|$ has a constant value $\gamma > 0$ on the

whole boundary S of Ω ; then the values of $|f(z)|$ in Ω are certainly less than γ , since otherwise this modulus would assume a maximum in Ω . Therefore $|f(z)|$ assumes in Ω its minimum, which cannot be >0 . We see that $f(z)$ has certainly a positive (and finite) number of zeros in Ω .

Since we have for the length of the gradient $F'(z)$ of $F(z)$, $|F'(z)| = 2|f'(z)||f(z)|$, the set Ω^* consists of all zeros of $f(z)f'(z)$ lying inside of Ω .

As to Λ^* , it is given by the formula

$$\Lambda^* = 2 \operatorname{Max}_{z \in \Omega+S} (|f'(z)|^2 + |f(z)f''(z)|).^\dagger \quad (25.28)$$

In this particular case it is simpler to use $\psi = \varphi$ so that the iteration formula is of the type

$$z_1 = z_0 - 2tf(z_0)\vec{f}'(z_0). \quad (25.29)$$

It will follow from the general results of the next chapter that in our case the sequence ξ_μ is always convergent in the usual sense either to a zero of $f(z)$ or to a zero of $f'(z)$.

Observe finally that if $f(z)$ is a polynomial, we need not care about the geometric configuration C, C_1, \dots , but can just start with any point z_0 and go on using the iteration formula (25.29).

\dagger As a matter of fact we have even

$$\Lambda^* = 2 \operatorname{Max}_{z \in S} (|f'(z)|^2 + \gamma |f''(z)|). \quad (25.28')$$

This follows from the fact that $2|f'|^2 + 2|ff''|$ is a so-called subharmonic function, and such functions have the “maximum property,” that is, if such a function is subharmonic on $\Omega + S$, it attains its maximum on S .

26

Method of Steepest Descent. Weakly Linear Convergence of the ξ_μ

THE DERIVED SET OF THE ξ_μ

1. If the set Ω^* in Theorem 25.1 consists of only one point, ζ , then from Theorem 25.1 it follows that $\xi_\mu \rightarrow \zeta$.

It is easily seen that the ξ_μ must be convergent in the usual sense, if Ω^* is a finite set. More general conditions can be obtained using

Theorem 26.1. *Assume a bounded sequence S of points ξ_v in R_n for which $\xi_{v+1} - \xi_v \rightarrow 0$. Then the derived set S' of S is a continuum, if ξ_v does not converge in the usual sense.*

2. Remark. We remind the reader that a *continuum* is defined as a closed set of points which cannot be decomposed into the sum of two disjoint closed sets.

3. Proof of Theorem 26.1. S' is obviously closed. Suppose we have $S' = C_1 + C_2$ where C_1 and C_2 are disjoint and both closed. Then there exists a positive p such that the distance of any point of C_1 from every point of C_2 is $> p$. By hypothesis we have for a certain n_0

$$|\xi_{v+1} - \xi_v| \leq p/3 \quad (v \geq n_0). \quad (26.1)$$

Take a point P from C_1 . There exist then arbitrarily large indices $m > n_0$ such that $|\xi_m, P| < p/3$. As the points ξ_v with $v > m$ have a cluster point in C_2 , there exist indices $k > m$ such that $|\xi_k, C_2| \leq 2p/3$. Assume that k is the smallest such index. Then we have certainly $|\xi_{k-1}, C_2| > 2p/3$, and therefore by (26.1) $|\xi_k, C_2| > p/3$. We see that

$$p/3 < |\xi_k, C_2| \leq 2p/3, \quad k > m. \quad (26.2)$$

4. There exists therefore an infinite sequence of indices k_1, k_2, \dots for which (26.2) holds. A cluster point ζ of this sequence belongs to S and satisfies the relation

$$p/3 \leq |\zeta, C_2| \leq 2p/3. \quad (26.3)$$

It therefore does not belong to C_2 . It must then lie in C_1 while its distance from C_2 is less than p . With this contradiction our theorem is proved.

WEAKLY LINEAR CONVERGENCE

5. We will now assume that under the conditions of Theorem 25.1 we have

$$\xi_\mu \rightarrow \zeta. \quad (26.4)$$

We are going to show that then under certain additional conditions we have the *weakly linear convergence* of the ξ_μ to ζ in the following sense: We will say that the convergence of the ξ_μ to ζ is *weakly linear* if there exists a positive integer N so that

$$\lim_{\mu \rightarrow \infty} \frac{|\xi_{\mu+N} - \zeta|}{|\xi_\mu - \zeta|} < 1. \dagger \quad (26.5)$$

6. Consider a function $f(\xi) = f(x_1, \dots, x_n)$ continuous with its first and second derivatives in a neighborhood of a point $\zeta(z_1, \dots, z_n)$. Assume that $f'(\zeta) = 0$. Then, as is well known, in order that $f(\xi)$ have in ζ a (local) *minimum*, it is *necessary* that the matrix $f''(\zeta)$ be positive (definite or semidefinite) and *sufficient* that $f''(\zeta)$ be positive definite. If in particular $f''(\zeta)$ is positive definite, we say that $f(\xi)$ has in ζ a *regular minimum*.

7. Theorem 26.2. *Assume under the conditions of Theorem 25.1 that ξ_μ tends to a limit ζ and that $f(\xi)$ has in ζ a regular minimum. Then the convergence of the ξ_μ to ζ is weakly linear.*

8. Proof. Without loss of generality, we can assume that $\zeta = 0$, $f(\zeta) = 0$ and we will denote $f''(\zeta) = f''(0)$ by f''_0 . Put

$$\lambda = \frac{1}{2}\lambda(f''_0), \quad \Lambda = 2\Lambda(f''_0). \quad (26.6)$$

Since the components of the matrix $f''(\xi)$ are continuous in the neighborhood of the origin, there exists such a neighborhood U of the origin that we have

$$\lambda(f''(\xi)) > \lambda, \quad \Lambda(f''(\xi)) < \Lambda \quad (\xi \in U). \quad (26.7)$$

We can therefore also assume that Λ^* introduced in (25.10) and used in (25.11) is equal to Λ .

† If we have a sequence α_v converging linearly to 0 and multiply the α_v by the constants c_v , such that for two positive numbers c, C we have $c \leq |c_v| \leq C$, the new sequence $c_v \alpha_v$ will not in general converge linearly to 0. On the other hand, if the α_v converge *weakly linearly* to 0, this property is preserved in the sequence $c_v \alpha_v$. It is this fact that makes the concept of the weakly linear convergence useful in numerical analysis.

9. We have, by what has been said at the end of Chapter 19,

$$\Lambda(f_0''^{-1}) = \frac{1}{\lambda(f_0'')} . \quad (26.8)$$

Further, obviously,

$$\Lambda(f''(\eta_1, \dots, \eta_n)^{-1}) \rightarrow \Lambda(f_0''^{-1}) = \frac{1}{\lambda(f_0'')} \quad (26.9)$$

if η_1, \dots, η_n tend independently to the origin. Taking the neighborhood U of the origin sufficiently small, we therefore also have

$$\Lambda(f''(\eta_1, \dots, \eta_n)^{-1}) < \frac{2}{\lambda(f_0'')} = \frac{1}{\lambda} \quad (\eta_1, \dots, \eta_n \in U). \quad (26.10)$$

We can further assume that all ξ_μ already lie in U .

10. Developing $f(\xi_\mu)$ at the origin with the remainder of the second order, we have, since $f'(0) = 0$,

$$f(\xi_\mu) = \frac{1}{2}(\xi_\mu, f''(\theta\xi_\mu)\xi_\mu'), \quad 0 \leq \theta \leq 1,$$

and therefore, using (26.7),

$$\lambda |\xi_\mu|_2^2 \leq 2f(\xi_\mu) \leq \Lambda |\xi_\mu|_2^2. \quad (26.11)$$

11. Apply to each component of $f'(\xi_\mu)$ the mean value theorem; we obtain

$$\frac{\partial f(\xi_\mu)}{\partial x_\kappa} = \sum_{v=1}^n \frac{\partial^2 f(\eta_\kappa)}{\partial x_\kappa \partial x_v} x_v^{(\mu)}, \quad \xi_\mu = (x_1^{(\mu)}, \dots, x_n^{(\mu)}),$$

where η_κ lies on the segment joining ξ_μ with the origin, that is, in U . We can therefore write, in notation (25.9),

$$f'(\xi_\mu) = f''(\eta_1, \dots, \eta_n) \xi_\mu$$

and, solving this with respect to ξ_μ and using (25.4),

$$\xi_\mu = f''(\eta_1, \dots, \eta_n)^{-1} f'(\xi_\mu) = \kappa_\mu f''(\eta_1, \dots, \eta_n)^{-1} \varphi_\mu.$$

Therefore by (23.13)

$$|\xi_\mu|_2 \leq \kappa_\mu \Lambda f''(\eta_1, \dots, \eta_n)^{-1}.$$

Using now (26.10) we get $|\xi_\mu|_2 \leq \kappa_\mu / \lambda$, and combining this with (26.11), we obtain

$$\kappa_\mu^2 \geq \lambda^2 |\xi_\mu|_2^2 \geq 2 \frac{\lambda^2}{\Lambda} f(\xi_\mu). \quad (26.12)$$

12. Using (26.11) we have, for any positive integer N ,

$$\frac{|\xi_{\mu+N}|_2^2}{|\xi_\mu|_2^2} \leq \frac{\Lambda}{\lambda} \frac{f(\xi_{\mu+N})}{f(\xi_\mu)}. \quad (26.13)$$

On the other hand, we have, from (25.18) and (26.12),

$$\Delta_\mu \geq \frac{1-\varepsilon^2}{2} \frac{\delta^2}{\Lambda} \left(\frac{2\lambda^2}{\Lambda} f(\xi_\mu) \right) = (1-\varepsilon^2) \left(\frac{\lambda\delta}{\Lambda} \right)^2 f(\xi_\mu),$$

and therefore, using (25.18),

$$f(\xi_{\mu+1}) = f(\xi_\mu) - \Delta_\mu \leq f(\xi_\mu) \left(1 - \frac{1-\varepsilon^2}{\Lambda^2} \lambda^2 \delta^2 \right).$$

This can be written, putting

$$\theta = \sqrt{1 - (1-\varepsilon^2)(\lambda\delta/\Lambda)^2}, \quad (26.14)$$

in the form

$$f(\xi_{\mu+1}) \leq \theta^2 f(\xi_\mu).$$

Therefore, for any positive integer N , we have

$$\frac{f(\xi_{\mu+N})}{f(\xi_\mu)} \leq \theta^{2N}.$$

Combining this with (26.13) we obtain

$$\frac{|\xi_{\mu+N}|_2^2}{|\xi_\mu|_2^2} \leq \theta^{2N} \frac{\Lambda}{\lambda}. \quad (26.15)$$

If we now choose N so large that we have

$$\theta^{2N} \frac{\Lambda}{\lambda} < 1,$$

relation (26.5) holds and Theorem 26.2 is proved.

CONDITION FOR THE REGULAR MINIMUM OF THE FUNCTION (25.3)

13. If Theorems 25.1 and 26.2 are applied to the solution of the system (25.2) using (25.3), the following theorem is useful:

Theorem 26.3. *Assume that Eqs. (25.2) with $n \geq 2$ are satisfied at a point ζ and that at this point the second derivatives of the functions $f_v(\xi)$ with respect to the variables x_1, \dots, x_n are continuous. Then in order that the function (25.3) have in ζ a regular minimum, it is necessary and sufficient that the Jacobian of the f_v with respect to the x_μ does not vanish in ζ .*

Proof. We have from (25.3)

$$\frac{1}{2}f'_{x_\kappa} = \sum_{v=1}^n f_v f'_{vx_\kappa}, \quad \frac{1}{2}f''_{vx_\kappa x_\lambda} = \sum_{v=1}^n f'_{vx_\lambda} f'_{vx_\kappa} + \sum_{v=1}^n f_v f''_{vx_\kappa x_\lambda}.$$

Then from (25.2) it follows that

$$f'(\zeta) = 0.$$

Form with the components x_1, \dots, x_n of the general point ξ the quadratic form corresponding to $f''(\zeta)$. We obtain

$$2 \sum_{v=1}^n \sum_{\kappa=1}^n f'_{vx_\kappa}(\zeta) x_\kappa \sum_{\lambda=1}^n f'_{vx_\lambda}(\zeta) x_\lambda + 2 \sum_{v=1}^n f_v(\zeta) \sum_{\kappa, \lambda=1}^n f''_{vx_\kappa x_\lambda}(\zeta) x_\kappa x_\lambda.$$

Here the second term vanishes, since all $f_v(\zeta) = 0$. The first term can be written as

$$2 \sum_{v=1}^n \left(\sum_{\kappa=1}^n f'_{vx_\kappa}(\zeta) x_\kappa \right)^2.$$

This is certainly ≥ 0 and can only be $= 0$ if we have

$$\sum_{\kappa=1}^n f'_{vx_\kappa}(\zeta) x_\kappa = 0 \quad (v = 1, \dots, n).$$

But these equations can only be satisfied by a nontrivial set of values of the x_ν if and only if the Jacobian of the f_v vanishes at ζ . This proves our theorem.

Observe that Theorem (26.3) is no longer true for $n = 1$.

ALGEBRAIC EQUATIONS WITH ONE UNKNOWN

14. Applying our results to the case of analytic functions as discussed in Sections 14–17 of the preceding chapter, it is an immediate corollary of Theorem 26.1 that the sequence z_μ constructed according to the rule of Theorem 25.1 is always convergent either to a zero of $f(z)$ or to a zero of $f'(z)$. Indeed, since $f(z)$ is assumed regular on $\Omega + S$, $f(z)\overline{f'(z)}$ has only a finite number of zeros on this closed set, and Ω^* is finite.

The occurrence of the zeros of $f'(z)$ in this connection is due to the fact that the singular points of the level curves of $|f(z)|$ lie only in the points where $f'(z)$ vanishes. So far, the geometric probability of the convergence to a zero of $f'(z)$ is zero. However, the situation in the real case could easily bring about such a convergence.

If, for instance, $f(z)$ is a polynomial with real coefficients and no real zeros, the iteration by (25.29) cannot lead to a complex zero of $f(z)$ if we start with a *real* z_0 . Therefore the corresponding sequence z_v must converge to a real zero of $f'(z)$. This difficulty can, however, be overcome, as will be shown in Chapter 28.

27

Method of Steepest Descent. Linear Convergence of the ξ_μ

CONDITIONS FOR STRICTLY LINEAR CONVERGENCE

1. By narrowing down the conditions of Theorem 26.2 we can even ensure that the convergence of the ξ_μ to ζ becomes *strictly linear*.

Theorem 27.1. *Assume about Ω , S , $f(\xi)$ and ξ_0 the conditions of Theorem 25.1. Form the sequence ξ_ν in (25.12), taking there*

$$r_\mu = \alpha_\mu \kappa_\mu, \quad (27.1)$$

where ψ_μ is restricted by the conditions $|\psi_\mu| = 1$ and (25.7), and κ_μ , φ_μ are given by (25.4). Assume that the ξ_μ tend to a point ζ inside Ω in which $f(\xi)$ has a regular minimum. Put

$$f''(\zeta) =: f_0'', \quad \Lambda(f_0'') =: \Lambda_0, \quad \lambda(f_0'') =: \lambda_0. \quad (27.2)$$

Take a positive number p satisfying the condition

$$p < \text{Min}\left(\frac{2}{\lambda_0 + \Lambda_0}, \frac{\lambda_0^2}{2\Lambda_0^2}\right) \quad (27.3)$$

and put

$$\delta = 1 - \frac{1}{2} \left(\frac{\lambda_0}{\Lambda_0} \right)^2 + p, \quad (27.4)$$

Assume further that the α_μ lie in the interval

$$p \leq \alpha_\mu \leq \frac{2}{\lambda_0 + \Lambda_0}, \quad (27.5)$$

and that we have $\delta_\mu \geq \delta$. Then the ξ_μ converge to ζ linearly and we have more precisely

$$\lim_{\mu \rightarrow \infty} \frac{|\xi_{\mu+1} - \zeta|}{|\xi_\mu - \zeta|} \leq 1 - \frac{p^2 \Lambda_0^2}{\lambda_0}. \quad (27.6)$$

2. Proof. Without loss of generality we can assume $\zeta = 0$.

Formula (25.12) can be written, using (27.1), as

$$\xi_{\mu+1} - \xi_{\mu} = -\alpha_{\mu} \kappa_{\mu} \psi_{\mu} = -\alpha_{\mu} \kappa_{\mu} \varphi_{\mu} + \alpha_{\mu} \kappa_{\mu} (\varphi_{\mu} - \psi_{\mu}). \quad (27.7)$$

Here we have, by (25.4), $\kappa_{\mu} \varphi_{\mu} = f'(\xi_{\mu})$. Apply to each component of the vector $f'(\xi_{\mu})$ the mean value theorem. Since $f'(0) = 0$, we can write

$$f'(\xi_{\mu}) = \left(\frac{\partial^2 f(\theta_i \xi_{\mu})}{\partial x_i \partial x_j} \right) \xi_{\mu},$$

where the numbers θ_i lie between 0 and 1. As the second derivatives of f are continuous at ζ , the elements of the matrix $(\partial^2 f(\theta_i \xi_{\mu}) / \partial x_i \partial x_j)$ tend to the corresponding elements of $f''(0) = f''_0$. We can therefore write

$$f'(\xi_{\mu}) = f''_0 \xi_{\mu} + o(|\xi_{\mu}|) \quad (27.8)$$

and, introducing this into (27.7),

$$\xi_{\mu+1} = (I - \alpha_{\mu} f''_0) \xi_{\mu} + \alpha_{\mu} \kappa_{\mu} (\varphi_{\mu} - \psi_{\mu}) + o(|\xi_{\mu}|). \quad (27.9)$$

3. Since we have $|\varphi_{\mu}| = |\psi_{\mu}| = 1$, we have, denoting the components of φ_{μ} by h_1, \dots, h_n and those of ψ_{μ} by k_1, \dots, k_n , $\sum_{i=1}^n h_i^2 = \sum_{i=1}^n k_i^2 = 1$. Hence we get, using (27.4)

$$\begin{aligned} |\varphi_{\mu} - \psi_{\mu}|^2 &= \sum_1^n (h_i - k_i)^2 = 2 - 2 \sum_1^n h_i k_i = 2 - 2\delta_{\mu} \leq 2 - 2\delta, \\ |\varphi_{\mu} - \psi_{\mu}| &= \sqrt{2 - 2\delta} = \frac{1}{\Lambda_0} \sqrt{\lambda_0^2 - 2p\Lambda_0^2}. \end{aligned} \quad (27.10)$$

Further, using (27.8) and (23.13),

$$\kappa_{\mu} = |f'(\xi_{\mu})| = |f''_0 \xi_{\mu}| + o(|\xi_{\mu}|) \leq \Lambda_0 |\xi_{\mu}| + o(|\xi_{\mu}|).$$

As to the first right-hand term of (27.9), we obtain, putting $M_{\mu} := \Lambda(I - \alpha_{\mu} f''_0)$,

$$|(I - \alpha_{\mu} f''_0) \xi_{\mu}| \leq M_{\mu} |\xi_{\mu}|. \quad (27.11)$$

4. Now I say that we have

$$M_{\mu} = \text{Max}(|1 - \lambda_0 \alpha_{\mu}|, |1 - \Lambda_0 \alpha_{\mu}|). \quad (27.12)$$

Indeed, the eigenvalues of $I - \alpha_{\mu} f''_0$ are not changed if we multiply this matrix from the left by an arbitrary orthogonal matrix U and from the right by its inverse U' . But as is well known, since f''_0 is a symmetric positive definite matrix, U can be chosen in such a way that the matrix $U f''_0 U'$ becomes a diagonal matrix with the eigenvalues of f''_0 along the diagonal

$$0 < \lambda_1 \leq \dots \leq \lambda_n.$$

Therefore the eigenvalues of $I - \alpha_\mu f_0''$ are

$$1 - \lambda_1 \alpha_\mu, \quad 1 - \lambda_2 \alpha_\mu, \dots, 1 - \lambda_n \alpha_\mu.$$

But now it follows that

$$\Lambda(I - \alpha_\mu f_0'') = \max_v |1 - \lambda_v \alpha_\mu| = \max(|1 - \lambda_1 \alpha_\mu|, |1 - \lambda_n \alpha_\mu|)$$

and (27.12) follows since $\lambda_1 = \lambda_0$, $\lambda_n = \Lambda_0$. Equation (27.12) is proved.

Since α_μ satisfies (27.5), we have

$$-(1 - \alpha_\mu \Lambda_0) \leq 1 - \alpha_\mu \lambda_0$$

and therefore by (27.12)

$$M_\mu = 1 - \alpha_\mu \lambda_0.$$

5. Introducing these estimates into (27.9), we get

$$\begin{aligned} \frac{|\xi_{\mu+1}|}{|\xi_\mu|} &\leq M_\mu + \alpha_\mu \sqrt{\lambda_0^2 - 2p\Lambda_0^2} + o(1), \\ \frac{|\xi_{\mu+1}|}{|\xi_\mu|} &\leq 1 - \alpha_\mu (\lambda_0 - \sqrt{\lambda_0^2 - 2p\Lambda_0^2}) + o(1). \end{aligned} \quad (27.13)$$

But here

$$\lambda_0 - \sqrt{\lambda_0^2 - 2p\Lambda_0^2} = \frac{2p\Lambda_0^2}{\lambda_0 + \sqrt{\lambda_0^2 - 2p\Lambda_0^2}} > \frac{p\Lambda_0^2}{\lambda_0}$$

and therefore, by (27.5),

$$\frac{|\xi_{\mu+1}|}{|\xi_\mu|} \leq 1 - p\alpha_\mu \frac{\Lambda_0^2}{\lambda_0} + o(1) \leq 1 - \frac{p^2\Lambda_0^2}{\lambda_0} + o(1).$$

We obtain relation (27.6).

AN EXAMPLE

6. We show now by an example that if we choose t in Section 6 of Chapter 25 each time so as to make $f(\xi_\mu - t\varphi_\mu)$ minimum, the convergence of the ξ_μ is still in the general case only linear. We take, in the two-dimensional plane, the ellipse E_0 :

$$f(\xi) := \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1, \quad a > b > 0.$$

Choose ξ_0 and ξ_0' on E_0 (see Fig. 6) in the first and in the fourth quadrant,

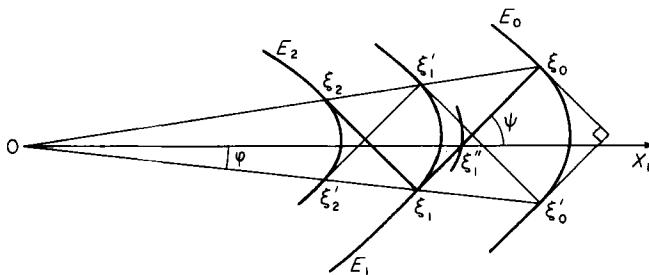


FIGURE 6

so that the tangents to E_0 in these points are orthogonal. If we go from ξ_0 along the normal so as to make $f(\xi_0 - t\varphi_0)$ a minimum, we come to a point ξ_1 at which the normal touches an ellipse E_1 similar and similarly situated to E_0 , and ξ_1 lies on the line $O\xi_0'$, so that we have

$$\xi_1 = \sigma \xi_0', \quad 0 < \sigma < 1.$$

Correspondingly, the normal to E_0 at ξ_0' touches E_1 at a point ξ_1' so that the configuration (E_0, ξ_0, ξ_0') is similar to the configuration (E_1, ξ_1, ξ_1') . We have then

$$\xi_1' = \sigma \xi_0.$$

7. Starting now from ξ_1 and ξ_1' we arrive in a similar way at the points ξ_2 and ξ_2' on the lines $O\xi_0$, $O\xi_0'$, so that

$$\xi_2 = \sigma \xi_1', \quad \xi_2' = \sigma \xi_1.$$

Continuing in the same way, we have generally

$$\xi_{v+1} = \sigma \xi_v', \quad \xi_{v+1}' = \sigma \xi_v. \quad (27.14)$$

Now put

$$q_v = |\xi_{v+1}| / |\xi_v|.$$

Then it follows from (27.14), eliminating the ξ' , that

$$q_{v+1} = q_{v-1}, \quad q_v q_{v+1} = \sigma^2.$$

Further, we easily obtain

$$\xi_{2v+i} = \sigma^{2v} \xi_i \quad (i = 0, 1; \quad v = 1, 2, \dots).$$

We see that the convergence of the ξ_v to zero is not better than linear. But it is not evident that both q_0, q_1 are < 1 .

8. We are going to prove now that the convergence of the ξ_v to zero is indeed *linear*, proving that

$$\frac{|\xi_{v+1}|}{|\xi_v|} \leq \frac{c^2}{a^2}, \quad c^2 = a^2 - b^2 \quad (v = 0, 1, \dots). \quad (27.15)$$

It is sufficient to prove (27.15) for $v = 0$.

Let the line $\xi_0 \xi_1$ cut the x_1 -axis at ξ''_1 . Denote the angles $\xi_0 \xi''_1 x_1$ by ψ and $\xi_1 0 x_1$ by φ . The equation of the normal to the ellipse E_0 at ξ_0 is in running coordinates u, v ,

$$\frac{ux_2}{b^2} - \frac{vx_1}{a^2} = \frac{c^2}{a^2 b^2} x_1 x_2,$$

where we have for the coordinates (x_1, x_2) of ξ_0 in the usual parametrical representation

$$x_1 = a \cos t, \quad x_2 = b \sin t, \quad 0 < t < \pi/2.$$

Putting $v = 0$, we obtain $u = (c^2/a^2)x_1$,

$$|\xi''_1| = \frac{c^2}{a^2} x_1. \quad (27.16)$$

On the other hand, denoting the coordinates of ξ'_0 by (x'_1, x'_2) , we have

$$x'_1 = a \cos t', \quad x'_2 = b \sin t', \quad -\pi/2 < t' < 0,$$

and since the tangents at ξ_0 and ξ'_0 are orthogonal, as the orthogonality condition

$$\tan t \tan |t'| = \frac{b^2}{a^2}. \quad (27.17)$$

9. We have now for the angle ψ from Fig. 6

$$\begin{aligned} \tan \psi &= \frac{x_2}{x_1 - |\xi''_1|} = \frac{x_2}{x_1} \frac{1}{1 - c^2/a^2}, \\ \tan \psi &= \frac{a^2 b}{b^2 a} \tan t = \frac{a}{b} \tan t. \end{aligned}$$

As to the angle φ , we have

$$\tan \varphi = \frac{|x'_2|}{x'_1} = \frac{b}{a} \tan |t'|$$

and by (27.17)

$$\tan \varphi = \frac{b^3}{a^3} \cot t.$$

It now follows that

$$\tan(\varphi + \psi) = \frac{(a/b) \tan t + (b^3/a^3) \cot t}{1 - (b^2/a^2)} = \frac{1}{c^2} \left(\frac{a^3}{b} \tan t + \frac{b^3}{a} \cot t \right).$$

Since the right-hand expression remains finite for $0 < t < \pi/2$, it follows that $\varphi + \psi$ remains either always in the interval $(0, \pi/2)$ or always in the interval $(\pi, 3\pi/2)$. On the other hand, we see from Fig. 6 that both angles φ and ψ are acute. Therefore $\varphi + \psi$ lies always between 0 and $\pi/2$ so that, as indicated in Fig. 6, the lines $0\xi_1, 0x_1, 0\xi_0$ follow in that order after the perpendicular from 0 to the normal $\xi_0 \xi_1$. We see in particular that $|\xi_1| \leq |\xi''_1|$ and from (27.16) we have now

$$\frac{|\xi_1|}{|\xi_0|} \leq \frac{|\xi''_1|}{|\xi_0|} = \frac{c^2}{a^2} \frac{1}{\sqrt{1 + (x_2/x_1)^2}} \leq \frac{c^2}{a^2}.$$

This indeed proves that the convergence of ξ_v is linear and also that

$$\sigma \leq \frac{c^2}{a^2}.$$

10. Figure 6 also illustrates the fact that it is not the best strategy in our case to go along the normal $\xi_0 \xi_1$ all the way through to ξ_1 .

Indeed, if we stop instead at ξ''_1 , the normal to the corresponding ellipse at this point goes directly through the center, so that only one further step is sufficient.

CONNECTION WITH THE NEWTON-RAPHSON PROCEDURE

11. We now discuss another example of the method of steepest descent in order to establish its connection with the Newton-Raphson method. We use the assumptions and formulas of Section 14 of Chapter 25, and assume that $f(z)$ has a simple zero at $z = 0$. Then if z is a starting value for the iteration with $f'(z) \neq 0$, it follows from (25.22) that the next value would be

$$z' = z - TZ(z) = z - 2Tf(z)\overline{f'(z)}$$

for a convenient choice of $T > 0$.

We easily see that by putting $t = 2T|f'(z)|^2$, z' becomes

$$z' = z - tf(z)/f'(z). \quad (27.18)$$

The choice $t = 1$ corresponds to the Newton-Raphson procedure. What can be said about the value of t_C corresponding to Cauchy's prescription?

We have to consider the first minimum of $F(z')$ if t increases beginning with $t = 0$. In order to obtain $\partial F(z')/\partial t$, observe that if we put $z' = x' + iy'$ it

follows from (25.22) that

$$\begin{aligned}\frac{\partial F(z')}{\partial t} &= F'_{x'} \frac{\partial x'}{\partial t} + F'_{y'} \frac{\partial y'}{\partial t} = R(F'_{x'} + iF'_{y'}) \left(\frac{\partial z'}{\partial t} \right) \\ &= -R(Z(z') \overline{f(z)/f'(z)}) = -2|f(z)|^2 R \left(\frac{Z(z')}{Z(z)} \right).\end{aligned}$$

If we put further $1-t =: \tau$, we obtain

$$\frac{1}{2|f(z)|^2} \frac{\partial F(z')}{\partial \tau} = R \left(\frac{Z(z')}{Z(z)} \right).$$

12. Since in our discussion we can multiply f and z by convenient non-vanishing constants, we can assume without loss of generality that $f'(0) = 1$, $f''(0) = 2$, disregarding the exceptional case where $f''(0) = 0$. Put

$$z = x + iy, \quad |z| =: \rho;$$

then we can write

$$f(z) = z + z^2 + O(\rho^3), \quad f'(z) = 1 + 2z + O(\rho^2). \quad (27.19)$$

It easily follows that

$$\begin{aligned}\frac{f(z)}{f'(z)} &= z(1-z) + O(\rho^3), \\ z' &= z(\tau + (1-\tau)z) + O(\rho^3).\end{aligned} \quad (27.20)$$

It then follows from (25.22) and (27.19) that

$$\frac{1}{2}Z(z) = f(z) \overline{f'(z)} = (z+z^2+2\rho^2) + O(\rho^3). \quad (27.21)$$

Using (27.20) it follows further, since $|z'| = O(\rho)$, that

$$\frac{1}{2}Z(z') = \tau z + (1-\tau)z^2 + \tau^2 z^2 + 2|z'|^2 + O(\rho^3).$$

On the other hand, from (27.20) it follows that $|z'| = \tau\rho + O(\rho^2)$ and therefore finally

$$\frac{1}{2}Z(z') = \tau z + (1-\tau)z^2 + \tau^2 z^2 + 2\tau^2 \rho^2 + O(\rho^3).$$

We now obtain

$$\begin{aligned}\frac{Z(z')}{Z(z)} &= \frac{\tau + (1-\tau+\tau^2)z + 2\tau^2 \bar{z} + O(\rho^2)}{1 + z + 2\bar{z} + O(\rho^2)} \\ &= \tau + ((1-\tau+\tau^2)z + 2\tau^2 \bar{z} - \tau z - 2\tau \bar{z}) + O(\rho^2), \\ R \left(\frac{Z(z')}{Z(z)} \right) &= \tau + (1-4\tau+3\tau^2)x + O(\rho^2).\end{aligned} \quad (27.22)$$

13. From this formula we see that $F(z')$ is monotonically decreasing with decreasing τ , that is, with increasing t , as long as t does not come into a neighborhood of 1, which can be taken arbitrarily small choosing ρ sufficiently small. But in the neighborhood of $t = 1$ there is a change of sign of the derivative $\partial F(z')/\partial t$.

There exists, therefore, one zero $\tau = \tau_C$ of (27.22) which tends to 0 with $\rho \downarrow 0$.

Putting this $\tau_C = \sigma - x$, introducing this into the right-hand expression in (27.22), and letting this = 0, we obtain easily

$$\sigma = 4\sigma x - 3\sigma^2 x + O(\rho^2).$$

Since, however, σ is bounded, it follows from this formula that $\sigma = O(\rho)$ and then again $\sigma = O(\rho^2)$. We see that we have, for the first minimum of $F(z')$,

$$t_C = 1 + x + O(\rho^2). \quad (27.23)$$

Our final result is that for sufficiently small ρ the value of t corresponding to the Cauchy procedure is obtained from (27.23).

The corresponding iteration procedure is also converging quadratically. Indeed, we have

$$z' = z - t_C \frac{f(z)}{f'(z)} = z - \frac{f(z)}{f'(z)} - x \frac{f(z)}{f'(z)} + O(\rho^2) = O(\rho^2),$$

since $z - f(z)/f'(z) = O(\rho^2)$ and $f(z)/f'(z) = O(\rho)$, as follows from Chapter 7.

28

Convergent Procedures for Polynomial Equations

THE FIRST STEP OF THE PROCEDURE

1. From what has been said in Chapter 27, we can set up a procedure, going in the direction of steepest descent, in the form

$$z' = z - t \frac{f(z)}{f'(z)} \quad (t > 0), \quad (28.1)$$

and the problem is to find a positive value of t such that $|f(z')| < |f(z)|$. Such a value of t can be obtained from the discussion in Chapter 27. In this chapter, however, we will proceed in a more direct way.

In our discussion we will need some asymptotic relations. Assume ζ a zero of $f(z)$ of exact multiplicity $p \geq 1$; then developing $f(z)$ at ζ , we have, as $z \rightarrow \zeta$,

$$f(z) \sim \frac{f^{(p)}(\zeta)}{p!} (z - \zeta)^p \quad (z \rightarrow \zeta). \quad (28.2)$$

Further, developing $f'(z)$ at ζ , we have

$$f'(z) \sim \frac{f^{(p)}(\zeta)}{(p-1)!} (z - \zeta)^{p-1} \quad (z \rightarrow \zeta). \quad (28.3)$$

Dividing, we obtain

$$\frac{f(z)}{f'(z)} \sim \frac{1}{p} (z - \zeta) \quad (z \rightarrow \zeta) \quad (28.4)$$

and

$$\frac{|f'(z)|^2}{|f(z)|} \sim \frac{p}{(p-1)!} |f^{(p)}(\zeta)| |z - \zeta|^{p-2} \quad (z \rightarrow \zeta). \quad (28.5)$$

2. The equation to be solved can be assumed in the “reduced form”:

$$f(z) := z^n + a_2 z^{n-2} + \cdots + a_n = 0, \quad |a_v| \leq 1 \quad (2 \leq v \leq n). \quad (28.6)$$

The reduction of a general polynomial of degree n to a “reduced form” is given in Appendix R. In this appendix we also prove three properties of reduced polynomials which will be important in our discussion:

(1) There exists a positive number ρ^* depending only on n and $<(\sqrt{5}+1)/2 = 1.6180\dots$ such that for $|z| > \rho^*$ we have everywhere $|f(z)| > 1$ and even

$$|f(z)| \geq \left| \frac{z}{\rho^*} \right|^n \quad (|z| \geq \rho^*). \quad (28.7)$$

(2) There exists a positive number M depending only on n such that

$$|f''(z)| \leq M \quad (|z| \leq \rho^*). \quad (28.8)$$

We give in Appendix R tables of values of ρ^* and M for $n = 3, \dots, 20$.

(3) For $0 \leq v < n$ the inequalities hold:

$$\frac{|f^{(v)}(z)|}{v!} \leq 2 \binom{n}{v} |z|^{n-v} \quad (|z| \geq \rho^*). \quad (28.9)$$

3. Theorem 28.1. Consider the circle $K(|z| < \rho^*)$ and assume that z lies in K and that $f(z)f'(z) \neq 0$; put

$$T := T(z) = \frac{|f'(z)|^2}{M|f(z)|}, \quad t^* := \text{Min}(1, T). \quad (28.10)$$

Take $0 < t \leq t^*$ and assume that

$$z - t \frac{f(z)}{f'(z)} \in K. \quad (28.11)$$

Then we have, in notation (28.1),

$$\left| \frac{f(z')}{f(z)} \right| \leq 1 - \frac{t}{2}. \quad (28.12)$$

4. Proof. The whole interval $\langle z, z' \rangle$ lies in K together with z and z' . Developing $f(z')$ in powers of t , we obtain, denoting by θ^* a complex number with $|\theta^*| \leq 1$,

$$\begin{aligned} f(z') &= f(z) - f'(z)t \frac{f(z)}{f'(z)} + \frac{t^2}{2}\theta^*M \frac{|f(z)|^2}{|f'(z)|^2} \\ &= f(z) \left(1 - t + \theta^* \frac{t^2}{2T} \right), \quad |\theta^*| \leq 1. \end{aligned}$$

Therefore, since $t \leq 1$,

$$\left| \frac{f(z')}{f(z)} \right| \leq 1 - t + \frac{t^2}{2T} = 1 - t \left(1 - \frac{t}{2T} \right).$$

But, by definition of t^* ,

$$1 - \frac{t}{2T} \geq 1 - \frac{t^*}{2T} \geq \frac{1}{2}$$

and (28.12) is proved.

5. Assume now that

$$|f(z)| \leq 1, \quad |z| < \rho^*. \quad (28.13)$$

Then it follows easily that

$$\left| z - t^* \frac{f(z)}{f'(z)} \right| < \rho^*, \quad (28.14)$$

$$\left| f\left(z - t^* \frac{f(z)}{f'(z)} \right) \right| \leq 1 - \frac{t^*}{2}. \quad (28.15)$$

Indeed, for sufficiently small positive $t < t^*$ (28.11) is obviously true, and therefore so is (28.12), by Theorem 28.1. If (28.14) were false, there would exist a t' with $0 < t' \leq t^*$ such that

$$\left| z - t' \frac{f(z)}{f'(z)} \right| = \rho^*,$$

while (28.11) remains true for $0 < t < t'$. But then it follows from (28.12), as $t \uparrow t'$, that

$$\left| f\left(z - t' \frac{f(z)}{f'(z)} \right) \right| \leq 1 - \frac{t'}{2} < 1$$

in contradiction with (28.7).

Therefore (28.11) also remains true for $t = t^*$, and (28.15) follows from Theorem 28.1.

CONVERGENCE OF THE ITERATION PROCEDURE

6. Assume now that (28.13) holds; then if we put

$$z_0 := z, \quad t_0 := t^*, \quad z_1 := z_0 - t_0 \frac{f(z_0)}{f'(z_0)},$$

we see that (28.13) remains true, replacing z with z_1 , and that we have also

$$|f(z_1)| \leq |f(z_0)| \left(1 - \frac{t_0}{2} \right) < 1.$$

If $f(z_1)f''(z_1) \neq 0$, we can therefore proceed in the same way to z_2 , and so on indefinitely, and obtain the sequence z_v defined by

$$z_{v+1} = z_v - t_v \frac{f(z_v)}{f'(z_v)}, \quad t_v = \text{Max}\left(1, \frac{|f'(z_v)|^2}{M|f(z_v)|}\right) \quad (v = 0, 1, \dots) \quad (28.16)$$

where $z_0 := z$, all z_v lie in K , and

$$|f(z_{v+1})| \leq |f(z_v)| \left(1 - \frac{t_v}{2}\right). \quad (28.17)$$

All this is so, however, only if $f(z_v)f'(z_v)$ remains $\neq 0$.

If then we obtain for a z_v :

$$f(z_v)f'(z_v) = 0,$$

the procedure stops with this z_v . Otherwise we obtain an infinite sequence z_v , and we are now going to show that then the sequence (28.16) always converges to a zero of $f(z)f'(z)$.

7. From (28.17) it follows that

$$|f(z_v)| \downarrow m \geq 0.$$

We consider the cases where $m > 0$ and $m = 0$ separately.

If $m > 0$, then by (28.17) $t_v \rightarrow 0$ and therefore, by definition of t_v , from a certain v on

$$t_v = \frac{|f'(z_v)|^2}{M|f(z_v)|} \rightarrow 0, \quad f'(z_v) \rightarrow 0. \quad (28.18)$$

We see that here all accumulation points of the z_v are zeros of $f'(z)$.

On the other hand, by (28.16) and (28.18)

$$|z_v - z_{v+1}| = t_v \frac{|f(z_v)|}{|f'(z_v)|} = \frac{|f'(z_v)|}{M} \rightarrow 0$$

and it follows from Theorem 25.1 that the z_v converge to a zero of $f'(z)$. In this case $f(z_v)/f'(z_v) \rightarrow \infty$ and we speak of the “spurious convergence” of the z_v .

8. In the case where $m = 0$ we have

$$|f(z_v)| \downarrow 0$$

and it follows that all accumulation points of the z_v are zeros of $f(z)$. But then $f(z_v)/f'(z_v)$ must tend to 0, since if for a partial sequence z_{v_k} we had

$$\frac{f(z_{v_k})}{f'(z_{v_k})} \rightarrow a \neq 0, \quad |a| \leq \infty,$$

we could assume, sieving, if necessary, the z_{v_k} once again through, that $z_{v_k} \rightarrow \zeta, f(\zeta) = 0$. But then it follows from (28.4)

$$\frac{f(z_{v_k})}{f'(z_{v_k})} \sim \frac{\zeta - z_{v_k}}{p}$$

if p is the exact multiplicity of ζ , that $a = 0$.

Since on the other hand $t_v \leq 1$, it follows that $z_v - z_{v+1} \rightarrow 0$ and by Theorem 25.1 the z_v must converge to a zero of $f(z)$.

We see further that z_v converges to a zero of $f(z)$ or to a zero of $f'(z)$, different from all zeros of $f(z)$, according as $f(z_v)/f'(z_v) \rightarrow 0$ or $f(z_v)/f'(z_v) \rightarrow \infty$.

SWITCHING OVER TO THE NEWTON-RAPHSON PROCEDURE

9. If we form the sequence z_v by the rule (28.16), we have to discuss whether the expression $|f'(z_v)|^2/(M|f(z_v)|)$ is < 1 or ≤ 1 . It is advisable at the same time to check whether this expression is > 2 . Indeed, if we have

$$\frac{|f'(z_v)|^2}{M|f(z_v)|} > 2, \quad (28.19)$$

it follows from Theorem 7.2 that, starting the Newton-Raphson procedure from the corresponding z_v , we obtain a quadratic convergence to a zero of $f(z)$. The corresponding command under the condition (28.19) has to be incorporated into the program of the computation.

THE Ω -TEST

10. Assuming that the sequence (28.16) tends to a zero ζ of $f(z)$, we are now going to indicate a rule for obtaining the distance of the approximation z_v from the nearest zero of $f(z)$ or the nearest group of zeros of $f(z)$, lying close together. This rule is based on the theorem:

Theorem 28.2. Assume that $\varphi(u)$ is defined in the circle $|u| \leq R$ for a positive R and can be represented in this circle by

$$\varphi(u) = \sum_{\mu=0}^m D_\mu u^\mu + \theta^* M_{m+1} u^{m+1}, \quad M_{m+1} > 0, \quad |\theta^*| \leq 1 \quad (|u| \leq R), \quad (28.20)$$

where $m > 0$ and M_{m+1} is a positive constant. Assume further that for a positive $\rho \leq R$ we have

$$\frac{|D_m|}{2M_{m+1}} \geq \rho \quad (28.21)$$

and for an r satisfying

$$0 < r \leq \rho, \quad (28.22)$$

$$2m \frac{|D_\mu|}{|D_m|} \leq r^{m-\mu} \quad (\mu = 0, \dots, m-1). \quad (28.23)$$

Assume finally that in at least one of the $m+2$ inequalities (28.21), (28.22), (28.23) the equality sign is excluded.

Then $\varphi(u)$ has exactly m zeros in $|u| \leq \rho$ and all these zeros lie in the circle $|u| < r$.

Remark. Observe that the assertion of the theorem is no longer necessarily true if the equality sign holds in all relations (28.21) and (28.23) and $r = \rho$. A counterexample is given by $\varphi(u) = u^2 + 2u + 1$, $m = \rho = r = M_2 = 1$.

11. Proof. Without loss of generality we can assume $D_m = 1$, as we can divide φ by D_m .

We have, from (28.20),

$$|\varphi(u) - u^m| \leq \sum_{\mu=0}^{m-1} |D_\mu| |u|^\mu + M_{m+1} |u|^{m+1}$$

and, using (28.21) and (28.23),

$$|\varphi(u) - u^m| \leq \frac{1}{2m} \sum_{\mu=0}^{m-1} r^{m-\mu} |u|^\mu + \frac{|u|}{2\rho} |u|^m \quad (28.24)$$

where we can even write $<$ instead of \leq , if in one of the $m+1$ relations (28.21), (28.23) the equality sign is excluded.

If we now assume $r \leq |u| \leq \rho$, it follows further that

$$|\varphi(u) - u^m| < \frac{1}{2m} \sum_{\mu=0}^{m-1} |u|^\mu + \frac{|u|^m}{2} = |u|^m, \quad (28.25)$$

since by our hypotheses, if we have the equality sign in (28.24), we must have $r < \rho$ and then either $|u| > r$ or $|u| < r$.

From (28.25) the assertion of our theorem follows immediately by Rouché's theorem.

12. Applying our theorem, we will put $\varphi(u) := f(z_v + u)$, $R_0 = 1.781 > (\sqrt{5} + 1)/2 + \frac{1}{10}$, and define M_{m+1} by

$$M_{m+1} = \text{Max} \frac{|f^{(m+1)}(z)|}{(m+1)!} \quad (|z| \leq R_0).$$

Then, since $|z_v| < (\sqrt{5} + 1)/2$, it follows that

$$\frac{|\varphi^{(m+1)}(u)|}{(m+1)!} \leq M_{m+1} \quad (|u| \leq \frac{1}{10}).$$

In practice, since it is not worthwhile computing the exact value M_{m+1} , it will be sufficient to use the value following from (28.9):

$$M_{m+1} := 2 \binom{n}{m+1} (1.7181)^{n-m-1}.$$

This M_{m+1} can be used in Theorem 28.2 if we take $\rho \leq \frac{1}{10}$. But then it follows from our theorem, as the Ω -test:

Assume that

$$\min\left(\frac{|f^{(m)}(z_v)|}{m! 2M_{m+1}}, \frac{1}{10}\right) =: \rho \geq r := \max_{0 \leq \mu \leq m-1} \left(2m \frac{|f^{(\mu)}(z_v)| m!}{|f^{(m)}(z_v)| \mu!}\right)^{1/(m-\mu)}, \quad (28.26)$$

where, however, if the left-hand side expression is $= |f^{(m)}(z_v)| / (2m! M_{m+1})$ and the m expressions in parentheses on the right-hand side are equal, $\rho \geq r$ must be replaced with $\rho > r$. Then there are in the closed ρ -neighborhood of z_v exactly m zeros of $f(z)$ and these zeros lie in the open r -neighborhood of z_v .

Of course, $\frac{1}{10}$ could be replaced with other positive constants, changing M_{m+1} correspondingly.

13. Here r gives the “precision” with which the z_v approximate a zero, or a group of zeros, of $f(z)$.

To make r small we must make $|D_0| = |f(z_v)|$ small and then try to find an m , beginning with $m = 1$, for which the Ω -test is positive. However, since for $m = 1$ the test is practically the same as has to be applied at every step to check the applicability of the Newton–Raphson procedure, it is sufficient to begin the Ω -test with $m = 2$. Further, for $m = n$ it is enough to consider

$$r = \max_{\mu} \left(2n \frac{|f^{(\mu)}(z_v)|}{\mu!} \right)^{1/(n-\mu)},$$

which gives in any case an upper limit for the greatest distance of z_v from all zeros of $f(z)$.

14. The Ω -test must be used at the end of the computation. Here we must distinguish two cases. If the computation is done with a *fixed number of decimals*, $z_v - z_{v+1}$ becomes finally undistinguishable from 0 and then we use the Ω -test, if necessary, with multiple precision, to find out what has been achieved in this case.

In the second case we give from the beginning the desirable order of magnitude of the error, that is, of r . Then we apply the Ω -test if it can be expected that the desired precision has been already attained. In order to recognize this, observe that $f(z_v)/f'(z_v) \rightarrow 0$ and, as soon as r/ρ in our test is sufficiently small, $|f(z_v)|/|f'(z_v)|$ does not essentially exceed r/m . This follows easily from formula (28.4).

In this case the program must include the command to apply the Ω -test as soon as $|f(z)|/|f'(z)|$ becomes less than a given small number.

15. If the sequence (28.16) breaks up, we have for the last z_v either $f(z_v) = 0$ or $f(z_v) \neq 0, f'(z_v) = 0$. In the first case, if $f(z_v)$ is exactly $= 0$, the aim of the computation is achieved. If, however, we only know that $f(z_v)$ is undistinguishable from 0 in the sense of the precision used and desire to obtain a completely secure information about the precision attained, the Ω -test ought to be applied with double, or if necessary with multiple, precision.

In the second case $f'(z_v)$ is 0 or undistinguishable from 0 and then this case has to be treated in the same way as the general case of spurious convergence:

$$z_v \rightarrow \zeta', \quad f'(\zeta') = 0, \quad f(\zeta') \neq 0.$$

Here we apply the so-called *J*-procedure, which will be described and discussed in the next chapter.

29

J-Test and *J*-Routine

BASIC THEOREM

1. We are now going to show how to modify the sequence z_v in the case of spurious convergence. Our method is based on the following theorem.

Theorem 29.1. *Assume, for an $R > 0$, $\varphi(u)$ analytic in $|u| \leq R$. Let its development at the origin be*

$$\varphi(u) = \sum_{\mu=0}^m C_\mu u^\mu + \theta^* M'_{m+1} u^{m+1}, \quad |\theta^*| \leq 1 \quad (|u| \leq R) \quad (29.1)$$

with $m > 1$ and

$$M'_{m+1} := \max_{|u|=R} \frac{|\varphi^{(m+1)}(u)|}{(m+1)!}.$$

Put $\kappa = \sqrt[m]{\frac{1}{3}}$ and assume $0 < r := \kappa \rho \leq \kappa R$.

Assume that the following relations hold:

$$M'_{m+1} \rho \leq \frac{m}{2m+2} |C_m|, \quad (29.2)$$

$$|C_\mu| \leq |C_m| \frac{r^{m-\mu}}{2(m^2-1)} \quad (1 \leq \mu \leq m-1), \quad (29.3)$$

$$|C_0| \geq |C_m| \rho^m. \quad (29.4)$$

Then (A) there exist in the circle $|u| \leq \rho$ exactly $m-1$ zeros of $\varphi'(u)$, $\zeta'_1, \dots, \zeta'_{m-1}$, and we have even

$$|\zeta'_v| < r \quad (v = 1, \dots, m-1); \quad (29.5)$$

(B) there exists a u_0 with $|u_0| = \rho$ such that

$$|\varphi(u_0)| < |\varphi(\zeta'_v)| \quad (v = 1, \dots, m-1). \quad (29.6)$$

2. Proof. In order to prove part (A) of the assertion it is sufficient to show that Theorem 28.2 is applicable to $\varphi'(u)$, replacing m with $m-1$. Writing

out the MacLaurin development of $\varphi'(u)$, we have obviously

$$\begin{aligned}\varphi'(u) &= \sum_{\mu=1}^{m-1} \mu C_\mu u^{\mu-1} + m C_m u^{m-1} + \theta^* M_m u^m, \quad |\theta^*| \leq 1, \\ M_m &= \underset{|u|=R}{\text{Max}} \frac{|\varphi^{(m+1)}(u)|}{m!} = (m+1) M'_{m+1}.\end{aligned}$$

If we now replace, in the conditions of Theorem 28.2, m with $m-1$, M'_{m+1} with $(m+1) M'_{m+1}$, and the D_0, \dots, D_{m-1} resp. with $C_1, 2C_2, \dots, mC_m$, the conditions corresponding to (28.21), (28.22) with $r < \rho$, and (28.23) are obviously satisfied and assertion (A) of our theorem follows.

3. In order to prove assertion (B), observe that we have for $|u| \leq \rho$, using (29.1), (29.2), and (29.3):

$$\begin{aligned}|\varphi(u) - C_0 - C_m u^m| &\leq \sum_{\mu=1}^{m-1} |C_\mu| |u|^\mu + M'_{m+1} \rho |u|^m \frac{|u|}{\rho} \\ &\leq \frac{|C_m|}{2(m^2-1)} \sum_{\mu=1}^{m-1} |u|^\mu r^{m-\mu} + \frac{m}{2m+2} |C_m| |u|^m, \\ |\varphi(u) - C_0 - C_m u^m| &\leq \frac{|C_m| |u|^m}{2m+2} \left(\frac{1}{m-1} \sum_{\mu=1}^{m-1} \left(\frac{r}{|u|} \right)^\mu + m \right) \quad (|u| \leq \rho). \quad (29.7)\end{aligned}$$

If we assume here $|u| = r$, the right-hand bound in (29.7) becomes $|C_m| r^m / 2$. Since φ is regular for $|u| \leq r$, it follows for $|u| \leq r$ that

$$|\varphi(u) - C_0 - C_m u^m| \leq \frac{|C_m| r^m}{2} \quad (|u| \leq r).$$

We obtain now from (29.5), since $r^m = \rho^m / 3$, as $|\zeta'| < r$,

$$\begin{aligned}|\varphi(\zeta_v')| &> |C_0| - |C_m| r^m - \frac{|C_m|}{2} r^m, \\ |\varphi(\zeta_v')| &> |C_0| - \frac{|C_m|}{2} \rho^m. \quad (29.8)\end{aligned}$$

4. On the other hand, it follows from (29.7) for $|u| = \rho$ that

$$\begin{aligned}|\varphi(u) - C_0 - C_m u^m| &\leq \frac{|C_m| \rho^m}{2}, \\ |\varphi(u)| &< |C_0 + C_m u^m| + |C_m| \frac{\rho^m}{2}.\end{aligned}$$

If $\psi := \arg C_0$, put

$$u_0 := \rho \exp\left(i \frac{\psi + \pi}{m}\right). \quad (29.9)$$

Then we obtain

$$|\varphi(u_0)| \leq |C_0| - |C_m| \rho^m + \frac{|C_m|}{2} \rho^m$$

and using (29.4)

$$|\varphi(u_0)| \leq |C_0| - |C_m| \rho^m + \frac{|C_m|}{2} \rho^m = |C_0| - \frac{|C_m|}{2} \rho^m.$$

Comparing this with (29.8), (29.6) follows and our theorem is proved.

THE J-TEST

5. In order to formulate the conditions for the applicability of the above theorem we have to eliminate ρ from the inequalities (29.2)–(29.4). Solving these inequalities with respect to ρ , we obtain

$$\begin{aligned} \rho &\geq \frac{1}{\kappa} \left(2(m^2 - 1) \left| \frac{C_\mu}{C_m} \right| \right)^{1/(m-\mu)} \quad (1 \leq \mu \leq m-1), \\ \rho &\leq \text{Min} \left(\left| \frac{C_0}{C_m} \right|^{1/m}, \frac{m}{2m+2} \frac{|C_m|}{M'_{m+1}} \right). \end{aligned}$$

Here we must fix a value for R , in order to be able to obtain M'_{m+1} . We choose $R = \frac{1}{10}$ and must therefore define in any case $\rho \leq \frac{1}{10}$. As on the other hand it is desirable to take ρ as large as possible, we can finally write our condition in the following way:

$$\text{Max}_{1 \leq \mu \leq m-1} \frac{1}{\kappa} \left(2(m^2 - 1) \left| \frac{C_\mu}{C_m} \right| \right)^{1/(m-\mu)} \leq \rho := \text{Min} \left(\frac{1}{10}, \left| \frac{C_0}{C_m} \right|^{1/m}, \frac{m}{2m+2} \frac{|C_m|}{M'_{m+1}} \right). \quad (29.10)$$

This criterion will be applied to the function $\varphi(u) := f(z_v + u)$ and can be then written in the form

$$\begin{aligned} \text{Min} \left(\frac{1}{10}, \left(m! \left| \frac{f(z_v)}{f^{(m)}(z_v)} \right| \right)^{1/m}, \frac{m}{2m+2} \frac{|f^{(m)}(z_v)|}{m! M'_{m+1}} \right) &= \rho \\ &\geq \text{Max}_{1 \leq \mu \leq m-1} 3^{1/m} \left(2(m^2 - 1) \left| \frac{f^{(\mu)}(z_v)}{f^{(m)}(z_v)} \right| \frac{m!}{\mu!} \right)^{1/(m-\mu)}. \end{aligned} \quad (29.11)$$

As to the value of M'_{m+1} , we can take it as

$$M'_{m+1} = \text{Max} \frac{|f^{(m+1)}(z)|}{(m+1)!} \quad (|z| \leq \rho^* + \frac{1}{10} < 1.7181)$$

or even, using (28.9), as simply

$$M'_{m+1} := 2 \binom{n}{m+1} (1.7181)^{n-m-1}.$$

For $m = n$, of course, we can take $M'_{m+1} = 0$ and in the corresponding condition (29.11) the last term under the sign of Minimum on the left can be omitted.

The conditions (29.11) with $m = 2, \dots, n$ together form what we call the *J-test*.

THE J_m -ROUTINE

6. If condition (29.11) is satisfied for an m , we form the next approximation z_{v+1} using the rule

$$\psi := \arg \frac{f(z_v)}{f^{(m)}(z_v)}, \quad u_0 := \rho \exp \left(i \frac{\psi + \pi}{m} \right), \quad z_{v+1} := z_v + u_0. \quad (29.12)$$

If the zeros of φ in $|u| \leq \rho$ are denoted by $\zeta'_1, \dots, \zeta'_{m+1}$, it follows from (29.6) that

$$|f(z_{v+1})| < |f(z_v + \zeta'_\mu)| \quad (\mu = 1, \dots, m-1). \quad (29.13)$$

Since for the following approximations $|f(z)|$ is strictly decreasing, we see that if we start from this z_{v+1} again the procedure (28.16), the convergence to the zeros of $f'(z)$ inside $|z - z_v| < \rho$ is now excluded. The procedure described in (29.12) is the *J_m-routine*. Obviously this routine cannot be applied more than $n-1$ times, so that we finally obtain a sequence converging to a zero of $f(z)$.

7. Conditions (29.11) have to be tried out successively for $m = 2, \dots, n$. Since, however, the complete *J-test* with all values of m is rather expensive in computation time, and on the other hand the case $m > 2$ is a very exceptional one, it is advisable to apply repeatedly only the test (29.11) with $m = 2$ and to use the complete *J-test* perhaps only every tenth time.

It is only worthwhile to begin applying this test if $|f(z_v)|$ is already small. Since in the case of spurious convergence $f(z_v)/f'(z_v) \rightarrow \infty$, in the program the application of the *J-test* ought to be programmed to begin perhaps as soon as $|f(z_v)/f'(z_v)| > 100$.

8. For the programming of the above test it is advisable to treat the case $C_1 = 0$ separately, since in this case the J -test is obviously always positive and the corresponding m is the subscript of the first C_v , $v > 1$, which is $\neq 0$.

On the other hand, the special test with $m = 2$ will be applied particularly often, and it is therefore useful in this case to derive this test with better constants than would follow by specialization of (29.10).

Indeed, with $m = 2$ (29.10) becomes

$$6\sqrt{3}\left|\frac{C_1}{C_2}\right| \leq \rho := \text{Min}\left(\frac{1}{10}, \left|\frac{C_0}{C_2}\right|^{1/2}, \frac{1}{3} \frac{|C_2|}{M_3'}\right). \quad (29.14)$$

Instead, we will prove that already, if the following inequality holds:

$$3\left|\frac{C_1}{C_2}\right| \leq \rho := \text{Min}\left(\frac{1}{10}, \left|\frac{C_0}{C_2}\right|^{1/2}, \frac{|C_2|}{2M_3'}\right), \quad (29.15)$$

the J_2 -routine is successful, with $r = \rho/5$. The condition (29.15) is obviously less restrictive than (29.14).

9. To prove this, assume that (29.15) holds. Then, proceeding as in Sections 2–4, we obtain

$$\begin{aligned} |\varphi'(u) - 2C_2 u| &\leq |C_1| + 3M_3' |u|^2 = |C_2| \rho \left(\left| \frac{C_1}{C_2} \right| \frac{1}{\rho} + 3\rho \frac{M_3'}{|C_2|} \left(\frac{|u|}{\rho} \right)^2 \right) \\ &\leq \left(\frac{1}{3} + \frac{3}{2} \left(\frac{|u|}{\rho} \right)^2 \right) |C_2| \rho \end{aligned}$$

and therefore for $|u| = \rho$ and $|u| = \rho/5$:

$$|\varphi'(u) - 2C_2 u| \leq \begin{cases} (11/12)|2C_2 u|, & |u| = \rho \\ (59/60)|2C_2 u|, & |u| = \rho/5. \end{cases}$$

It follows now, by Rouché's theorem, that $\varphi'(u)$ has in $|u| \leq \rho$ exactly one zero, ζ' , and that even $|\zeta'| < \rho/5$.

10. We therefore obtain, for $\varphi(\zeta')$,

$$\begin{aligned} |\varphi(\zeta')| &\geq |C_0| - |C_1| \frac{\rho}{5} - |C_2| \frac{\rho^2}{25} - M_3' \frac{\rho^3}{125} \\ &= |C_0| - \frac{1}{15} |C_2| \left(3 \left| \frac{C_1}{C_2} \right| \rho + \frac{3}{5} \rho^2 + \frac{3}{25} \frac{M_3'}{|C_2|} \rho^3 \right); \end{aligned}$$

this is, using (29.15),

$$\geq |C_0| - \frac{1}{15} |C_2| \left(\rho^2 + \frac{3}{5} \rho^2 + \frac{3}{25} \rho^2 \frac{1}{2} \right)$$

$$= |C_0| - \frac{83}{750} |C_2| \rho^2$$

$$> |C_0| - \frac{1}{9} |C_2| \rho^2.$$

We obtain

$$|\varphi(\zeta')| > |C_0| - \frac{1}{9} |C_2| \rho^2. \quad (29.16)$$

On the other hand, choosing u_0 with $|u_0| = \rho$ according to (29.9),

$$\begin{aligned} |\varphi(u_0)| &\leq |C_0 + C_2 u_0^2| + |C_1| \rho + M_3' \rho^3 \\ &= (|C_0| - |C_2| \rho^2) + |C_2| \rho^2 \left(\left| \frac{C_1}{C_2} \right| \frac{1}{\rho} + \frac{M_3'}{|C_2|} \rho \right); \end{aligned}$$

by virtue of (29.15) this is

$$\begin{aligned} &\leq |C_0| - |C_2| \rho^2 + \frac{5}{6} |C_2| \rho^2 \\ &= |C_0| - \frac{1}{6} |C_2|^2 \end{aligned}$$

and it follows that

$$|\varphi(u_0)| \leq |C_0| - \frac{1}{6} |C_2| \rho^2 < |\varphi(\zeta')|.$$

This corresponds to (29.6).

30

q-Acceleration. The Practice of the Procedure

THE DEFINITION OF *q*-ACCELERATION

1. The direct application of procedure (28.16) with the modifications indicated in Chapter 29 leads in all circumstances to a sequence converging to a zero of $f(z)$. However, this convergence will usually be very slow. It is therefore important to replace the original procedure (28.16) by a conveniently accelerated one. This can be done in different ways, for instance, by using the Steffensen acceleration procedure discussed in Appendix E. In our case, however, the most convenient method of speeding up the convergence appears to be the following procedure, which we will call *q*-acceleration.

2. Assume a fixed $q > 1$. While in Chapter 28 the next approximation was computed by the rule, in which $f(z)f'(z)$ is assumed to be $\neq 0$,

$$z' := z - t^* \frac{f(z)}{f'(z)}, \quad t^* := \text{Min}(1, T(z)), \quad T(z) := \frac{|f'(z)|^2}{M|f(z)|},$$

we consider instead the sequence of numbers

$$z^{(v)} := z - q^{v-1} t^* \frac{f(z)}{f'(z)} \quad (v = 1, \dots) \quad (30.1)$$

and go in this sequence to the last index $v = k = k(z)$ for which

$$|f(z)| > |f(z')| > |f(z^{(2)})| > \dots > |f(z^{(k)})|, \quad (30.2)$$

so that

$$|f(z^{(k)})| \leq |f(z^{(k+1)})|. \quad (30.3)$$

Such an index $k(z)$ certainly exists if $f(z)f'(z) \neq 0$, since obviously

$$|f(z^{(v)})| \rightarrow \infty \quad (v \rightarrow \infty). \quad (30.4)$$

We can therefore define

$$G_q(z) := z^{(k(z))} = z - q^{k(z)-1} t^* \frac{f(z)}{f'(z)} \quad (f(z)f'(z) \neq 0). \quad (30.5)$$

We then replace the iteration by (28.16) with the iteration

$$z_{v+1} = G_q(z_v) \quad (v = 0, 1, \dots; \quad f(z_v)f'(z_v) \neq 0). \quad (30.6)$$

THE BASIC LEMMA

3. We now have to discuss the convergence of the sequence z_v defined by (30.6). First we prove the following lemma.

Lemma. Denote by d the smallest distance of two different zeros of $f(z)$. Assume

$$0 < \varepsilon < \text{Min}\left(\frac{d}{2q+2}, \frac{d}{2n+2}\right). \quad (30.7)$$

Consider the circles $|z - \zeta_v| = \varepsilon$, none of which contains a zero of $f(z)$ by (30.7), and denote by $M_0 > 0$ the minimum of $|f(z)|$ on the complete set of these circles.

Let now ζ be a fixed one of the zeros of $f(z)$ and assume that z lies in the open ε -neighborhood of ζ , $z \in U_\varepsilon(\zeta)$, and that

$$|f(z)| < M_0.$$

Then we have

$$\left| \frac{f(z)}{f'(z)} \right| < 2\varepsilon \quad (30.8)$$

and further, if we form the sequence (30.1), all numbers

$$z', \quad z^{(2)}, \dots, z^{(k)} = G_q(z) \quad (30.9)$$

lie in $U_\varepsilon(\zeta)$, while for any integer $g > 0$

$$|z^{(k+g)} - \zeta| < (2q^g + 1)\varepsilon. \quad (30.10)$$

4. Proof. We have, if p is the exact multiplicity of the zero ζ of $f(z)$,

$$\frac{f'(z)}{f(z)} = \frac{p}{z - \zeta} + \sum_{v=1}^{n-p} \frac{1}{z - \zeta_v}$$

where ζ_v runs through all zeros of $f(z)$, different from ζ . Therefore, if $|z - \zeta| \leq \varepsilon$,

$$\left| \frac{f'(z)}{f(z)} - \frac{p}{z - \zeta} \right| \leq \frac{n}{d - \varepsilon},$$

$$\frac{f'(z)}{f(z)} = \frac{p}{z - \zeta} + \frac{n\theta^*}{d - \varepsilon}, \quad |\theta^*| \leq 1,$$

$$\frac{f(z)/f'(z)}{z-\zeta} = \frac{1}{p+n\theta^*\varepsilon/(d-\varepsilon)},$$

$$\left| \frac{f(z)/f'(z)}{z-\zeta} \right| \leq \frac{1}{1-n/(2n+1)} < 2,$$

and (30.8) follows.

5. From our definition of M_0 it follows that each time when $|f(z^*)| < M_0$, the point z^* lies within the ε -distance of one of the zeros of $f(z)$, since otherwise $|f(z)|$ would have a minimum $\neq 0$.

If our assertion about the sequence (30.9) is not true, then for a certain first κ , $1 \leq \kappa \leq k$, the expression $z^{(\kappa)}$ lies, by what just has been said, in the ε -neighborhood of a zero $\zeta_1 \neq \zeta$ of $f(z)$. We have therefore

$$\begin{aligned} |z^{(\kappa)} - \zeta_1| &\leq \varepsilon, \\ |z^{(\kappa)} - \zeta| &\geq d - \varepsilon > (2q+1)\varepsilon. \end{aligned} \tag{30.11}$$

On the other hand, if $\kappa > 1$, it follows from (30.1)

$$z^{(\kappa)} - z^{(\kappa-1)} = -(q-1)q^{(\kappa-2)}t^* \frac{f(z)}{f'(z)} = (q-1)(z^{(\kappa-1)} - z).$$

But, by our assumption about κ , $z^{(\kappa-1)}$ lies in $U_\varepsilon(\zeta)$ and therefore

$$\begin{aligned} |z^{(\kappa-1)} - z| &\leq 2\varepsilon, \quad |z^{(\kappa)} - z^{(\kappa-1)}| \leq 2(q-1)\varepsilon, \\ |z^{(\kappa)} - \zeta| &< (2q-2)\varepsilon + \varepsilon = (2q-1)\varepsilon, \end{aligned}$$

in contradiction with (30.11).

6. If on the other hand $\kappa = 1$, it follows from (30.11) that

$$|z' - z| \geq |\zeta_1 - \zeta| - |z - \zeta| - |z' - \zeta_1| \geq d - 2\varepsilon > 2q\varepsilon.$$

Further, by (30.8), since $t^* \leq 1$,

$$|z' - z| = t^* \left| \frac{f(z)}{f'(z)} \right| < 2\varepsilon < 2q\varepsilon.$$

With this contradiction our assertion is also proved for $\kappa = 1$, and therefore the complete assertion of our lemma about the numbers (30.9).

7. As to the assertion (30.10), we have, by definition (30.1),

$$z^{(k+g)} - z = q^g(z^{(k)} - z)$$

and since we already proved that $z^{(k)}$ lies in $U_\varepsilon(\zeta)$, it follows that

$$|z^{(k+g)} - z| \leq 2q^g\varepsilon, \quad |z^{(k+g)} - \zeta| \leq (2q^g+1)\varepsilon.$$

Equation (30.10), and with it our lemma, is proved.

THE CONVERGENCE DISCUSSION

8. Consider now the sequence z_v , defined by (30.6).

If the sequence breaks up, the last z_v is either a zero of $f(z)$, and then our aim is achieved; or $f'(z_v) = 0$, and then the *J*-routine can be applied and the zero z_v of $f'(z)$ is eliminated from the competition.

Otherwise we have

$$|f(z_v)| \downarrow m \geq 0.$$

9. If $m = 0$, $|f(z_v)| \downarrow 0$, then for any ε satisfying the conditions of the above lemma, a certain z_v' lies in the ε -neighborhood of a zero ζ of $f(z)$ and, by virtue of our lemma, all following z_v lie in this neighborhood. Since ε can be taken as small as we will, the convergence of z_v to ζ is proved.

10. Assume now that $m > 0$, $|f(z_{v+1})/f(z_v)| \rightarrow 1$. Since we have

$$|f(z_{v+1})/f(z_v)| \leq |f(z_v')/f(z_v)| < 1,$$

it follows that

$$|f(z_v')/f(z_v)| \rightarrow 1.$$

By (28.12) it follows that $t_v \rightarrow 0$,

$$T(z_v) = t_v \rightarrow 0, \quad |f'(z_v)|^2/f(z_v) \rightarrow 0.$$

Since $f(z_v) > m$, we have

$$f'(z_v) \rightarrow 0.$$

But then from a certain v on, the z_v come so close to one of the zeros of $f'(z)$, different from all zeros of $f(z)$, that the *J*-test is satisfied. Applying the *J*-routine, the corresponding zero of $f'(z)$ is eliminated from the competition, and after at most $n - 1$ applications of the *J*-routine we obtain a sequence tending to a zero of $f(z)$.

SPEED OF CONVERGENCE

11. We will now discuss how good a convergence can be achieved by using the q -acceleration method. We will tackle this problem using asymptotic relations for the error and shall therefore assume from the beginning that the sequence of the z_v is no longer interrupted by the *J*-routines and therefore converges to a zero ζ of $f(z)$ of exact multiplicity $p \geq 1$.

The expression $T(z_v)$ defined by (28.10) satisfies by virtue of (28.5) the

asymptotic relation

$$T(z_v) \sim A |z_v - \zeta|^{p-2}, \quad A = \frac{p |f^{(p)}(\zeta)|}{(p-1)! M} \quad (z_v \rightarrow \zeta). \quad (30.12)$$

In the case where $p = 1$, we have $T(z_v) \rightarrow \infty$, and from a v on $t_v = 1$, $|f(z_{v+1})/f(z_v)| \leq \frac{1}{2}$, by virtue of (28.12), and by (28.2)

$$\overline{\lim} \left| \frac{z_{v+1} - \zeta}{z_v - \zeta} \right| \leq \frac{1}{2}.$$

We have a *linear convergence*. But as a matter of fact, from a certain v on, we switch over to the Newton-Raphson procedure and obtain quadratic convergence.

From now on, therefore, we assume $p \geq 2$.

12. Assume now $p = 2$. From (30.12) and (28.16) we have

$$T(z_v) \rightarrow A, \quad t_v \rightarrow a := \min(1, A).$$

It follows therefore, by (28.12) and (28.2), that

$$\begin{aligned} |f(z_{v+1})/f(z_v)| &\leq 1 - \frac{a}{2}, \\ \overline{\lim} |z_{v+1} - \zeta|/|z_v - \zeta| &\leq \left(1 - \frac{a}{2}\right)^{1/2}. \end{aligned} \quad (30.13)$$

We have therefore *linear convergence* even before, and *a fortiori* after, the q -acceleration.

13. Consider now the case where $p > 2$. From (30.12) and (28.16) it follows that

$$T(z_v) \rightarrow 0, \quad t_v \rightarrow 0.$$

In this case a convergence, although guaranteed by (28.12), will usually be too slow. Applying the q -acceleration, we now have to check $\overline{\lim}_{v \rightarrow \infty} (G_q(z_v) - \zeta)/(z_v - \zeta)$. It can be proved that this expression, for any sequence z_v tending to ζ , is $\leq (q-1)/(q+1)$. We give the proof of this result in Appendix S, as it is too long to be given here.

We have therefore in any case *linear convergence* which is the better the smaller $q-1 > 0$ is.

However, if q is taken too small, the number of terms in (30.2), $k(z_v)$, is too large. Taking $q = 2$, we obtain a reasonably good value $\frac{1}{3}$ of $(q-1)/(q+1)$, even if the number of terms in the sequence (30.2) may be still too large.

14. We can, however, get around this difficulty by also using, simultaneously with the given q ,

$$Q = q^\gamma, \quad \gamma > 1. \quad (30.14)$$

Then if we consider the sequence (30.2) corresponding to Q , the number of terms is roughly divided by γ . After we get the optimal value corresponding to Q ,

$$G_Q(z) = z - Q^{K-1} t^* f(z)/f'(z),$$

we can then try to interpolate between Q^{K-2} and Q^K one of the terms

$$q^\nu Q^{K-2} \quad (0 \leq \nu \leq 2\gamma - 1)$$

and obtain in this way a good approximation to $k_q(z)$. This follows from the following formula, which will also be proved in Appendix S and in which we assume γ as an integer > 1 :

$$\lim_{z \rightarrow \zeta} |k_q(z) - \gamma k_Q(z)| < \gamma. \quad (30.15)$$

FLOW CHARTS

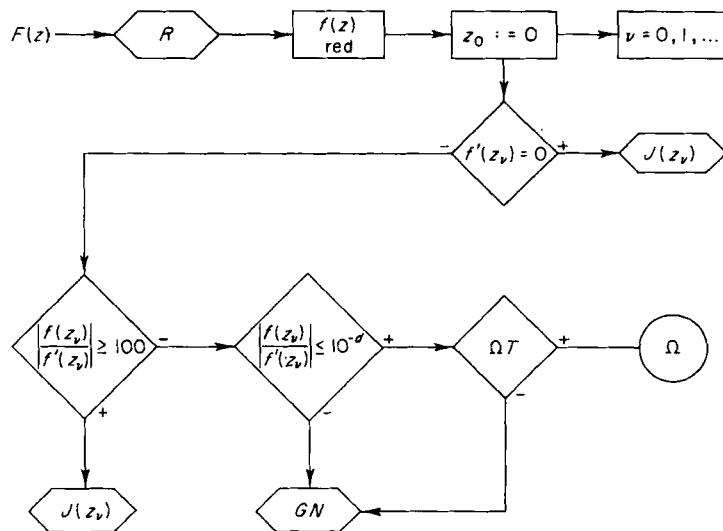
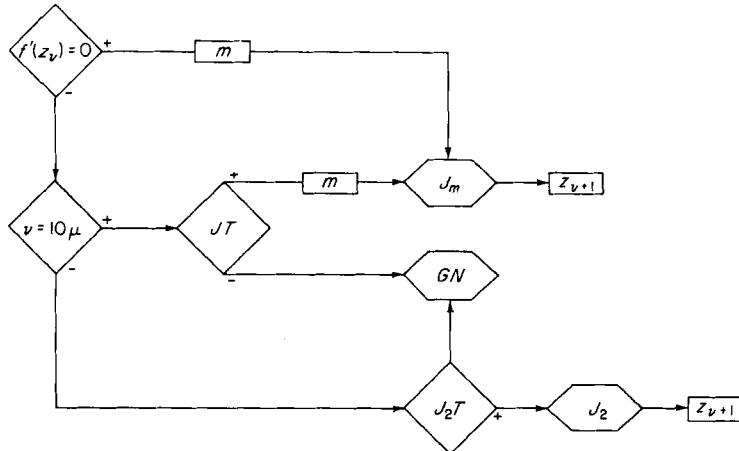
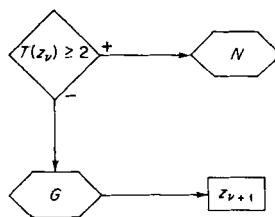
15. The organization of the computation according to the above rules is indicated in the flow charts in Sections 16 and 17. In these flow charts R is the routine giving the reduction of a general polynomial of degree n , $F(z)$, to the reduced polynomial $f(z)$ (see Section 1 in Appendix R). The routine N describes the Newton–Raphson procedure. It is a part of the routine GN , Section 17. The test ΩT is the test given in Section 12 of Chapter 28. This test is applied if $|f(z_\nu)| \leq 10^{-d}$ where the value of d is to be chosen according to the exigencies and possibilities of the computational equipment. Ω in the circle denotes the end of the computation.

The G -routine describes either the computation of $G_q(z)$ or the procedure described in Section 14 of this chapter. Here $q = 2$, $\gamma = 3$, $Q = 8$ will probably be the values usually chosen. The computation of t_ν by (28.16) is included in G .

The routine GN described in the flow chart in Section 17 gives the procedure to be used if no J -routine is applied.

As to the routine $J(z_\nu)$, it is described in more detail in the flow chart in Section 17. Here we apply either the general test JT given by (29.11) for $m = 2, \dots, n$ with the corresponding J_m -routine (29.12) if JT is positive, or the partial test $J_2 T$ corresponding to $m = 2$ according to (29.15). If this is positive, we apply the routine J_2 .

It is not necessary here to give the description of the routine N . Of course, this routine also contains an error estimate following from Theorem 7.2 of Chapter 7.

16.**17. Routine $J(z_\nu)$** **Routine GN** 

31

Normed Linear Spaces

LINEAR SPACES

1. Consider a set S of elements α, β, \dots denoted by Greek minuscules, which forms an Abelian group with the operation $+$ ("addition"). We remind the reader what this means:

For any couple of elements α, β from S there exists a uniquely determined third element of S , denoted by $\alpha + \beta$ ("sum" of α and β). It is further assumed that we have generally

$$\alpha + \beta = \beta + \alpha \quad (\text{commutativity})$$

and for any further element γ of S

$$\gamma + (\alpha + \beta) = (\gamma + \alpha) + \beta \quad (\text{associativity}).$$

Further there exists a uniquely determined "zero element," which we will denote by 0 , such that for any α from S

$$\alpha + 0 = \alpha.$$

Finally, to any α from S there exists a uniquely determined element of S , denoted by $-\alpha$, with the property that

$$\alpha + (-\alpha) = 0.$$

For any couple of elements from S , $\alpha, \beta, \gamma := \beta + (-\alpha)$ (the "difference" of β and α) is usually written as $\beta - \alpha$ and is the uniquely determined element of S such that $\alpha + \gamma = \beta$.

Observe that the properties enumerated above are not all independent. Perhaps the simplest example of such an Abelian group is the set of all real numbers. A little more sophisticated example is given by the set of all n -dimensional vectors c with complex components, \mathbb{C}^n .

2. Consider, together with S as described above, a number field F which is either the field of all real numbers, \mathbb{R} , or the field of all complex numbers, \mathbb{C} . Assume that S is a *modulus* with respect to F . This means that to any couple

of an element ξ of S and an element a of F corresponds a uniquely determined element of S , denoted indifferently by $a\xi$ or ξa (the “product” of a and ξ), such that the following properties are satisfied:

- (1) $a\xi$ is 0 if $a = 0$ or $\xi = 0$;
- (2) $a(b\xi) = (ab)\xi$ ($a \wedge b \in F$);
- (3) $(a+b)\xi = a\xi + b\xi$ ($a \wedge b \in F$);
- (4) $a(\xi + \eta) = a\xi + a\eta$ ($\xi \wedge \eta \in S$).

Then S is called a *linear space with respect to F* , and we will denote F by C_S .

A pretty general example of such a linear space is the set of all n -dimensional vectors with complex components, \mathbb{C}^n , with respect to the coefficient field of all real numbers, \mathbb{R} .

If ξ_1, ξ_2 are two elements of S , we denote as *interval* $\langle \xi_1, \xi_2 \rangle$ (*closed interval*) the set of all elements of S given by

$$\langle \xi_1, \xi_2 \rangle := \{\xi_1 + t(\xi_2 - \xi_1) \mid 0 \leq t \leq 1\}. \quad (31.1)$$

Obviously $\langle \xi_1, \xi_2 \rangle = \langle \xi_2, \xi_1 \rangle$. The *open interval* (ξ_1, ξ_2) is defined similarly.

NORMS

3. Assume now further that in the linear space S with respect to the coefficient field F to any element ξ of S corresponds a uniquely determined nonnegative number, called the *norm of ξ* and denoted by $\|\xi\|$, so that the following properties are satisfied:

- (1) $\|\xi\| = 0$ iff $\xi = 0$;
- (2) $\|a\xi\| = |a| \cdot \|\xi\|$ ($a \in F$);
- (3) $\|\alpha + \beta\| \leq \|\alpha\| + \|\beta\|$ ($\alpha \wedge \beta \in S$) (triangle inequality).

Then the space S is called *normed*. Of course, a given linear space can be normed in different ways and we obtain in this way *different* normed spaces. We use for “normed linear space” the abbreviation *NLS*.

An example of a linear space which can be normed in different ways is given by a generalization of \mathbb{C}^n . We consider the linear space l which is defined as the set of all infinite sequences $\xi := (x_1, \dots, x_v, \dots)$ where the single “components” x_v run through all complex numbers. For any $p \geq 1$ we consider the set of all elements of l with the convergent sum $\sum_{v=1}^{\infty} |x_v|^p$. Using (19.6) it is easy to see that this set is a linear space, which becomes an *NLS*, l_p , if we put

$$\|\xi\| := \|\xi\|_p := \left(\sum_{v=1}^{\infty} |x_v|^p \right)^{1/p}. \quad (31.2)$$

For $p = 2$ we have the so-called Hilbert space H .

It is clear that the linear space corresponding to the normed space l_1 can be also normed by the above formula for any $p > 1$. Further, in the limiting case corresponding to $p = \infty$, we define as l_∞ the subspace of l consisting of all *bounded* sequences ξ , and put

$$\|\xi\| := \|\xi\|_\infty := \sup_{v \geq 1} |\xi_v|. \quad (31.3)$$

Further examples of NLS are considered in Section 6 of this chapter.

By the symbol $U_r(\xi_0)$ we denote the *closed* ball centered in ξ_0 with the radius r , that is, the set of all ξ from S with $\|\xi - \xi_0\| \leq r$. In most cases, this is a canonical form of a closed neighborhood of ξ_0 .

CONVERGENCE

4. If S is an NLS, the definition of convergence in S is immediate. We say that a sequence ξ_v from S converges to an element ξ of S if $\|\xi_v - \xi\|$ tends to 0:

$$\xi_v \rightarrow \xi, \quad \lim_{v \rightarrow \infty} \xi_v = \xi \quad \text{iff} \quad \|\xi_v - \xi\| \rightarrow 0.$$

ξ is then called the *limit* of ξ_v , and is *uniquely determined*. Indeed, if ξ as well as ξ' are both limits of $\{\xi_v\}$, it follows from the triangle inequality that

$$\|\xi - \xi'\| = \|(\xi - \xi_v) - (\xi' - \xi_v)\| \leq \|\xi - \xi_v\| + \|\xi' - \xi_v\| \rightarrow 0,$$

so that we have $\|\xi - \xi'\| = 0$.

It follows further from the triangle inequality that

$$\|\xi\| - \|\xi_v - \xi\| \leq \|\xi_v\| \leq \|\xi\| + \|\xi_v - \xi\|$$

and therefore

$$\|\xi_v\| \rightarrow \|\xi\| \quad (\xi_v \rightarrow \xi). \quad (31.4)$$

Further, again by the triangle inequality,

$$\|\xi_v - \xi_\mu\| = \|(\xi_v - \xi) - (\xi_\mu - \xi)\| \leq \|\xi_v - \xi\| + \|\xi_\mu - \xi\|,$$

and we see that a *necessary* condition for $\xi_v \rightarrow \xi$ is

$$\|\xi_v - \xi_\mu\| \rightarrow 0 \quad (\min(v, \mu) \rightarrow \infty),$$

the *Cauchy–Bolzano condition*. However, if this condition is satisfied—the sequence ξ_v is then called a *Cauchy sequence*—it does not necessarily follow that there exists a ξ in S such that $\xi_v \rightarrow \xi$.

COMPLETENESS AND COMPACTNESS

5. A subset of an NLS in which every Cauchy sequence has a limit is called *complete*. Today a complete NLS is usually called a *Banach space*.

It follows easily from relation (31.4) for a Cauchy sequence ξ_v :

If a subset X_0 of an NLS is complete, then every closed ball $U_r(\xi_0)$ which is $\subset X_0$ is complete.

If ξ_v is a Cauchy sequence in an NLS which is not complete, it still follows that the sequence of the norms converges to a nonnegative limit. Indeed, from the triangle inequality follows at once

$$\|\xi_v\| - \|\xi_\mu\| \leq \|\xi_v - \xi_\mu\|$$

and by symmetry

$$\|\xi_\mu\| - \|\xi_v\| \leq \|\xi_v - \xi_\mu\|.$$

Therefore

$$|\|\xi_v\| - \|\xi_\mu\|| \leq \|\xi_v - \xi_\mu\| \rightarrow 0$$

and we see that the sequence $\|\xi_v\|$ is a Cauchy sequence in \mathbb{R} .

Another important property which can be valid or not in an NLS is the *compactness*. An NLS S is called *compact* iff any infinite set of elements of S contains a Cauchy sequence. It is easy to see that neither the n -dimensional vector space \mathbb{C}^n nor any of the spaces l_p ($p \geq 1$) is compact.

Consider for instance the sequence $\xi_v = (x_1^{(v)}, x_2^{(v)}, \dots)$ in l_p such that $x_v^{(v)} = 1$, $x_v^{(\mu)} = 0$ ($v \neq \mu$); then it is easily seen that we have generally

$$\|\xi_v - \xi_\mu\|_p = 2^{1/p} \quad (v \neq \mu),$$

and therefore this sequence cannot contain any Cauchy sequence. On the other hand, it may be mentioned that any of the spaces l_p ($1 \leq p \leq \infty$) is complete.

EXAMPLES

6. We will now discuss some examples of linear spaces and NLS consisting of complex functions of a real variable. We denote by J the finite closed interval $a \leq x \leq b$.

For any integer $k \geq 1$, the class $C^k(J)$ is defined as the set of all functions which are continuous in J and have there continuous derivatives of the order $1, 2, \dots, k$. Of course, in the points a and b only suitable one-sided derivatives and one-sided continuity have to be considered. The class $C^\infty(J)$ is defined similarly.

The class $C^0(J)$ is then the class of all complex functions continuous in J . A subclass of $C^0(J)$ is the class $L(M, J) = C(0, M, J)$ for an arbitrary positive M . This is the class of functions continuous in J and satisfying there the *Lipschitz inequality* with the *Lipschitz constant* M :

$$|f(x) - f(y)| \leq M|x - y| \quad (x, y \in J). \quad (31.5)$$

More generally, for an integer $k \geq 0$ the class $C(k, M, J)$ is the subclass of $C^k(J)$ consisting of all $f(x) \in C^k(J)$ for which $f^{(k)} \in L(M, J)$. Obviously the classes $C^k(J)$ and $C(k, M, J)$ are linear spaces with respect to $F = \mathbb{C}$. The classes $C(k, M, J)$ are particularly useful in discussions concerning the compactness.

The simplest norm to use in the classes $C^k(J)$ is the norm

$$\|f\|_0^* := \operatorname{Max}_J |f(x)|. \quad (31.6)$$

The convergence in the sense of this norm is obviously *the uniform convergence on J* .

One has often to consider the approximations to a function $f(x)$ of an order $k > 0$, that is to say, such that for $\kappa = 0, 1, \dots, k$ its derivatives $f^{(\kappa)}(x)$ are also uniformly approximated by the corresponding derivatives of the approximating function to $f(x)$. In order to deal with such problems, a more sophisticated norm than (31.6) is defined by

$$\|f\|_\kappa^* := \operatorname{Max}_J \sum_{\lambda=0}^\kappa |f^{(\lambda)}(x)|, \quad (31.7)$$

is useful. This norm can be used in $C^k(J)$ for any $k \geq \kappa$.

SPACES $C^k(J)$

7. It is easy to prove that each of the spaces $C^k(J)$, if normed by $\|f\|_k^*$, is complete. Consider first the case where $k = 0$. Here we make the assumption that

$$\operatorname{Max}_J |f_\mu(x) - f_\nu(x)| \rightarrow 0 \quad (\mu \wedge \nu \rightarrow \infty), \quad (31.8)$$

where all $f_\nu(x)$ are continuous on J , and have to prove that $f_\nu(x)$ tends uniformly to a continuous function over J . By virtue of the Cauchy–Bolzano convergence criterion there exists an $F(x)$ on J such that for each $x \in J$

$$f_\nu(x) \rightarrow F(x) \quad (x \in J).$$

From (31.8) it follows that to any positive ε corresponds an integer $N =$

$N(\varepsilon)$ such that

$$|f_\mu(x) - f_v(x)| \leq \varepsilon \quad (x \in J, \quad \mu \wedge v \geq N(\varepsilon)). \quad (31.9)$$

If we let μ tend to ∞ here, it follows that

$$|F(x) - f_v(x)| \leq \varepsilon \quad (x \in J, \quad v \geq N(\varepsilon)).$$

This signifies that $f_v(x)$ tends to $F(x)$ uniformly on J to $F(x)$. Therefore, by a well-known theorem, $F(x)$ is continuous on J .

8. Assume now $k > 0$. Then if a sequence $f_v(x)$ from $C^k(J)$ is in the sense of the norm $\|f\|_k^*$ a Cauchy sequence, it follows that

$$\|f_v - f_\mu\|_k^* \rightarrow 0 \quad (v \wedge \mu \rightarrow \infty)$$

and therefore by the definition of $\|f\|_k^*$

$$\operatorname{Max}_J |f_\mu^{(k)}(x) - f_v^{(k)}(x)| \rightarrow 0 \quad (\kappa = 0, 1, \dots, k; \quad \mu \wedge v \rightarrow \infty)$$

and therefore, by what has already been proved,

$$f_v^{(\kappa)}(x) \Rightarrow F_\kappa(x) \quad (x \in J, \quad \kappa = 0, 1, \dots, k, \quad v \rightarrow \infty) \quad (31.10)$$

where each of the functions $F_\kappa(x)$ is continuous on J .

Integrating (31.10), we obtain

$$\begin{aligned} \int_a^x f_v^{(\kappa)}(x) dx &\Rightarrow \int_a^x F_\kappa(x) dx = \lim_{v \rightarrow \infty} (f_v^{(\kappa-1)}(x) - f_v^{(\kappa-1)}(a)) \\ &= F_{\kappa-1}(x) - F_{\kappa-1}(a). \end{aligned}$$

It follows therefore that

$$F'_{\kappa-1}(x) = F_\kappa(x) = F_0^{(\kappa)}(x)$$

and now it follows immediately that

$$f_v(x) \rightarrow F_0(x)$$

in the sense of the norm $\|f\|_k^*$.

SPACES $L_\alpha(G)$

9. Further, for the reader familiar with the Lebesgue theory of measure and integration, we ought to mention, again without proofs, the linear spaces $L_\alpha(G)$ ($\alpha \geq 1$) defined as the sets of functions $f(x)$ with the convergent integral $\int_G |f(x)|^\alpha dx$, where G is a measurable set on the x line. The norm can then be

defined by

$$\|f(x)\|_\alpha := \left(\int_G |f(x)|^\alpha dx \right)^{1/\alpha}.$$

In conformity with the above definition, $L_\infty(G)$ is defined as the set of functions of x uniformly bounded, $|f(x)| \leq M_N$, on the set $G - N$ where N is a conveniently chosen zero set, and M_N a constant depending on N . The norm is then given by

$$\|f(x)\|_\infty = \min_{m(N)=0} M_N.$$

32

Metric Spaces

DEFINITION OF METRIC SPACES

1. For the development of some results which we will need in the following, the concept of the NLS is too narrow. We will consider therefore in this chapter a more general class of spaces, *metric spaces*, which in the following will be denoted by MS. A collection of elements X is called a *metric space* if to any couple of its elements ξ, η there corresponds a real number, their “distance”, which will be denoted by $|\xi, \eta|$, satisfying the following postulates:

- (1) $|\xi, \eta| = 0$ iff $\xi = \eta$ ($\xi \wedge \eta \in X$),
- (2) $|\xi, \eta| \leq |\xi, \zeta| + |\eta, \zeta|$ ($\xi \wedge \eta \wedge \zeta \in X$)

(the triangle inequality). From (2) it follows for $\zeta = \xi$: $|\xi, \eta| \leq |\eta, \xi|$ and therefore

$$|\xi, \eta| = |\eta, \xi| \quad (\xi \wedge \eta \in X);$$

further, for $\xi = \eta$: $2|\xi, \zeta| \geq 0$,

$$|\xi, \zeta| \geq 0 \quad (\xi \wedge \zeta \in X).$$

2. We say that the sequence ξ_v *tends to* ξ (has ξ as its *limit*), $\xi_v \rightarrow \xi$, if $|\xi_v, \xi| \rightarrow 0$, where $\xi_v \wedge \xi$ belong to X .

By virtue of this definition, we have, if $\xi_v \rightarrow \xi$, $\xi_v \rightarrow \eta$, by the triangle inequality

$$|\xi, \eta| \leq |\xi_v, \xi| + |\xi_v, \eta| \rightarrow 0,$$

and we see that the limit of a convergent sequence in an MS is uniquely determined.

If we have in X : $\xi_v \rightarrow \xi$, $\eta_\mu \rightarrow \eta$, it follows from the triangle inequality that

$$||\xi_v, \eta_\mu| - |\xi, \eta|| \leq |\xi_v, \xi| + |\eta_\mu, \eta|$$

and therefore

$$|\xi_v, \eta_\mu| \rightarrow |\xi, \eta| \quad (v \rightarrow \infty, \mu \rightarrow \infty).$$

The sequence ξ_v from an MS is called a *Cauchy sequence* if

$$|\xi_v - \xi_\mu| \rightarrow 0 \quad (v \geq \mu \rightarrow \infty).$$

For a convergent sequence ξ_v it follows immediately from the triangle inequality that it is a Cauchy sequence. If in an MS every Cauchy sequence is convergent, this MS is called *complete*. If an MS, U , has a subset U' such that every Cauchy sequence from U' is convergent in U , U' will be called *complete in U*.

An NLS, X , becomes an MS if we define for its elements:

$$|\xi, \eta| := \|\xi - \eta\|. \quad (32.1)$$

Then the concepts of convergence of a Cauchy sequence and of limit, as defined in the NLS, X , remain invariant if X is *metricized* by (32.1).

3. An open *ball* in an MS, X , with the *center* ξ_0 and the *radius* r is defined as the set of all elements $\xi \in X$ for which $|\xi, \xi_0| < r$, and will be denoted by the symbol $(U_r(\xi_0))$.

A set $S \subset X$ is called *open* if to any element $\xi_0 \in S$ there exists an $r > 0$ such that $(U_r(\xi_0))$ belongs to S . It follows from the triangle inequality that each $(U_r(\xi_0))$ is an open set.

4. Consider a mapping $f(X \rightarrow Y)$ of an MS, X , into an MS, Y . We denote generally the element of Y corresponding to $\xi \in X$ by $f(\xi)$. f is called *continuous in X* if to any convergent sequence from X , $\xi_v \rightarrow \xi$, $\xi \in X$, corresponds $f(\xi_v) \rightarrow f(\xi)$.

The above concepts have been defined earlier for an NLS X . They remain invariant if X is metricized by (32.1).

PRINCIPLE OF CONTRACTING OPERATORS

5. Consider an MS, X , and a mapping $T(X \rightarrow X)$. T is called a *contracting operator in X* if there exists a $q < 1$, a *contraction bound* of T , such that

$$|T(\xi), T(\eta)| \leq q |\xi, \eta| \quad (\xi \wedge \eta \in X). \quad (32.2)$$

An important result, due to Banach and usually called the *principle of contracting operator*, states that if X is complete, it follows from (32.2) that there exists in X exactly one point ξ^* for which

$$T(\xi^*) = \xi^*, \quad (32.3)$$

and that further we have for any $\xi_0 \in X$, defining recurrently

$$\xi_v := T(\xi_{v-1}) \quad (v = 1, 2, \dots),$$

$\xi_v \rightarrow \xi^*$. In what follows, we use the following more elaborate statement.

6. Theorem 32.1. Let X be an MS, $X_0 \subset X$, $T(X_0 \rightarrow X)$. Consider a $\xi_0 \in X_0$ such that

$$(a) \quad \xi_1 := T(\xi_0) \in X_0.$$

Assume that

$$(b) \quad 0 < q < 1, \quad \rho := \frac{q}{1-q} |\xi_0, \xi_1|,$$

$$(c) \quad U_\rho(\xi_1) \subset X_0,$$

and assume further that X_0 is complete and that T is a contracting operator in X_0 with the contraction bound q :

$$(d) \quad |T(\xi''), T(\xi')| \leq q |\xi'', \xi'| \quad (\xi'' \wedge \xi' \in X_0).$$

Then, if we define recurrently

$$(e) \quad \xi_{v+1} := T(\xi_v) \quad (v = 1, 2, \dots),$$

all ξ_v lie in X_0 , we have $\xi_v \rightarrow \xi^* \in X_0$, where ξ^* satisfies (32.3) and

$$|\xi^*, \xi_n| \leq \frac{q}{1-q} |\xi_n, \xi_{n-1}| \leq \frac{q^n}{1-q} |\xi_0, \xi_1|, \quad (32.4)$$

while ξ^* is the unique solution in X_0 of Eq. (32.3).

7. Proof. The unicity statement of the above theorem is almost immediate. Indeed, if we had in X_0 two elements ξ^*, η^* with $T(\xi^*) = \xi^*$, $T(\eta^*) = \eta^*$, it would follow from (d) that

$$|\xi^*, \eta^*| = |T(\xi^*), T(\eta^*)| \leq q |\xi^*, \eta^*|$$

and, since $q < 1$, we see that $|\xi^*, \eta^*| = 0$, $\xi^* = \eta^*$.

If ξ_1 in (a) were $= \xi_0$, then we would have $\xi_0 = \xi^*$ and (32.4) is true, as $\xi_n = \xi^*$. We can therefore assume from now on that $\xi_1 \neq \xi_0$, $\rho > 0$.

8. Observe that from (d) it follows that $T(\xi)$ is continuous on X_0 . Assume that for an $n \geq 1$ we already know that

$$|\xi_n, \xi_1| < \rho$$

(this is certainly true for $n = 1$), so that $\xi_n \in X_0$. Then it follows that

$$|\xi_n, \xi_0| < |\xi_0, \xi_1| + \frac{q}{1-q} |\xi_0, \xi_1| = \frac{|\xi_0, \xi_1|}{1-q},$$

and further by (d):

$$|\xi_{n+1}, \xi_1| = |T(\xi_n), T(\xi_0)| \leq q |\xi_n, \xi_0| < \frac{q}{1-q} |\xi_0, \xi_1| = \rho;$$

we see that all ξ_v lie in X_0 .

It now follows, repeatedly applying (d), that for $\mu > \kappa \geq 1$,

$$|\xi_\mu, \xi_{\mu-1}| \leq q |\xi_{\mu-1}, \xi_{\mu-2}| \leq \cdots \leq q^{\mu-\kappa} |\xi_\kappa, \xi_{\kappa-1}| \quad (32.5)$$

and therefore, for $v > n \geq 1$:

$$\begin{aligned} |\xi_v, \xi_n| &\leq \sum_{\mu=n+1}^v |\xi_\mu, \xi_{\mu-1}| \leq |\xi_n, \xi_{n-1}|(q^{v-n} + \cdots + q), \\ |\xi_v, \xi_n| &< \frac{q}{1-q} |\xi_n, \xi_{n-1}| \leq \frac{q^n}{1-q} |\xi_0, \xi_1| \quad (v > n \geq 1). \end{aligned} \quad (32.6)$$

9. Since this $\rightarrow 0$ with $n \rightarrow \infty$, we see that ξ_v tends to a ξ^* lying in the complete space X_0 . We obtain now from (32.6), as $v \rightarrow \infty$, relations (32.4).

On the other hand, it follows from (32.5) that, with $\kappa = 1$, $|T(\xi_{\mu-1}), \xi_{\mu-1}| \leq q^{\mu-1} |\xi_0, \xi_1|$, and with $\mu \rightarrow \infty$, $|T(\xi^*), \xi^*| = 0$, that is (32.3). Theorem 32.1 is proved.

In applying the unicity statement of Theorem 32.1, one will of course try to choose X_0 as large as possible. On the other hand, if we want to localize ξ^* as well as possible, we will choose X_0 as “small” as possible, that is, the set of all ξ with $|\xi, \xi_1| \leq \rho$.

10. We apply the above theorem to prove the following lemma, which will be used later.

Theorem 32.2. Consider a Banach space X and put $X_0 := U_Q(0)$, for a certain positive Q . Consider a mapping $\Omega(X_0 \rightarrow X)$ such that $\Omega(0) = 0$ and generally, for an α , $0 < \alpha < 1$:

$$\|\Omega(\xi'') - \Omega(\xi')\| \leq \alpha \|\xi'' - \xi'\| \quad (\xi' \wedge \xi'' \in X_0), \quad 0 < \alpha < 1. \quad (32.7)$$

Then for every $\eta \in X_0$ with

$$\|\eta\| \leq (1-\alpha) Q \quad (32.8)$$

there exists in X_0 exactly one solution ζ of

$$\zeta - \Omega(\zeta) = \eta. \quad (32.9)$$

Proof. We define the mapping $T(X_0 \rightarrow X)$ by

$$T(\xi) := \eta + \Omega(\xi)$$

and put in Theorem 32.1 $\xi_0 := \eta$, $q := \alpha$. Then, in order to be able to apply this theorem, we have to prove (a), (c), and

$$\rho := \frac{\alpha}{1-\alpha} \|T(\eta) - \eta\| < Q,$$

since relation (d) follows from (32.7). But, by virtue of (32.7),

$$\|\xi_1 - \xi_0\| = \|T(\eta) - \eta\| = \|\Omega(\eta) - \Omega(0)\| \leq \alpha \|\eta\| \leq \alpha(1-\alpha) Q.$$

Therefore

$$\rho \leq \frac{\alpha}{1-\alpha} \alpha(1-\alpha)Q = \alpha^2 Q < Q,$$

$$\rho + \|\xi_1 - \xi_0\| = \|\xi_1 - \xi_0\| \left(\frac{\alpha}{1-\alpha} + 1 \right) = \frac{\|\xi_1 - \xi_0\|}{1-\alpha} \leq \alpha Q < Q$$

we see that conditions (a) and (c) of Theorem 32.1 are satisfied. It follows, therefore, that there exists in X_0 a unique ζ with

$$\zeta = T(\zeta) = \eta + \Omega(\zeta)$$

and this is Eq. (32.9). Our theorem is proved.

33

Operators in Normed Linear Spaces

MAPPINGS AND OPERATORS

1. Consider two NLS X and Y , where it could be also $Y = X$, and a subset, U , of X . If we want to define a “function” f with the “argument” running through U and the “values” from a subset, V , of Y , we have to interpret it as a *mapping* of U into V by which to *any element* $\xi \in U$ corresponds its *image*, an element η of V . Here “into” means that the images of the elements of U do not necessarily cover the whole V —otherwise we say “onto” instead of “into.” In this connection it is usual in functional analysis to speak, not of a *function* in U , but of an *operator* acting on U . Further, instead of $f(\xi) = \eta$ we write $f\xi = \eta$, which then means that η is the image of ξ by the mapping f , that is, that f transforms ξ into η . Instead of the locution *the operator* f *mapping* U *into* V we will simply write $f(U \rightarrow V)$.

The operator which, applied to the elements of X , maps each element into itself is called the identical operator and is usually denoted by I .

If $g(U \rightarrow Y)$ is another operator of the same kind, we define for any a, b from C_Y and for any ξ from U :

$$(af + bg)\xi := a(f\xi) + b(g\xi). \quad (33.1)$$

We have here obviously a new operator mapping U into Y which is denoted by $af + bg$. In the sense of the addition and multiplication operations defined in this way, the set of all operators mapping U into Y is a linear space. The *zero element* of this space is the *zero operator* which maps all elements of U into the zero element of Y .

BOUNDED OPERATORS

2. The operator $f(U \rightarrow Y)$ is called *bounded* on U if there exists a $B \geq 0$ such that

$$\|f\xi\| \leq B \|\xi\| \quad (\xi \in U). \quad (33.2)$$

In (33.2), of course, the left- and right-side norms are taken respectively in Y and X . If f is a bounded operator, denote by B_f^* the Infimum of all B in (33.2). Then (33.2) remains valid if B is replaced by B_f^* , as follows at once by going conveniently to the limit. *The set $S_{U,Y}$, of all bounded operators mapping U into Y is an NLS if we put*

$$\|f\| := B_f^* \quad (f \in S_{U,Y}). \quad (33.3)$$

Indeed, we have, if $a \in C_Y$, $b \in C_Y$, $f \in S_{U,Y}$, $g \in S_{U,Y}$,

$$\begin{aligned} \|(af + bg)\xi\| &\leq \|af\xi\| + \|bg\xi\| \leq |a|B_f^*\|\xi\| + |b|B_g^*\|\xi\| \\ &\leq (|a|B_f^* + |b|B_g^*)\|\xi\|. \end{aligned}$$

We see that $af + bg$ is also a bounded operator and it follows further that

$$B_{af+bg}^* \leq |a|B_f^* + |b|B_g^*. \quad (33.4)$$

For $a = b = 1$ we have the triangle inequality. For $b = 0$, $a \neq 0$ it follows, using that $f = (1/a)(af)$, that

$$B_{af}^* \leq |a|B_f^* \leq |a|\frac{1}{|a|}B_{af}^* = B_{af}^*, \quad B_{af}^* = |a|B_f^*,$$

which gives the homogeneity relation. Further, if $B_f^* = 0$, it follows for all $\xi \in U$: $\|f\xi\| = 0$, $f\xi = 0$, $f = 0$.

The norm of f defined as above is called *induced* by the norms in X and in Y .

The simplest bounded operator mapping X into X is $\eta = a\xi$, realized by multiplying any element ξ of X with a fixed $a \in C_X$.

If $X = \mathbb{R}$, $y = \sin x$ is a bounded operator ($\mathbb{R} \rightarrow \mathbb{R}$), if \mathbb{R} is normed by the moduli; the induced norm of $\sin x$ is $\|\sin x\| = 1$.

On the other hand, the operator ($\mathbb{R} \rightarrow \mathbb{R}$) given by $y = x^3$ is obviously not a bounded operator if \mathbb{R} is normed by the moduli of its elements.

LINEAR OPERATORS

3. An operator ($X \rightarrow Y$) is called *additive* if we have

$$f(\xi_1 + \xi_2) = f\xi_1 + f\xi_2 \quad (\xi_1 \in X, \xi_2 \in X).$$

An operator ($X \rightarrow Y$) which is both bounded and additive is called *linear*. The simplest linear operator ($X \rightarrow X$) is realized by $a\xi$ where the “constant” a belongs to C_X . If $X = \mathbb{R}$, this is also the most general linear operator. This is, however, no longer necessarily true if X is, for instance, \mathbb{C} . In this case we can define a linear operator in the following way: Consider arbitrary nonnegative

ε, δ ; put, for $\xi = a + bi$,

$$f\xi := ae + b\delta = R(\varepsilon - i\delta)\xi.$$

It is easily seen that this formula defines a linear operator. However, this can only be expressed in the form $f\xi \equiv (\alpha + i\beta)\xi$ for fixed real α and β , if we have $\delta = \varepsilon = 0$.

4. Observe finally that if L is a linear operator in X , we have, for $\xi \in X$, $\zeta \in X$,

$$L(\xi + \zeta) - L(\xi) = L(\zeta), \quad \|L(\xi + \zeta) - L(\xi)\| \leq \|L\| \|\zeta\|.$$

We see that such an operator satisfies in the whole space X a “Lipschitz condition” with the “Lipschitz constant,” $\|L\|$. It follows in particular that L is a “continuous function” of its argument, ξ . On the other hand, it follows from the additivity immediately that for any integer p , $L(p\xi) = pL(\xi)$, and therefore, for any rational r ,

$$L(r\xi) = rL(\xi).$$

If now a is an irrational real number, we have for a sequence of rational real numbers r_v , convergent to a , $L(r_v \xi) = r_v L(\xi)$ and therefore in the limit, using the continuity property of L ,

$$L(a\xi) = aL(\xi) \quad (a \geq 0). \quad (33.5)$$

STRONG AND WEAK CONVERGENCE

5. Assume that we have in $S_{U,Y}$

$$f_v \rightarrow f, \quad \|f_v - f\| \rightarrow 0. \quad (33.6)$$

This is called the *strong convergence* in $S_{U,Y}$. From (33.6) it follows for any $\xi \in U$ that $\|f_v \xi - f \xi\| \leq \|f_v - f\| \|\xi\| \rightarrow 0$,

$$f_v \xi \rightarrow f \xi \quad (\xi \in U), \quad (33.7)$$

and even the uniform convergence:

$$f_v \xi \Rightarrow f \xi \quad (\xi \in U, \|\xi\| \leq C). \quad (33.8)$$

On the other hand, if we have, for $f \in S_{U,Y}$, $f_v \in S_{U,Y}$, the relation (33.7), we say that the sequence f_v converges weakly to f . While, as we have seen, the weak convergence follows from the strong convergence, the converse is not necessarily true. However, we can prove the

Lemma. *If the sequence f_v from $S_{U,Y}$ is a Cauchy sequence and is weakly convergent to an $f \in S_{U,Y}$, f_v also strongly converges to f .*

6. Proof. Put $g_v := f_v - f$. Then we have

$$\|g_v - g_\mu\| \rightarrow 0 \quad (\min(v, \mu) \rightarrow \infty), \quad (33.9)$$

$$g_\mu \xi \rightarrow 0 \quad (\xi \in U) \quad (33.10)$$

and have to prove that $g_v \rightarrow 0$, that is, $\|g_v\| \rightarrow 0$.

Since $\lim \|g_v\|$ exists by Section 4 of Chapter 31, we will assume that our assertion is false and therefore that

$$\|g_\mu\| \rightarrow p > 0. \quad (33.11)$$

By the definition of the norm there follows from (33.11) the existence of a sequence of nonzero elements ξ_v from U such that

$$\frac{\|g_v \xi_v\|}{\|\xi_v\|} \rightarrow p \quad (v \rightarrow \infty).$$

Since further by (33.9)

$$\frac{\|(g_\mu - g_v) \xi_v\|}{\|\xi_v\|} \rightarrow 0 \quad (\min(\mu, v) \rightarrow \infty),$$

we have, since also $\|\|g_\mu \xi_v\| - \|g_v \xi_v\|\|/\|\xi_v\| \rightarrow 0$,

$$\frac{\|g_\mu \xi_v\|}{\|\xi_v\|} \rightarrow p \quad (\min(\mu, v) \rightarrow \infty).$$

But then there exists an integer N such that

$$\frac{\|g_\mu \xi_v\|}{\|\xi_v\|} > \frac{p}{2} \quad (\mu \geq N, v \geq N)$$

and therefore in particular taking $v = N$

$$\|g_\mu \xi_N\| > \frac{p}{2} \|\xi_N\| \quad (\mu \geq N)$$

and this contradicts (33.10). Our lemma is proved.

7. As a corollary of the lemma in Section 5 we prove now

Theorem 33.1. *If Y is complete, then $S_{U,Y}$ is also complete.*

Indeed, consider a Cauchy sequence $f_v: \|f_v - f_\mu\| \rightarrow 0$ ($\min(v, \mu) \rightarrow \infty$); then for any fixed ξ

$$\|f_v \xi - f_\mu \xi\| \leq \|f_v - f_\mu\| \|\xi\| \rightarrow 0$$

and therefore $f_v \xi$ is a Cauchy sequence. From the completeness of Y it follows that $f_v \xi$ tends to an η contained in Y . In this way we have a mapping of ξ on η

which gives us an operator $f(U \rightarrow Y)$ with $\eta = f\xi$. Put, using the remark in Section 4 of Chapter 31,

$$\omega := \lim_{v \rightarrow \infty} \|f_v\|;$$

then it follows that

$$\|\eta\| = \lim_{v \rightarrow \infty} \|f_v \xi\| \leq \lim_{v \rightarrow \infty} \|f_v\| \|\xi\| = \omega \|\xi\|,$$

and we see that f is a bounded operator and belongs therefore to $S_{U,Y}$. But then f_v is weakly convergent to f and our assertion follows from the lemma in Section 5.

8. Consider two n -dimensional vector spaces \mathbb{C}^n : $X(\xi := (x_1, \dots, x_n))$, $Y(\eta := (y_1, \dots, y_n))$. If $f(X \rightarrow Y)$ is a linear operator, it can easily be expressed by a *matrix*. Indeed, denote the coordinate unity vectors by $\xi^{(v)} := (\delta_{1v}, \delta_{2v}, \dots, \delta_{nv})$ ($v = 1, \dots, n$), where δ_{kv} is Kronecker's δ . Then $\xi = \sum_{v=1}^n x_v \xi^{(v)}$. Put now $f\xi^{(v)} = (a_{1v}, \dots, a_{nv})$ and consider the matrix $A := (a_{\mu v})$. Then in the relation $\eta = f\xi$ the components of η are given by

$$y_\mu = \sum_{v=1}^n a_{\mu v} x_v \quad (\mu = 1, \dots, n)$$

and this is the matrix formula

$$\eta' = A\xi' \tag{33.12}$$

which describes the most general linear operator mapping X into Y .

34

Inverse Operators

DEFINITION OF THE INVERSE OPERATOR

1. Assume that we have three NLS X, Y, Z , two subsets $U \subset X, V \subset Y$, and two operators $f_1(U \rightarrow V), f_2(V \rightarrow Z)$. For any $\xi \in U$ we have $f_1 \xi \in V, f_2(f_1 \xi) \in Z$. In this way, an operator mapping U into Z is well defined and we will denote it by $f_2 f_1$:

$$(f_2 f_1) \xi := f_2(f_1 \xi).$$

If both operators f_1, f_2 are bounded on U resp. V , we have for any $\xi \in U$,

$$\|f_2 f_1 \xi\| \leq \|f_2\| \|f_1 \xi\| \leq \|f_1\| \|f_2\| \|\xi\|$$

and we see that $f_1 f_2$ is also a bounded operator on U satisfying

$$\|f_2 f_1\| \leq \|f_1\| \|f_2\|. \quad (34.1)$$

If in particular $U = V$ and f maps U into itself, then all natural powers of f, f^v , can be defined and we have obviously from (34.1)

$$\|f^v\| \leq \|f\|^v \quad (v = 2, 3, \dots). \quad (34.2)$$

A product of two linear operators, if it is defined, is linear. Obviously, $If = fI = f$.

2. Assume that $f(U \rightarrow V)$ maps U onto V and gives a *one-to-one mapping* between U and V . This will be denoted by the symbol $f(U \leftrightarrow V)$. This implies a one-to-one mapping of V onto U , which will be denoted by f^{-1} .

f^{-1} is called the *inverse operator* to f . We have obviously

$$ff^{-1} = I, \quad f^{-1}f = I$$

where I in the first formula is the identical operator in V and in the second relation the identical operator in U .

EXISTENCE OF THE INVERSE OPERATOR

3. Theorem 34.1. Let $f(X \rightarrow X)$ be a linear operator mapping the Banach space X into itself, and assume that

$$\omega := \|f\| < 1. \quad (34.3)$$

Then the operator $I-f$ maps X onto itself and has an inverse in X :

$$(I-f)^{-1} = \sum_{v=0}^{\infty} f^v, \quad (34.4)$$

$$(I-f)^{-1} = I + \frac{\omega}{1-\omega} \theta, \quad \|\theta\| \leq 1, \quad (34.5)$$

where θ is an operator from $S_{X,X}$.

4. Proof. If we put

$$s_n := \sum_{v=0}^n f^v,$$

the right-hand expression in (34.4) is defined as $\lim_{n \rightarrow \infty} s_n$. In order to prove that this limit exists, consider for $m > n > 0$ the norm of $s_m - s_n$:

$$\|s_m - s_n\| \leq \sum_{v=n+1}^m \|f\|^v \leq \sum_{v=n+1}^m \omega^v \leq \frac{\omega^{n+1}}{1-\omega},$$

and this tends to 0 with $n \rightarrow \infty$. Therefore $s := \lim_{n \rightarrow \infty} s_n$ exists. On the other hand, from the definition of s_n , we have

$$(I-f)s_n = s_n(I-f) = I - f^{n+1}.$$

Here, however, for $n \rightarrow \infty$, $f^{n+1} \rightarrow 0$ by (34.3) and therefore, indeed,

$$(I-f)s = I, \quad s(I-f) = I. \quad (34.6)$$

Let $V \subset X$ be the image of X by s . From the first relation (34.6) it follows that

$$s(X \leftrightarrow V), \quad (I-f)(V \leftrightarrow X).$$

Let $U \subset X$ be the image of X by $I-f$. From the second relation (34.6) it follows that the mapping is one-to-one. From the three relations

$$(I-f)(V \leftrightarrow X), \quad (I-f)(X \leftrightarrow U), \quad V \subset X,$$

we see that $U = V = X$. This proves (34.4).

Further, it follows from (34.4) that

$$(I-f)^{-1} - I = \sum_{v=1}^{\infty} f^v, \quad \|(I-f)^{-1} - I\| \leq \sum_{v=1}^{\infty} \omega^v = \frac{\omega}{1-\omega},$$

and, putting

$$\theta := \{(I-f)^{-1} - I\} \frac{1-\omega}{\omega},$$

we see that indeed $\|\theta\| \leq 1$.

5. A more sophisticated criterion for the existence of f^{-1} is given by the following theorem.

ANOTHER EXISTENCE THEOREM

Theorem 34.2. Let f, g be two linear operators mapping X into Y where X and Y are NLS. Assume that $f(X \leftrightarrow Y)$; that further one of the following inequalities is satisfied on X , resp. Y ,

$$(a) \|f^{-1}(g-f)\| < 1, \quad (b) \|(g-f)f^{-1}\| < 1; \quad (34.7)$$

and that the corresponding space X or Y is complete. Then $g(X \leftrightarrow Y)$ and we have on Y

$$g^{-1} = f^{-1} + \frac{\omega \|f^{-1}\|}{1-\omega} \theta, \quad \|\theta\| \leq 1, \quad (34.8)$$

where $\theta(Y \rightarrow X)$ is a linear operator and ω one of the left-hand expressions in (34.7), which is < 1 .

Obviously, both inequalities are satisfied if $\|g-f\| \cdot \|f^{-1}\| < 1$.

6. Proof. Define, according as we use (a) or (b) in (34.7), the (linear) operators

$$h := f^{-1}(f-g) = I - f^{-1}g; \quad h' := (f-g)f^{-1} = I - gf^{-1}. \quad (34.9)$$

Then obviously

$$g = f(I-h) = (I-h')f. \quad (34.10)$$

By the corresponding assumption (34.7) and Theorem 34.1 the inverses exist:

$$s := (I-h)^{-1}(X \leftrightarrow X) \quad \text{resp.} \quad s' := (I-h')(Y \leftrightarrow Y). \quad (34.11)$$

Further, the corresponding difference $s - I$ or $s' - I$ is

$$\frac{\omega}{1-\omega} \theta, \quad \|\theta\| < 1. \quad (34.12)$$

We have in the corresponding cases identically

$$g = f(I-h), \quad g = (I-h')f.$$

Here, the factors f , $I - h$ resp. $I - h'$ give one-to-one mappings $X \leftrightarrow Y$, $X \leftrightarrow X$, $Y \leftrightarrow Y$. Therefore, in any case, $g(X \leftrightarrow Y)$ and from the definition of s and s' follows $g^{-1} = sf^{-1}$ resp. $g^{-1} = f^{-1}s'$. In both cases, (34.8) follows immediately from the estimate (34.12).

A BANACH THEOREM

7. We finally mention an important theorem due to Banach:

Theorem 34.3. *Assume that the operator $f(X \leftrightarrow Y)$ is linear on the Banach space X and that Y is also a Banach space. Then $f^{-1}(Y \leftrightarrow X)$ is linear on Y .*

The proof of the *additivity* of f^{-1} is immediate. Put for two arbitrary elements of Y , η_1, η_2 :

$$f^{-1}\eta_1 = \xi_1, \quad f^{-1}\eta_2 = \xi_2, \quad f\xi_1 = \eta_1, \quad f\xi_2 = \eta_2.$$

Since f is additive, it follows further that

$$f(\xi_1 + \xi_2) = \eta_1 + \eta_2, \quad f^{-1}(\eta_1 + \eta_2) = \xi_1 + \xi_2.$$

We see that indeed $f^{-1}(\eta_1 + \eta_2) = f^{-1}\eta_1 + f^{-1}\eta_2$.

The proof of the boundedness of f^{-1} is rather sophisticated. The reader will find this proof in any detailed book on Banach spaces.

It follows from $I = ff^{-1}$ that

$$1 \leq \|f\| \|f^{-1}\|, \quad \|f^{-1}\| \geq 1/\|f\|.$$

8. Let $L(X \leftrightarrow Y)$ be a linear operator mapping X onto Y and Λ its inverse operator. Referring to the definition of the operator norm in Section 2 of Chapter 32, we can write

$$\sup_{\xi \in X} \frac{\|L\xi\|}{\|\xi\|} = \sup_{\|\xi\|=1} \|L\xi\| = \|L\| \quad (34.13)$$

and further, applying the corresponding relations to Λ ,

$$\inf_{\xi \in X} \frac{\|L\xi\|}{\|\xi\|} = \inf_{\|\xi\|=1} \|L\xi\| = \frac{1}{\|\Lambda\|} \quad (34.14)$$

where, if Λ is not bounded, all three terms of this relation have the value 0.

9. We shall need later the following result, which asserts roughly speaking that the inverse of a linear operator continuous with respect to a parameter, is also continuous with respect to this parameter.

Theorem 34.4. *Assume that a linear operator $L_\zeta(U \leftrightarrow V)$ mapping an NLS, U , onto an NLS, V , depends on a parameter ζ running through a neighborhood of ζ_0 in a metric space Z , and possesses the inverse Λ_ζ for all ζ from this neighborhood. If then Λ_ζ is bounded for all these ζ by a constant γ , $\|\Lambda_\zeta\| \leq \gamma$, and L_ζ is continuous in ζ at ζ_0 , then Λ_ζ is continuous in ζ at ζ_0 .*

Proof. Put

$$P_\zeta := (\Lambda_\zeta - \Lambda_{\zeta_0}) L_\zeta = I - \Lambda_{\zeta_0} L_\zeta \rightarrow 0$$

as $\zeta \rightarrow \zeta_0$. But then, from

$$\Lambda_\zeta - \Lambda_{\zeta_0} = P_\zeta \Lambda_\zeta, \quad \|\Lambda_\zeta - \Lambda_{\zeta_0}\| \leq \gamma \|P_\zeta\|$$

the assertion of the theorem follows immediately.

35

Operators Mapping a Linear Interval

A REFINEMENT OF BOREL'S COVERING THEOREM

1. We begin with a lemma containing a refinement of the famous *Borel Covering Theorem*.

Lemma 35.1. *Assume that a function $\delta(t)$ is defined in a finite interval $J: (a \leq t \leq b)$ and is everywhere positive in J . Then there exists a finite set of t_v ,*

$$a \leq t_1 < t_2 < \dots < t_n \leq b \quad (35.1)$$

such that

$$\begin{aligned} t_1 - a &\leq \delta(t_1), & b - t_n &\leq \delta(t_n), \\ t_{v+1} - t_v &\leq \text{Max}\{\delta(t_v), \delta(t_{v+1})\} & (v = 1, 2, \dots, n-1). \end{aligned} \quad (35.2)$$

Proof. Denote generally

$$J(t) := \langle t - \frac{1}{2}\delta(t), t + \frac{1}{2}\delta(t) \rangle \quad (a \leq t \leq b). \quad (35.3)$$

Then by the classical Borel Covering Theorem there exists a finite number of points t_v from J such that the whole interval J is covered by the union of the corresponding $J(t_v)$:

$$J \subset \bigcup_{v=1}^n J(t_v). \quad (35.4)$$

Obviously, we can delete in (35.4) any component $J(t_v)$ if, after its deletion, (35.4) remains true. We can therefore assume from the beginning, without loss of generality, that none of the intervals $J(t_v)$ can be deleted and that the t_v are ordered as in (35.1). We are now going to prove that, for such a sequence of t_v , (35.2) is true. We put generally $J_v := J(t_v)$.

2. If it were not true that $t_1 - a \leq \delta(t_1)$ then a certainly would not be covered by J_1 . Then for a $v > 1$, $a \in J_v$. We would then have the situation as in Fig. 7, where the left halves of the intervals J_1 and J_v are indicated. A glance at Fig. 7 shows that then $J_1 \subset J_v$ and could be deleted from (35.4), contrary to our assumption.



FIGURE 7

If $b - t_n$ were not $\leq \delta(t_n)$, then we would have for a $v < n : b \in J_v$, and the situation would correspond to that in Fig. 8, where the right halves of J_n and J_v are marked, and again obviously J_n would be contained in J_v , contrary to our assumption.

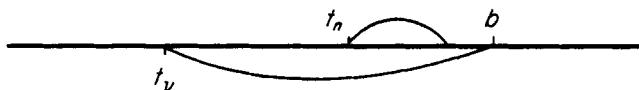


FIGURE 8

3. Assume now that for a certain v , $1 \leq v \leq n-1$, we would have, contrary to (35.2),

$$t_{v+1} - t_v > \text{Max}\{\delta(t_v), \delta(t_{v+1})\} \geq \frac{1}{2}\{\delta(t_v) + \delta(t_{v+1})\};$$

then we would have

$$t_{v+1} - \frac{1}{2}\delta(t_{v+1}) > t_v + \frac{1}{2}\delta(t_v).$$

The situation would then correspond to that in Figs. 9 and 10 where the left

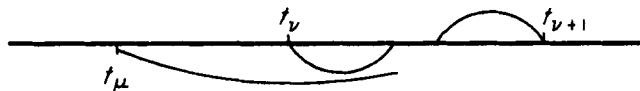


FIGURE 9



FIGURE 10

half of J_{v+1} and the right half of J_v are marked. Then a point from (t_v, t_{v+1}) which is not covered by $J_v \cup J_{v+1}$ is covered by another J_μ , $\mu < v$ or $\mu > v+1$. If $\mu < v$, we have the situation of Fig. 9, where the right half of J_μ is partly marked, and a glance at Fig. 9 shows that $J_v \subset J_\mu$, contrary to our hypothesis. In the case where $\mu > v+1$, the situation is symmetric and Fig. 10 shows that $J_{v+1} \subset J_\mu$, which again contradicts our hypothesis. Lemma 35.1 is proved.

LIPSCHITZ CONDITION FOR $H(t)$

4. In the following sections of this chapter we consider an operator $H(t)$ mapping the finite interval

$$J := (a \leq t \leq b)$$

into an NLS, Y , and we denote the image of t by the symbol $H(t)$. $H(t)$ is then a “function” of the real argument t with “values” from Y .

Theorem 35.1. *Assume that for a constant $M \geq 0$, $H(t)$ satisfies the condition*

$$\limsup_{h \rightarrow 0} \frac{\|H(t+h) - H(t)\|}{|h|} \leq M \quad (t \wedge (t+h) \in J). \quad (35.5)$$

Then $H(t)$ satisfies the “Lipschitz condition” with the “Lipschitz constant” M :

$$\|H(T_2) - H(T_1)\| \leq M(T_2 - T_1) \quad (a \leq T_1 < T_2 \leq b). \quad (35.6)$$

5. Proof. Take an arbitrary $M' > M$. Then by (35.5) to any $t \in J$ corresponds a positive $\delta(t)$ such that

$$\frac{\|H(t') - H(t)\|}{|t' - t|} \leq M' \quad (|t' - t| \leq \delta(t), \quad t \wedge t' \in J). \quad (35.7)$$

Apply Lemma 35.1 to the positive function $\delta(t)$ and the interval J . It follows the existence of t_1, \dots, t_n such that (35.1) and (35.2) are valid. Applying now (35.7) to each of the intervals

$$\langle a, t_1 \rangle, \quad \langle t_1, t_2 \rangle, \dots, \langle t_{n-1}, t_n \rangle, \quad \langle t_n, b \rangle,$$

we have

$$\|H(t_1) - H(a)\| \leq M'(t_1 - a),$$

$$\|H(t_{v+1}) - H(t_v)\| \leq M'(t_{v+1} - t_v) \quad (v = 1, \dots, n-1),$$

$$\|H(b) - H(t_n)\| \leq M'(b - t_n).$$

Then by the triangle inequality it follows that

$$\begin{aligned} \|H(b) - H(a)\| &\leq \|H(t_1) - H(a)\| + \sum_{v=1}^{n-1} \|H(t_{v+1}) - H(t_v)\| + \|H(b) - H(t_n)\| \\ &\leq \left[(t_1 - a) + \sum_{v=1}^{n-1} (t_{v+1} - t_v) + (b - t_n) \right] M' = M'(b - a), \\ \|H(b) - H(a)\| &\leq M'(b - a). \end{aligned}$$

Since this is true for any $M' > M$, we obtain

$$\|H(b) - H(a)\| \leq M(b-a).$$

This is relation (35.6) for $T_1 = a$, $T_2 = b$.

But now we immediately obtain relation (35.6) in the general case, since in our hypothesis we can replace J by the interval $\langle T_1, T_2 \rangle$.

6. Theorem 35.2. *Assume that for a certain nonnegative function $M(t)$, for which the Riemann integral $\int_a^b M(t) dt$ exists, we have*

$$\lim_{h \rightarrow 0} \frac{\|H(t+h) - H(t)\|}{|h|} \leq M(t) \quad (t \wedge (t+h) \in J). \quad (35.8)$$

Then

$$\|H(T_2) - H(T_1)\| \leq \int_{T_1}^{T_2} M(t) dt \quad (a \leq T_1 < T_2 \leq b). \quad (35.9)$$

7. Proof. As in the proof of Theorem 35.1, it suffices to prove (35.9) in the case $T_1 = a$, $T_2 = b$. Since the Riemann integral $\int_a^b M(t) dt$ exists, to an arbitrary positive ε we can find such a subdivision of J ,

$$a = t_0 < t_1 < \cdots < t_n = b, \quad (35.10)$$

that for the corresponding “upper sum,” S , of $M(t)$ for the partition (35.10) we have

$$S := \sum_{v=0}^{n-1} (t_{v+1} - t_v) \sup_{\langle t_v, t_{v+1} \rangle} M(t) < \int_a^b M(t) dt + \varepsilon. \quad (35.11)$$

On the other hand, in the general interval $\langle t_v, t_{v+1} \rangle$, condition (35.5) of Theorem 35.1 is satisfied if we replace M there with

$$\sup_{\langle t_v, t_{v+1} \rangle} M(t).$$

Therefore we have

$$\|H(t_{v+1}) - H(t_v)\| \leq (t_{v+1} - t_v) \sup_{\langle t_v, t_{v+1} \rangle} M(t).$$

Summing this over $v = 0, 1, \dots, n-1$, we obtain, using (35.11),

$$\sum_{v=0}^{n-1} \|H(t_{v+1}) - H(t_v)\| \leq S < \int_a^b M(t) dt + \varepsilon.$$

Therefore

$$\|H(b) - H(a)\| \leq \sum_{v=0}^{n-1} \|H(t_{v+1}) - H(t_v)\| < \int_a^b M(t) dt + \varepsilon.$$

In the inequality

$$\|H(b) - H(a)\| < \int_a^b M(t) dt + \varepsilon,$$

which we thus obtained, we can now let ε diminish to 0 and obtain (35.9) for $T_1 = a$, $T_2 = b$. Theorem 35.2 is proved.

TAYLOR DEVELOPMENT

8. The derivative of $H(t)$, $H'(t)$, is defined by the usual formula:

$$\frac{d}{dt} H(t) = H'(t) := \lim_{h \rightarrow 0} \frac{H(t+h) - H(t)}{h} \quad (t \wedge (t+h) \in J).$$

The higher derivatives $H^{(v)}(t)$ are defined recurrently in a similar way.

Theorem 35.3. Assume that in J the n derivatives $H'(t), \dots, H^{(n)}(t)$ exist everywhere and that further the n th derivative, $H^{(n)}(t)$, satisfies the relation

$$\varlimsup_{h \rightarrow 0} \frac{\|H^{(n)}(t+h) - H^{(n)}(t)\|}{|h|} \leq M(t) \quad (t \wedge (t+h) \in J), \quad (35.12)$$

where the Riemann integral $\int_a^b M(t) dt$ exists. Form for $t \in J$ the functions

$$M_v(t) = \frac{1}{v!} \int_a^t (t-\tau)^v M(\tau) d\tau \quad (v = 0, 1, \dots, n). \quad (35.13)$$

Then

$$H(t) = \sum_{v=0}^n \frac{(t-a)^v}{v!} H^{(v)}(a) + \theta^*(t) M_n(t), \quad \|\theta^*(t)\| \leq 1, \quad (35.14)$$

where $\theta^*(t)$ maps J into Y .

9. Proof. We prove first for the functions (35.13) the relations

$$\frac{d}{dt} M_v(t) = M_{v-1}(t) \quad (v = 1, \dots, n). \quad (35.15)$$

Indeed, from (35.13) it follows, for $t \wedge (t+h) \in J$, $v \geq 1$, that

$$\begin{aligned} v!(M_v(t+h) - M_v(t)) &= \left(\int_a^{t+h} - \int_a^t \right) (t+h-\tau)^v M(\tau) d\tau \\ &\quad + \int_a^t [(t+h-\tau)^v - (t-\tau)^v] M(\tau) d\tau = I + II, \end{aligned}$$

where, if $|M(t)| \leq C$ ($t \in J$) (the existence of the bound C follows from the Riemann integrability),

$$\begin{aligned} I &= \int_t^{t+h} (t+h-\tau)^v M(\tau) d\tau = \int_0^h (h-x)^v M(t+x) dx = \theta C |h|^{v+1}, \\ II &= \int_a^t [vh(t-\tau)^{v-1} + \binom{v}{2} h^2 (t-\tau + \theta_1 h)^{v-2}] M(\tau) d\tau \\ &= v! h M_{v-1}(t) + h^2 \theta_2 * \binom{v}{2} (|t| + |a| + |h|)^{v-2} C, \end{aligned}$$

so that now it follows that

$$\frac{1}{h} (M_v(t+h) - M_v(t)) \rightarrow M_{v-1}(t).$$

10. Observe now that by Theorem 35.2 the relation (35.12) has in particular as a consequence

$$\|H^{(n)}(t) - H^{(n)}(a)\| \leq M_0(t) \quad (t \in J). \quad (35.16)$$

Form the expression

$$G(t) := H(t) - \sum_{v=0}^n \frac{(t-a)^v}{v!} H^{(v)}(a).$$

Then we verify immediately that

$$G^{(v)}(a) = 0 \quad (v = 0, 1, \dots, n)$$

and that further

$$\|G^{(n)}(t)\| \leq M_0(t), \quad (35.17)$$

and we have to prove

$$G(t) = \theta^*(t) M_n(t), \quad \|\theta^*(t)\| \leq 1 \quad (t \in J). \quad (35.18)$$

11. Since

$$\lim_{h \rightarrow 0} \frac{G^{(n-1)}(t+h) - G^{(n-1)}(t)}{h} = G^{(n)}(t),$$

it follows from (35.17) that

$$\lim_{h \rightarrow 0} \frac{\|G^{(n-1)}(t+h) - G^{(n-1)}(t)\|}{|h|} \leq M_0(t).$$

(Here and in the following formulas we assume, without mentioning it every time, that t , as well as $t+h$, lies in J .) But then, applying Theorem 35.2 to

$G^{(n-1)}(t)$, we obtain

$$\|G^{(n-1)}(t)\| \leq \int_a^t M_0(\tau) d\tau = M_1(t),$$

and proceeding in the same way, finally

$$\|G(t)\| \leq \int_a^t M_{n-1}(\tau) d\tau = M_n(t).$$

Putting

$$\theta^*(t) = \frac{G(t)}{M_n(t)}$$

it follows that $\|\theta^*(t)\| \leq 1$ and this proves (35.18).

12. If we specialize Theorem 35.3 for $n = 1$, we can write the result in the form

$$\|H(t) - H(a) - (t-a) H'(a)\| \leq M_1(t) := \int_a^t (t-\tau) M(\tau) d\tau \quad (35.19)$$

where $M(\tau)$ is assumed to be Riemann integrable in J and is given by

$$M(\tau) := \overline{\lim}_{h \rightarrow 0} \frac{\|H'(t+h) - H'(t)\|}{|h|} \quad (t \wedge (t+h) \in J). \quad (35.20)$$

Of course, in the case of formulas (35.12) and (35.20), if the corresponding derivatives exist and are Riemann integrable, $M(t)$ can be replaced with the modulus of the corresponding derivatives.

36

The Directional Derivatives and Gradients of Operators

DIRECTIONAL DERIVATIVES

1. Consider, for the whole of Chapter 36 the operator $f(X \rightarrow Y)$ where X and Y are NLS, and let ξ and $\Delta \neq 0$ be two elements of X .

If, for real t , $(d/dt)f(\xi + t\Delta)$ at $t = 0$ exists, it is called the *G-differential* or *Gateau differential* of f at ξ in the direction Δ and will be denoted by

$$df(\xi; \Delta) := \left. \frac{d}{dt} f(\xi + t\Delta) \right|_{t=0}. \quad (36.1)$$

From (36.1) follows at once

$$df(\xi + t\Delta; \Delta) = \left. \frac{d}{dt} f(\xi + t\Delta) \right|_{t=0}, \quad (36.2)$$

for each (real) t for which the right-hand derivative exists.

From our definition it follows that if $df(\xi; \Delta)$ exists, then for any real $c \neq 0$, $df(\xi; c\Delta)$ exists too, and

$$df(\xi; c\Delta) = c df(\xi; \Delta) \quad (c \neq 0). \quad (36.3)$$

2. Theorem 36.1. Assume that for $a \leq t \leq b$ the G-differential of f at $\xi + t\Delta$ in the direction Δ exists and that for a nonnegative $N(t)$ which is Riemann integrable in $\langle a, b \rangle$ we have

$$\|df(\xi + t\Delta; \Delta)\| \leq N(t)\|\Delta\| \quad (a \leq t \leq b). \quad (36.4)$$

Then

$$\|f(\xi + t_2\Delta) - f(\xi + t_1\Delta)\| \leq \|\Delta\| \int_{t_1}^{t_2} N(t) dt \quad (a \leq t_1 < t_2 \leq b). \quad (36.5)$$

Indeed, putting $H(t) := f(\xi + t\Delta)$, (36.4) becomes

$$\lim_{|h| \rightarrow 0} \frac{\|H(t+h) - H(t)\|}{|h|} = \|df(\xi + t\Delta; \Delta)\| \leq N(t)\|\Delta\|.$$

But then condition (35.8) of Theorem 35.2 with $M(t) := N(t)\|\Delta\|$ holds and follows immediately from relation (35.9) in this theorem.

3. Generalizing (36.1), we put

$$d^v f(\xi; \Delta) := \left. \frac{d^v}{dt^v} f(\xi + t\Delta) \right|_{t=0}, \quad (36.6)$$

assuming that the v th derivative exists.

Theorem 36.2. Take a t with $a \leq t \leq b$ and assume that $d^v f(\xi + \tau\Delta; \Delta)$ exists for $v = 1, \dots, n$ and $0 \leq (\tau - a)/(t - a) \leq 1$. Assume further that

$$\overline{\lim}_{|h| \rightarrow 0} \frac{\|d^n f(\xi + (\tau + h)\Delta; \Delta) - d^n f(\xi + \tau\Delta; \Delta)\|}{|h|^n} \leq N(t) \|\Delta\|^{n+1} \quad \left(0 \leq \frac{\tau - a}{t - a} \leq 1\right). \quad (36.7)$$

Then

$$f(\xi + t\Delta) - f(\xi + a\Delta) = \sum_{v=1}^n (t - a)^v d^v f(\xi; \Delta) + \theta^* \frac{\|\Delta\|^{n+1}}{n!} \int_a^t (t - \tau)^n N(\tau) d\tau, \quad (36.8)$$

$$\|\theta^*\| \leq 1, \quad \theta^* \in Y.$$

Indeed, (36.8) follows from Theorem 35.3, putting again $H(\tau) := f(\xi + \tau\Delta)$. Then formula (35.14) of this theorem yields (36.8) if we take $M(\tau) := N(\tau) \|\Delta\|^{n+1}$.

If $N(\tau)$ in (36.7) is constant, $= M$, the remainder in formula (36.8) becomes

$$\theta^* M \frac{(t - a)^{n+1}}{(n+1)!} \|\Delta\|^{n+1}.$$

GATEAU GRADIENT

4. Assume now that the operator f is G -differentiable at the point ξ in any direction $\Delta \neq 0$. Then the expression $df(\xi; \Delta)$ can be considered as the result of an operator acting on any $\Delta \neq 0$ from X . This operator is the *gradient of f at the point ξ* and denoted by $\text{grad } f(\xi)$. If we postulate that this operator maps the zero element of X into the zero element of Y , we can write

$$\text{grad } f(\xi) \Delta := df(\xi; \Delta) \quad (\Delta \neq 0), \quad \text{grad } f(\xi) 0 = 0. \quad (36.9)$$

Here, the point ξ is to be considered as a “parameter.”

We will then say that f has at ξ a *G-gradient* or *Gateau gradient* or also that f is *weakly differentiable* at ξ .

5. Using formula (36.3), it follows that $\text{grad } f$ is *homogeneous* with respect to real multipliers:

$$\text{grad } f(\xi)(c\Delta) = c \text{ grad } f(\xi) \Delta \quad (c \geq 0). \quad (36.10)$$

But a G -gradient is not necessarily a linear operator in the sense of Section 3 of Chapter 33.

If, however, $\text{grad}f(\xi)$ is a *linear operator*, we speak then of an *L -gradient* and say that the operator f is *L -differentiable at ξ* .

As to the *boundedness condition* implied in the linearity of $\text{grad}f(\xi)$, it is clear that the inequality $\|\text{grad}f(\xi)\| \leq N$ is equivalent with

$$\lim_{|t| \rightarrow 0} \frac{\|f(\xi + t\Delta) - f(\xi)\|}{|t|} \leq N \|\Delta\| \quad (\Delta \in X).$$

This condition is certainly satisfied if we have the “Lipschitz condition at ξ with the Lipschitz constant N ”:

$$\|f(\xi + \Delta) - f(\xi)\| \leq N \|\Delta\| \quad (\Delta \in X). \quad (36.11)$$

6. Usually, $\text{grad}f(\xi)$ is an *additive operator*,

$$\text{grad}f(\xi)(\Delta_1 + \Delta_2) = \text{grad}f(\xi)\Delta_1 + \text{grad}f(\xi)\Delta_2 \quad (36.12)$$

for all $\Delta_1 \wedge \Delta_2 \in X$.

A necessary and sufficient condition for the additivity of the gradient of an operator f can be obtained conveniently by adapting a discussion due to Vainberg.

Theorem 36.3. Consider the operator $f(X \rightarrow Y)$ where $X \wedge Y$ are NLS. Assume that the G -gradient $\text{grad}f(\xi)$, $\xi \in X$, exists. Then the operator $\text{grad}f(\xi)$ is additive iff for real t :

$$\frac{f(\xi + t(\Delta_1 + \Delta_2)) + f(\xi) - f(\xi + t\Delta_1) - f(\xi + t\Delta_2)}{t} \rightarrow 0 \quad (t \rightarrow 0, \quad \Delta_1 \wedge \Delta_2 \in X). \quad (36.13)$$

This follows immediately from the identity

$$\begin{aligned} & \frac{f(\xi + t(\Delta_1 + \Delta_2)) - f(\xi)}{t} - \frac{f(\xi + t\Delta_1) - f(\xi)}{t} - \frac{f(\xi + t\Delta_2) - f(\xi)}{t} \\ &= \frac{f(\xi + t(\Delta_1 + \Delta_2)) - f(\xi + t\Delta_1)}{t} - \frac{f(\xi + t\Delta_2) - f(\xi)}{t}. \end{aligned}$$

7. The classical case of the L -gradient is the gradient of a function $f(P) = f(x_1, x_2, \dots, x_n)$ of a point P in the n -dimensional space with the coordinates x_1, \dots, x_n . In this case we obtain at once

$$\left. \frac{d}{dt} f(P + t\Delta) \right|_{t=0} = \sum_{v=1}^n d_v \frac{\partial f(P)}{\partial x_v}, \quad \Delta := (d_1, d_2, \dots, d_n).$$

The gradient is then given by the vector

$$L\text{-grad } f(P) = (f'_{x_1}(P), \dots, f'_{x_n}(P))$$

which operates on the vector Δ by forming the inner product with Δ , and maps in this way Δ into the real number $(\Delta, \text{grad } f(P))$.

A more sophisticated example is that of an m -dimensional vector function of an n -dimensional point P :

$$\Phi(P) = (f_1(P), \dots, f_m(P)).$$

Here, we obtain at once

$$d\Phi(P; \Delta) = \left[\frac{d}{dt} f_\mu(P + t\Delta) \right]_{t=0} = \left[\sum_{v=1}^n \frac{\partial f_\mu(P)}{\partial x_v} d_v \right]_{\mu=1, \dots, m}.$$

We can therefore write

$$L\text{-grad } \Phi(P) \Delta = d\Phi(P; \Delta) = \left(\frac{\partial f_\mu}{\partial x_v}(P) \right) \Delta.$$

The gradient in this case can be interpreted as the Jacobian matrix of the components of Φ with respect to the components of P .

F-DIFFERENTIALS AND F-GRADIENTS

8. If $\text{grad } f(\xi)$ is a linear operator, which may be denoted for the moment by L , we have

$$f(\xi + t\Delta) - f(\xi) - t(L\Delta) = o(t) \quad (36.14)$$

where the quotient of the right-hand expression by t tends to 0 but *not necessarily uniformly with respect to Δ* . In many cases the right-hand expression in (36.14) can be proved as being $o(t\|\Delta\|)$ with $t\|\Delta\| \rightarrow 0$. In this case the operator L will be called the *strong gradient* or *F-gradient* or *Fréchet gradient*. It is also very often denoted as the *Fréchet derivative*. In some cases, however, the verification of the additional condition contained in the definition of the *F-gradient* presents difficulties, while in some applications, the existence of the *G-* or *L-gradient* is already sufficient.

9. However, the following sufficient condition for the existence of the *F-gradient* can often be conveniently used.

Theorem 36.4 Assume that $L(\xi) := L\text{-grad } f(\xi)$ for $\xi \in U, (\xi_0)$ is continuous in ξ at ξ_0 . Then $L(\xi)$ is the *F-gradient* of f at ξ_0 .

Proof. Put $\Delta := \xi - \xi_0$ and, for $0 \leq t \leq 1$,

$$H(t) := f(\xi_0 + t\Delta) - f(\xi_0) - tL(\xi_0)\Delta.$$

Then it follows from (36.2) and (36.9) that

$$H'(t) = [L(\xi_0 + t\Delta) - L(\xi_0)] \Delta.$$

On the other hand, by the continuity assumption, to any ε with $0 < \varepsilon \leq r$ there corresponds a $\delta > 0$ such that from $\|\xi - \xi_0\| \leq \delta$ it follows that $\|L(\xi) - L(\xi_0)\| \leq \varepsilon$. We have then, for all t with $0 \leq t \leq 1$, $\|H'(t)\| \leq \|\Delta\| \varepsilon$, and from Theorem 35.1 it now follows that $H(1) - H(0) = H(1) = o(1)$ ($\|\Delta\| \rightarrow 0$), which is the assertion of the theorem.

10. If L is a linear operator in an NLS, X , we have for arbitrary ξ and Δ from X and an arbitrary real t : $L(\xi + t\Delta) = L(\xi) + tL(\Delta)$. It follows, differentiating with respect to t , by virtue of (36.1), that

$$dL(\xi + t\Delta; \Delta) = L(\Delta), \quad (36.15)$$

$$F\text{-grad } L(\xi) = L. \quad (36.16)$$

We see that the F -gradient of a linear operator in X is, at any element ξ of X , again the same operator, so that in particular a linear operator is everywhere its own F -gradient.

11. We are now going to prove a theorem about F -gradients.

Theorem 36.5. Assume that $f(\xi)$ has at ξ_0 the F -gradient L . Then

$$\overline{\lim}_{\xi \rightarrow \xi_0} \frac{\|f(\xi) - f(\xi_0)\|}{\|\xi - \xi_0\|} = \|L\|. \quad (36.17)$$

If, further, $\Lambda := L^{-1}$ exists in a $U_r(\xi_0)$, then

$$\lim_{\xi \rightarrow \xi_0} \frac{\|f(\xi) - f(\xi_0)\|}{\|\xi - \xi_0\|} = \frac{1}{\|\Lambda\|}, \quad (36.18)$$

where, if Λ is not bounded, $1/\|\Lambda\|$ is to be replaced with 0.

12. Proof. We have by definition of L

$$f(\xi) - f(\xi_0) = L(\xi - \xi_0) + \|\xi - \xi_0\| o(1). \quad (36.19)$$

Put $\xi - \xi_0 = u \|\xi - \xi_0\|$ where u is arbitrary with $\|u\| = 1$. If on both sides of (36.19) we take the norm and divide by $\|\xi - \xi_0\|$,

$$\frac{\|f(\xi) - f(\xi_0)\|}{\|\xi - \xi_0\|} = \|Lu\| + o(1) \quad (\xi \rightarrow \xi_0).$$

Then it is clear that $\overline{\lim}$ of the left-hand expression, as $\xi \rightarrow \xi_0$, is $\leq \|L\|$. On the other hand, we can let u run through a sequence of elements of X for which $\|Lu\|$ tends to $\|L\|$ and (36.17) follows. Further, if Λ exists, we can use (34.14) and in the same way obtain (36.18).

37

Central Existence Theorem

FORMULATION OF THE CENTRAL EXISTENCE THEOREM

1. We are now going to derive, for the general Banach spaces, an analog of Theorems 2.2 and 2.3.

If in the formulation of Theorem 2.2 we replace η with r, m with $1/\gamma$, x_0 with ξ_0 , and x with ξ , and drop the unicity part of the theorem, we obtain the following result.

If for an $r > 0$, $f(\xi) \wedge f'(\xi)$ exist in the interval $J_{\xi_0}: \langle \xi_0 - r \leq \xi \leq \xi_0 + r \rangle$ and if in this interval $1/|f'(\xi)| \leq \gamma \leq r/|f(\xi_0)|$, where $f(\xi_0) \neq 0$, then there exists in J_{ξ_0} a ζ such that $f(\zeta) = 0$.

This statement can now be almost literally generalized to the case of the general Banach spaces.

2. Theorem 37.1. (Central Existence Theorem) Consider a mapping $\Omega(X \rightarrow Y)$ where X and Y are NLS, a $\xi_0 \in X$ and assume that $\Omega(\xi_0) =: \Omega_0 \neq 0$. For a positive γ assume that the closed ball $K: \|\xi - \xi_0\| \leq \gamma \|\Omega_0\|$ is complete. Assume further that Ω has in K the L-gradient $L(\xi)$ which has an inverse $\Lambda(\xi)$ such that $\|\Lambda(\xi)\| \leq \gamma$ ($\xi \in K$), and that $L(\xi)$ is continuous with respect to the parameter ξ for $\xi \in K$. Then there exists a ζ in K such that $\Omega(\zeta) = 0$.

This theorem is equivalent with the following theorem, which is an analog to Theorem 2.3°.

Theorem 37.1°. Consider a mapping $f(X \rightarrow Y)$, $X \wedge Y$ being NLS. Assume that $f(0) = 0$ and that for an $r > 0$ the L-gradient of f exists with its inverse $f^*(\xi)$ and is continuous with respect to the parameter ξ in the ball $\|\xi\| \leq r$; assume further that the ball $\|\xi\| \leq r$ is complete and that in this ball $\|f^*\| \leq \gamma$, $\gamma < \infty$. Then for all $\eta \in Y$ with $\|\eta\| \leq r/\gamma$ the equation $f(\xi) = \eta$ can be solved with a ξ in X satisfying $\|\xi\| \leq r$.

The equivalence of this statement with the above theorem is easily established by using the transformation

$$f(\xi) := \Omega(\xi_0 + \xi) - \Omega_0, \quad \eta := \Omega_0$$

in order to derive Theorem 37.1 from Theorem 37.1°, and the transformation

$$\Omega(\xi) := f(\xi) - \eta, \quad \xi_0 := 0$$

in order to derive Theorem 37.1° from Theorem 37.1.

A LOCAL EXISTENCE THEOREM

3. Before proving Theorem 37.1 we have to prove a similar statement of more “local character.”

Theorem 37.2. (Local Existence Theorem) Consider a mapping $f(X \rightarrow Y)$, where X and Y are NLS, and positive R, γ, r . Assume that the ball $K: \|\xi - \xi_0\| \leq R$ is complete in X . For a ξ_1 with $\|\xi_0 - \xi_1\| < R$ put $f(\xi_1) =: \eta_1$. Assume that $f(\xi)$ has in K an L-gradient $L(\xi)$ with an inverse $\Lambda(\xi)$ such that throughout $K: \|\Lambda(\xi)\| \leq \gamma$ and that $L(\xi)$ is continuous with respect to the parameter ξ at $\xi = \xi_1$.

Assume finally that in an interval $J_r: |t - t_1| \leq r$ there exists an $\eta(t) \in Y$ such that $\eta(t_1) = \eta_1$ and $\eta'(t)$ exists in J_r and is continuous at t_1 .

Then, for any sufficiently small positive Q there exists a positive $\rho \leq R$ with the following property.

Denote with J_ρ the interval $|t - t_1| \leq \rho$. For any $t \in J_\rho$ there exists a $\xi(t) \in U_Q(\xi_1)$ such that

$$f(\xi(t)) = \eta(t) \quad (t \in J_\rho), \quad \xi(t_1) = \xi_1; \quad (37.1)$$

$\xi(t)$ is the unique solution from $U_Q(\xi_1)$ of the equation $f(\xi) = \eta(t)$ and is continuous at $t = t_1$; $\xi'(t_1)$ exists and

$$\xi'(t_1) = \Lambda(\xi_1)\eta'(t_1), \quad \|\xi'(t_1)\| \leq \gamma \|\eta'(t_1)\|. \quad (37.2)$$

4. Proof. Put

$$\xi - \xi_1 =: x, \quad \eta - \eta_1 =: y, \quad \varphi(x) := f(\xi_1 + x) - \eta_1, \quad y(t) := \eta(t) - \eta_1, \quad (37.3)$$

$$L(\xi_1) =: L_1, \quad \Lambda(\xi_1) =: \Lambda_1. \quad (37.4)$$

If we define the operator Ω acting on x by

$$\Omega(x) := \Lambda_1(L_1x - \varphi(x)), \quad (37.5)$$

it follows that

$$\Omega(x) = x - \Lambda_1\varphi(x), \quad \varphi(x) = L_1(x - \Omega(x)). \quad (37.6)$$

From (37.3) and (37.6) it follows immediately that the equation $f(\xi) = \eta$ is equivalent to the equation

$$x - \Omega(x) = \Lambda_1y. \quad (37.7)$$

5. From the assumed continuity of $L(\xi)$ at ξ_1 follows the existence of a positive Q such that we have for $\|x\| \leq Q$:

$$\|L(\xi_1 + x) - L_1\| \leq \frac{1}{2\gamma} \quad (\|x\| \leq Q), \quad (37.8)$$

and this Q can be chosen $< R - \|\xi_1 - \xi_0\|$. If we denote therefore by X_0 the ball $\|\xi - \xi_1\| \equiv \|x\| \leq Q$, this ball is contained in K .

6. We want now to obtain an estimate for $\text{grad } \Omega(x)$ in X_0 . We have by (37.5) and (37.3)

$$\text{grad } \Omega = \Lambda_1(L_1 - L(\xi_0 + x))$$

and therefore, by virtue of (37.8),

$$\|\text{grad } \Omega\| \leq \gamma \frac{1}{2\gamma} = \frac{1}{2} \quad (\xi \in X_0).$$

It follows, therefore, using Theorem 35.1, that

$$\|\Omega(x'') - \Omega(x')\| \leq \frac{1}{2} \|x'' - x'\| \quad (x' \wedge x'' \in X_0). \quad (37.9)$$

But then our $\Omega(x)$ satisfies the conditions of Theorem 32.2 with $\alpha = \frac{1}{2}$ and Eq. (37.7) is uniquely solvable with a $\xi \in X_0$ as soon as

$$\|\Lambda_1 y\| \leq Q/2. \quad (37.10)$$

7. Since by (37.3) $y = \eta(t) - \eta_1$, we have

$$\|\Lambda_1 y\| \leq \gamma \|y(t) - y(t_1)\|. \quad (37.11)$$

Put now $\gamma_1 := \|\eta'(t_1)\|$. Since $\eta'(t)$ is assumed continuous at $t = t_1$, it follows that there exists a positive δ such that we have

$$\|\eta'(t)\| \leq 1 + \gamma_1 \quad (|t - t_1| \leq \delta). \quad (37.12)$$

Take now

$$\rho := \text{Min}\left(\delta, \frac{Q}{2\gamma(1 + \gamma_1)}\right). \quad (37.13)$$

Then it follows that

$$\rho \|\eta'(t)\| \leq \frac{Q}{2\gamma} \quad (t \in J_\rho). \quad (37.14)$$

On the other hand, we have from (37.11) if $t \in J_\rho$, using Theorem 35.1,

$$\|\Lambda_1 y(t)\| \leq \gamma |t - t_1| \frac{Q}{2\gamma\rho} \leq \frac{Q}{2}$$

and we see that condition (37.10) is indeed satisfied for $t \in J_\rho$. Therefore, by Theorem (32.2), Eq. (37.7) is uniquely solvable in X_0 with $x = x(t)$ for

$$y = \eta(t) - \eta_1 \quad (t \in J_\rho). \quad (37.15)$$

8. Since Q can be chosen arbitrarily small, we see that $\xi(t) := x(t) + \xi_1$ is continuous at $t = t_1$.

In order to prove the existence of $\xi'(t_1)$ put

$$P := \frac{\xi(t) - \xi_1}{t - t_1}, \quad S := \Lambda_1 \frac{\eta(t) - \eta_1}{t - t_1}.$$

We have obviously $S \rightarrow \Lambda_1 \eta'(t_1)$ as $t \rightarrow t_1$. On the other hand, since the L -gradient $L(\xi)$ is assumed continuous at ξ_1 , by Theorem 36.3, L_1 is an F -gradient of f at ξ_1 and we have therefore

$$\eta(t) = f(\xi(t)) = \eta_1 + L_1(\xi(t) - \xi_1) + o(\|\xi(t) - \xi_1\|).$$

Bringing η_1 to the left and dividing on both sides by $t - t_1$, it follows that

$$\frac{\eta(t) - \eta_1}{t - t_1} = L_1 P + o(\|P\|) \quad (t \rightarrow t_1).$$

Multiply this on both sides by Λ_1 ; we obtain further

$$S = P + o(\|P\|), \quad (37.16)$$

as $\|\Lambda_1\| \leq \gamma$.

From (37.16) it follows, for sufficiently small $|t - t_1|$, that

$$\|S\| \geq \|P\| - \frac{\|P\|}{2} = \frac{\|P\|}{2}.$$

Denote $\Lambda_1 \eta'(t_1)$ by β ; then we have further, since $S = \beta + o(1)$, $\|S\| = \|\beta\| + o(1)$, that $\lim \|P\| \leq 3\|\beta\|$ as $t \rightarrow t_1$, and therefore $o(\|P\|)$ in (37.16) is $o(1)$. But now it follows from (37.16) that $P \rightarrow \beta$, that is, (37.2). Theorem 37.2 is now proved.

9. We will need in what follows a

Corollary to Theorem 37.2. *Assume, in the assumptions of Theorem 37.2 that $L(\xi)$ is continuous in K and $\eta'(t)$ is continuous in J_r . Then the function $\xi(t)$, the existence of which was proved in Theorem 37.2, is continuous and has a continuous derivative $\xi'(t) = \Lambda(\xi(t))\eta'(t)$ for all t with $|t - t_1| < \rho$.*

Indeed, for any t_2 with $|t_2 - t_1| < \rho$ we can apply Theorem 37.2, replacing t_1 with t_2 . Then we obtain again the same function $\xi(t)$ in a neighborhood of t_2 and it follows the continuity of $\xi(t)$ at t_2 , the existence of $\xi'(t_2)$, and the

relation

$$\xi'(t_2) = \Lambda(\xi(t_2))\eta'(t_2).$$

Since the right-hand expression is continuous for all t_2 with $|t_2 - t_1| < \rho$, the same follows for the left-hand expression. Our corollary is proved.

PROOF OF THEOREM 37.1

10. Proof of Theorem 37.1. Put

$$f(\xi) := \Omega(\xi) - \Omega_0, \quad \eta(t) := -t\Omega_0 \quad (|t| \leq 1). \quad (37.17)$$

We consider the set S of all T with the following properties:

- (1) It is $0 < T \leq 1$;
- (2) for any t with $0 \leq t < T$ there exists a $\xi(t) \in K$ with a continuous derivative $\xi'(t)$, $\xi(0) = \xi_0$, and

$$f(\xi(t)) = \eta(t) = -t\Omega_0. \quad (37.18)$$

We prove first that S is not empty. Indeed, every sufficiently small positive number is a T with the above property. This follows at once from the Theorem 37.2 and its corollary by putting there

$$\xi_1 := \xi_0, \quad t_1 := 0, \quad \eta_1 := 0, \quad r := 1, \quad \eta'(t) = -\Omega_0, \quad \gamma := R/\|\Omega_0\|.$$

Indeed, for these values the conditions of Theorem 37.2 and of its corollary are satisfied and it follows that any ρ , the existence of which was proved in this theorem, is a T .

11. Consider now two arbitrary different values T_1, T_2 from S ,

$$0 < T_1 < T_2 \leq 1, \quad (37.19)$$

and denote the corresponding $\xi(t)$ -functions with $\xi_1(t), \xi_2(t)$. Then we are going to prove that

$$\xi_1(t) = \xi_2(t) \quad (0 \leq t \leq T_1). \quad (37.20)$$

Indeed, it is clear that Eq. (37.20) holds for all sufficiently small t , since $\xi_1(t)$ and $\xi_2(t)$ are continuous at $t = 0$ and for $t_1 = 0$ we can apply the unicity statement of Theorem 37.2. Therefore, if relation (37.20) is not true for all values of $t < T_1$, the values of t for which $\xi_2(t) \neq \xi_1(t)$ have a positive lower bound τ , so that we have $\xi_2(t) = \xi_1(t)$ ($0 \leq t < \tau$), while every neighborhood of τ contains points t with $\xi_2(t) \neq \xi_1(t)$. Both functions, $\xi_1(t)$ and $\xi_2(t)$, are continuous at $\tau \leq T_1 < T_2$. Therefore we must also have $\xi_1(\tau) = \xi_2(\tau)$. But now, taking in Theorem 37.2 and its corollary $t_1 = \tau$, it follows that Eq.

(37.18) has a unique continuous solution in a neighborhood of τ , contrary to the definition of τ . We see that (37.20) is indeed true.

12. Consider now the supremum T^* of all T from S . Then it follows from the relation (37.20) that T^* is also an element of S .

Obviously $0 < T^* \leq 1$. It follows then from the corollary to Theorem 37.2 that

$$\begin{aligned}\|\xi'(t)\| &\leq \|\Lambda(\xi(t))\| \|\eta'(t)\| \leq (R/\|\Omega_0\|) \|\Omega_0\| = R, \\ \|\xi'(t)\| &\leq R \quad (0 \leq t < T^*).\end{aligned}\quad (37.21)$$

Consider now a sequence t_v increasing strictly monotonically to T^* , $t_v \uparrow T^*$. Then we obviously have, using (37.21),

$$\|\xi(t_v) - \xi(t_\mu)\| \leq R |t_v - t_\mu|$$

and it follows that $\xi(t_v)$ form a Cauchy sequence, which converges to a $\xi^* \in K$.

On the other hand, we have

$$\|\xi^* - \xi_0\| = \lim_{v \rightarrow \infty} \|\xi(t_v) - \xi_0\| \leq \lim_{v \rightarrow \infty} Rt_v = RT^*.$$

If T^* were < 1 , it would follow from Eq. (37.18) for $t_v \uparrow T^*$ that

$$f(\xi^*) = \eta(T^*)$$

and, since ξ^* is an inner point of K , we could again apply Theorem 37.2 and its corollary, taking $t_1 := T^*$, $\xi_1 := \xi^*$. But then the function $\xi(t)$ could be continued beyond the value of T^* , contrary to the definition of T^* . Therefore we must have $T^* = 1$ and it follows again from (37.18), writing it for t , and going to the limit, that

$$f(\xi^*) = \eta(1) = -\Omega_0.$$

Our ξ^* is a zero of Ω in K . Theorem 37.1 is proved.

38

Newton–Raphson Iteration in Banach Spaces. Statement of the Theorems

DEFINITION OF THE α_v

1. Throughout Chapters 38–40 we will use the following setup without repeating it in the assumptions of the individual theorems.

Consider a Banach space X and a point $\xi_0 \in X$. Let an operator f map a neighborhood of ξ_0 into a normed linear space Y and assume $f_0 := f(\xi_0) \neq 0$.

We assume that the F -gradient of f exists at ξ_0 ; put

$$P_0 := \text{grad } f(\xi_0), \quad (38.1)$$

and assume also the existence of $Q_0 := P_0^{-1}$. Writing

$$h_0 := -Q_0 f_0, \quad \|h_0\| \leq \|Q_0\| \|f_0\|, \quad (38.2)$$

we put further

$$\xi_1 := \xi_0 + h_0, \quad S_0 := \langle \xi_0, \xi_1 \rangle, \quad (38.3)$$

2. In the following we will use a fixed $\alpha \geq 2$. Writing

$$\alpha = 1 + \cos \varphi = 1 + \frac{e^\varphi + e^{-\varphi}}{2}, \quad \varphi \geq 0,^† \quad (38.4)$$

φ is uniquely determined by α . Put generally

$$\alpha_v := 1 + \cos 2^v \varphi \quad (v = 0, 1, \dots), \quad \alpha_0 := \alpha. \quad (38.5)$$

From (38.5) it follows that $\alpha_1 = 1 + \cos 2\varphi = 1 + \cos^2 \varphi + \sin^2 \varphi = 2 \cos^2 \varphi$,

$$\alpha_1 = 2(\alpha_0 - 1)^2. \quad (38.6)$$

Replacing φ by $2^v \varphi$ here, it follows generally that

$$\alpha_{v+1} = 2(\alpha_v - 1)^2 \quad (v \geq 0). \quad (38.7)$$

† We denote by Sin, Cos, Tg and Ctg the corresponding hyperbolic sine, cosine, tangent and cotangent functions, that is

$$\text{Sin } \varphi := \frac{e^\varphi - e^{-\varphi}}{2}, \quad \text{Cos } \varphi := \frac{e^\varphi + e^{-\varphi}}{2}, \quad \text{Tg } \varphi := \frac{e^\varphi - e^{-\varphi}}{e^\varphi + e^{-\varphi}}, \quad \text{Ctg } \varphi := \frac{e^\varphi + e^{-\varphi}}{e^\varphi - e^{-\varphi}}.$$

Using the well-known formulas

$$1 + \cos 2u = 2 \cos^2 u, \quad \cos u = \frac{\sin 2u}{2 \sin u},$$

we have further, if $\varphi > 0$,

$$\alpha_0 = 2 \cos^2(\varphi/2) = \frac{\frac{1}{2} \sin^2 \varphi}{\sin^2(\varphi/2)} = \frac{\sin \varphi}{\operatorname{Tg}(\varphi/2)}, \quad (38.8)$$

$$\frac{\alpha_0}{\alpha_0 - 1} = \frac{\operatorname{Tg} \varphi}{\operatorname{Tg}(\varphi/2)}, \quad (38.9)$$

$$\alpha_v = \frac{\frac{1}{2} \sin^2 2^v \varphi}{\sin^2 2^{v-1} \varphi}. \quad (38.10)$$

Further, it follows immediately, if $\varphi > 0$, that

$$\prod_{v=p}^{q-1} \cos 2^v \varphi = \frac{1}{2^{q-p}} \frac{\sin 2^q \varphi}{\sin 2^p \varphi} \quad (q > p \geq -1) \quad (38.11)$$

and, using $\cos^2 u / \cos 2u = \operatorname{Tg} 2u / 2 \operatorname{Tg} u$,

$$\prod_{v=p+1}^q \frac{\cos^2 2^{v-1} \varphi}{\cos 2^v \varphi} = 2^{p-q} \frac{\operatorname{Tg} 2^q \varphi}{\operatorname{Tg} 2^p \varphi} \quad (q > p \geq -1).$$

Finally, by virtue of (38.10),

$$\prod_{v=p+1}^q \alpha_v = \frac{1}{2^{q-p}} \frac{\sin^2 2^q \varphi}{\sin^2 2^p \varphi} \quad (q > p \geq -1). \quad (38.12)$$

If $\alpha_0 = 2$, we have $\varphi = 0$, $\alpha_0 = \alpha_1 = \alpha_2 = \dots = 2$, while in formulas (38.8)–(38.12) and in similar formulas later on the quotients of \sin functions or of Tg functions have to be replaced with their limits as $\varphi \downarrow 0$.

FORMULATION OF THEOREMS 38.1–38.3

3. Theorem 38.1. Put

$$\rho_0 := e^{-\varphi} \|h_0\| \quad (38.13)$$

and

$$\sigma_0 := \alpha_0 \|f_0\| \|Q_0\|^2. \quad (38.14)$$

Consider the closed ball around ξ_1 with the radius ρ_0 ,

$$(K_0) \quad (\|\xi - \xi_1\| \leq \rho_0), \quad (38.15)$$

and put

$$C_0 := S_0 \cup K_0, \quad (38.16)$$

where C_0 has a shape similar to a ping-pong paddle (Fig. 11 in Section 5).

Assume now that f maps a neighborhood of every point of C_0 into Y and that throughout C_0 the F -gradient

$$P(\xi) := \operatorname{grad} f(\xi) \quad (\xi \in C_0) \quad (38.17)$$

exists and satisfies the Lipschitz condition on S_0 as well as on K_0 :

$$\|P(\xi') - P(\xi)\| \leq \|\xi' - \xi\|/\sigma_0 \quad (\xi' \wedge \xi \in S_0 \vee K_0). \quad (38.18)$$

Then the procedure leading from ξ_0 to ξ_1 can be iterated indefinitely; the resulting sequence ξ_v ($v \geq 1$) lies in K_0 and tends to a zero $\zeta \in K_0$ off(ζ). Further, the following inequalities hold:

$$\|\xi_v - \zeta\| \leq \exp(-2^{v-1}\varphi) \frac{\sin \varphi}{\sin 2^{v-1}\varphi} \|h_0\| \quad (v \geq 1), \quad (38.19)$$

$$\|\xi_v - \zeta\| \leq 2^{1-v} \|h_0\| \quad (v \geq 1), \quad (38.19a)$$

$$\|\xi_{v+1} - \zeta\| \leq \exp(-2^v\varphi) \|\xi_{v+1} - \xi_v\|. \quad (38.20)$$

4. Theorem 38.2. In the hypotheses and the notations of Theorem 38.1 we have, for $v = 1, 2, \dots$, if $f(\xi_n) \neq 0$,

$$\frac{\|f(\xi_{n+v})\|}{\|f(\xi_n)\|} \leq \frac{\sin^2 2^{n-1}\varphi}{\sin^2 2^{n+v-1}\varphi} \leq \frac{1}{2^{2v}}; \quad (38.21)$$

further, putting for $v = 0, 1, \dots$

$$\begin{aligned} P_v &:= P(\xi_v), & Q_v &:= P_v^{-1}, & f_v &:= f(\xi_v), \\ h_v &:= Q_v f_v, & \sigma_v &:= \alpha_v \|f_v\| \|Q_v\|^2: \\ \|h_{v+1}\| &\leq \|h_v\|/(2\alpha_v - 2), \end{aligned} \quad (38.22)$$

$$\frac{\alpha_v}{\alpha_v - 1} \geq \frac{\|Q_{v+1}\|}{\|Q_v\|} \geq \frac{\alpha_v}{\alpha_v + 1} > \frac{\alpha_v - 1}{\alpha_v} \quad (v = 0, 1, \dots), \quad (38.23)$$

$$\frac{\operatorname{Tg} 2^{v-1}\varphi}{\operatorname{Tg} 2^{v-1}\varphi} \leq \frac{\|Q_\mu\|}{\|Q_v\|} \leq \frac{\operatorname{Tg} 2^{\mu-1}\varphi}{\operatorname{Tg} 2^{v-1}\varphi} \quad (\mu > v), \quad (38.24)$$

$$\frac{\|h_{v+1}\|}{\|h_v\|^2} \leq \frac{\|Q_{v+1}\|}{2\sigma_0} \leq \frac{\|Q_0\|}{2\sigma_0} \operatorname{Ctg} \frac{\varphi}{2} \quad (v = 0, 1, \dots, \varphi > 0, \alpha_0 > 2). \quad (38.25)$$

If $\alpha = 2, \varphi = 0$, the right-hand expressions in (38.21), (38.24) have to be replaced respectively with $2^{-2v}, 2^{\mu-v}$, that is, with their limits as $\varphi \downarrow 0$.

5. We will prove Theorems 38.1 and 38.2 together with the following theorem.

Theorem 38.3. *Let B be an open, connected, and convex set in X , containing ξ_0 . Put (see Fig. 11)*

$$K_0^* := K_0 \cap B$$

and assume that, if ξ goes from K_0^ to the points of the boundary ∂B of B , lying in K_0 , the following relation holds:*

$$\underline{\lim} \|f(\xi)\| \geq \|f_0\| \quad (\xi \rightarrow (\partial B) \cap K_0, \quad \xi \in K_0^*). \quad (38.26)$$

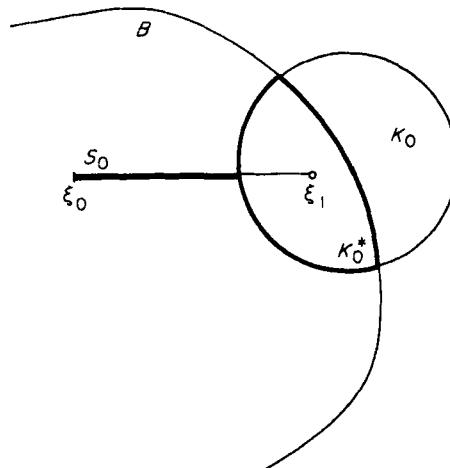


FIGURE 11

Assume that the assumptions of Theorem 38.1 are satisfied if we replace in them K_0 with K_0^* .

Then the procedure leading from ξ_0 to ξ_1 can be iterated indefinitely; the resulting sequence ξ_v ($v \geq 1$) lies in K_0^* and tends to a zero $\zeta \in K_0^*$ of $f(\xi)$. Further, (38.19), (38.19a), (38.20), and all the relations of Theorem 38.2 hold.

A LEMMA

6. We shall use the following lemma in the proof of our theorems.

Lemma 38.1. *Assume a τ with $0 < \tau \leq 1$ and denote the closed interval between ξ_0 and $\xi_0 + \tau h_0$ by*

$$T_\tau := \langle \xi_0, \xi_0 + \tau h_0 \rangle.$$

Assume that $f(\xi)$ exists in a neighborhood of every point of T_τ and $P(\xi) := \text{grad}f(\xi)$ exists on T_τ and there satisfies relation (38.18). Further, assume the notation (38.14).

Then we have in all points, $\xi_0 + th_0$, of T_τ :

$$\|f(\xi_0 + th_0)\| \leq (1-t) \|f_0\| + \frac{t^2}{2\sigma_0} \|h_0\|^2, \quad (38.27)$$

$$\|f(\xi_0 + th_0)\| \leq \|f_0\| \left(1 - t + \frac{t^2}{2\sigma_0} \|Q_0\| \|h_0\| \right), \quad (38.28)$$

$$\|f(\xi_0 + th_0)\| \leq \|f_0\| \left(1 - \frac{t}{2} \right)^2. \quad (38.29)$$

In (38.28) and (38.29) the equality sign for $t = \tau$ is possible only if we have

$$\|h_0\| = \|Q_0\| \|f_0\|. \quad (38.30)$$

If we have the equality sign for $t = \tau$ in one of the relations (38.27)–(38.29), then, putting

$$H(t) := f(\xi_0 + th_0), \quad H'(t) = P(\xi_0 + th_0)h_0, \quad (38.31)$$

we have generally

$$\|H'(t_2) - H'(t_1)\| = |t_2 - t_1| \|h_0\|^2 / \sigma_0 \quad (0 \leq t_1 \leq t_2 \leq \tau). \quad (38.32)$$

7. Proof of Lemma 38.1. Put $M := \|h_0\|^2 / \sigma_0$. It follows from (38.18), if a', b' lie between 0 and τ , that

$$\begin{aligned} H'(b') - H'(a') &= (P(\xi_0 + b'h_0) - P(\xi_0 + a'h_0))h_0, \\ \|H'(b') - H'(a')\| &\leq |b' - a'| M. \end{aligned} \quad (38.33)$$

If there exists a couple $a < b$, both from the open interval $(0, \tau)$, for which (38.33) is a strict inequality, we can assume, as $H'(t)$ is continuous, that for a $p > 0$

$$\|H'(b) - H'(a)\| \leq (b - a) M (1 - p). \quad (38.34)$$

If no such couple exists, we can still write (38.34), taking there $a = b = \tau/2$.

Put now

$$M(u) = \begin{cases} uM & (0 \leq u \leq b) \\ (u - p(b-a))M & (b < u \leq \tau). \end{cases} \quad (38.35)$$

Then we have, by (38.33) and (38.34),

$$\|H'(u) - H'(0)\| \leq M(u) \quad (0 \leq u \leq t).$$

As the left-hand expression here is $\|(H(u) - uH'(0))'_u\|$, it follows that

$$\|H(t) - tH'(0) - H(0)\| \leq \int_0^t M(u) du = M\left(\frac{t^2}{2} - \Delta(t)\right),$$

where

$$\Delta(t) = \begin{cases} 0 & (t \leq b) \\ (t-b)(b-a)p & (b < t \leq \tau), \end{cases} \quad (38.36)$$

or

$$H(t) = H(0) + tH'(0) + \theta M\left(\frac{t^2}{2} - \Delta(t)\right), \quad \|\theta\| \leq 1. \quad (38.37)$$

8. Since $H(0) = f_0$, $H'(0) = P_0 h_0 = -P_0 Q_0 f_0 = -f_0$, we can rewrite (38.7) as

$$f(\xi_0 + th_0) = (1-t)f_0 + \theta M\left(\frac{t^2}{2} - \Delta(t)\right). \quad (38.38)$$

As $\Delta(t) \geq 0$, (38.27) follows immediately from (38.38) if we introduce the value of M . The equality sign in (38.27) is obviously possible for $t = \tau$ only if $\Delta(\tau) = 0$, that is, if (38.32) holds.

Using (38.2), the bound in (38.27) is majorated by the bound in (38.28). The bound in (38.28) is again majorated by the bound in (38.29), since

$$\frac{t^2}{2} \|Q_0\| \|h_0\|/\sigma_0 \leq \frac{t^2}{2} \|Q_0\|^2 \|f_0\|/\sigma_0 = \frac{t^2}{2\alpha_0} \leq \frac{t^2}{4}.$$

Therefore it is clear that from the equality sign for $t = \tau$ in (38.28) or (38.29) there follows the one in (38.27), and therefore (38.32). On the other hand, passing from (38.27) to (38.28), we use the inequality (38.2) and it follows therefore from the equality sign in (38.28) or in (38.29) for $t = \tau$ that (38.30) holds. Lemma 38.1 is proved.

39

Proof of Theorems 38.1–38.3

A FURTHER LEMMA

1. Lemma 39.1. *Assume, in the hypotheses of Lemma 38.1, that $\tau = 1$, $T_1 = S_0$. Then $Q_1 := P_1^{-1}$ exists and*

$$\frac{\alpha_0 - 1}{\alpha_0} \|Q_0\| < \frac{\alpha_0}{\alpha_0 + 1} \|Q_0\| \leq \|Q_1\|, \quad (39.1)$$

$$\|Q_1\| \leq \frac{\alpha_0}{\alpha_0 - 1} \|Q_0\|, \quad (39.2)$$

$$\|f_1\| \leq \frac{\|f_0\| \|Q_0\| \|h_0\|}{2\sigma_0}, \quad (39.3)$$

$$\|f_1\| \leq \frac{\|f_0\|}{2\alpha_0}, \quad (39.4)$$

$$\|h_1\| \leq \frac{\|h_0\|}{e^\varphi + e^{-\varphi}}, \quad (39.5)$$

$$\|h_1\| \leq \frac{\|h_0\|}{2}, \quad (39.6)$$

$$\sigma_1 = \alpha_1 \|Q_1\|^2 \|f_1\| \leq \sigma_0. \quad (39.7)$$

2. Proof of Lemma 39.1. Put

$$\omega := 1 - Q_0 P_1 = Q_0 (P_0 - P_1).$$

Then it follows from (38.18) and (38.14) that

$$\|\omega\| = \|Q_0(P_1 - P_0)\| \leq \|Q_0\| \|h_0\| / \sigma_0 \leq \|Q_0\|^2 \|f_0\| / \sigma_0 = 1/\alpha_0 \leq \frac{1}{2}. \quad (39.8)$$

Therefore the geometric series in powers of ω is convergent and

$$U := \sum_{v=0}^{\infty} \omega^v = \frac{\theta}{1 - \|\omega\|}, \quad \|\theta\| \leq 1.$$

But now

$$U(1-\omega) = 1, \quad UQ_0 P_1 = 1, \quad Q_1 = UQ_0, \quad Q_0 = (1-\omega)Q_1,$$

$$\|Q_1\| \leq \frac{\|Q_0\|}{1-\|\omega\|} \leq \frac{\alpha_0}{\alpha_0-1} \|Q_0\|,$$

$$\|Q_0\| \leq \|Q_1\| \|1-\omega\| \leq \|Q_1\| (1+\|\omega\|) \leq \frac{\alpha_0+1}{\alpha_0} \|Q_1\| < \frac{\alpha_0}{\alpha_0-1} \|Q_1\|;$$

relations (39.1) and (39.2) follow immediately.

3. Relation (39.3) follows immediately from (38.28) for $t = 1$, while by introducing the estimate (38.2) of h_0 into (39.3) and using (38.14), we have (39.4):

$$\|f_1\| \leq \frac{\|f_0\| \|Q_0\| \|Q_0\| \|f_0\|}{2\sigma_0} = \frac{\|f_0\|}{2\alpha_0}.$$

Relation (39.5) follows from (39.2), (38.14), and (39.3):

$$\|h_1\| \leq \|Q_1\| \|f_1\| \leq \frac{\alpha_0}{\alpha_0-1} \|Q_0\| \frac{\|f_0\| \|Q_0\| \|h_0\|}{2\sigma_0} = \frac{\|h_0\|}{2(\alpha_0-1)}.$$

Relation (39.6) follows immediately from (39.5) since $e^\varphi + e^{-\varphi} \geq 2$.

Finally, from the value of σ_1 in (39.7) and from (39.2), (39.4) and (38.6) it follows that

$$\sigma_1 \leq 2(\alpha_0-1)^2 \frac{\alpha_0^2}{(\alpha_0-1)^2} \|Q_0\|^2 \frac{\|f_0\|}{2\alpha_0} = \sigma_0,$$

that is, (39.7). Lemma 39.1 is proved.

SPECIALIZATION FOR QUADRATIC POLYNOMIALS

4. It is of interest for later discussions to specialize the argument used in the proofs of Lemmas 38.1 and 39.1 to the case where $f(x)$ is a *real quadratic polynomial* in the *real variable* ξ ,

$$f(\xi) := \xi^2 + a\xi + b.$$

We assume in particular that we have, for the starting value ξ_0 ,

$$f_0 = f(\xi_0) > 0, \quad P_0 = f'(\xi_0) < 0, \quad Q_0 = 1/f'(\xi_0) < 0 \quad (39.9)$$

and further that

$$f_0 Q_0^2 \leq \frac{1}{4}. \quad (39.10)$$

If we choose then

$$\alpha_0 := \frac{1}{2f_0 Q_0^2},$$

α_0 is ≥ 2 and it follows from (38.14) that $\sigma_0 = \frac{1}{2}$.

On the other hand, since $f'(\xi)$ is linear and $f''(\xi) \equiv 2$, we have

$$f'(\xi'') - f'(\xi') = 2(\xi'' - \xi') \quad (39.11)$$

and (38.18) is verified, indeed, with $\sigma_0 = \sigma = \frac{1}{2}$.

5. Formula (38.27) obviously remains satisfied if all moduli are replaced by the corresponding numbers, and for $t = 1$ it follows that

$$f_1 = f(\xi_0 + h_0) = h_0^2. \quad (39.12)$$

On the other hand, it follows from (39.11) that

$$\begin{aligned} f'(\xi_1) &= f'(\xi_0) + 2h_0 = \frac{1}{Q_0} - 2Q_0 f_0 = \frac{1 - 2f_0 Q_0^2}{Q_0} = \frac{1 - 1/\alpha_0}{Q_0}, \\ Q_1 &= \frac{Q_0}{1 - 1/\alpha_0} = \frac{\alpha_0}{\alpha_0 - 1} Q_0 < 0. \end{aligned} \quad (39.13)$$

It follows now that

$$\begin{aligned} f_1 Q_1^2 &= h_0^2 Q_0^2 \alpha_0^2 / (\alpha_0 - 1)^2 = f_0^2 Q_0^4 \alpha_0^2 / (\alpha_0 - 1)^2 \\ &= \sigma_0^2 / (\alpha_0 - 1)^2 \leq \sigma_0^2 \leq \frac{1}{4}, \\ f_1 Q_1^2 &\leq \frac{1}{4} \end{aligned} \quad (39.14)$$

and

$$\begin{aligned} \sigma_1 &= \alpha_1 (f_1 Q_1^2) = 2(\alpha_0 - 1)^2 \frac{\sigma_0^2}{(\alpha_0 - 1)^2} = 2\sigma_0^2 = \sigma_0, \\ \sigma_1 &= \sigma_0 = \frac{1}{2}. \end{aligned} \quad (39.15)$$

Finally, it follows from (39.13) and (39.12) that

$$\begin{aligned} h_1 &= Q_1 f_1 = Q_0 \frac{\alpha_0}{\alpha_0 - 1} h_0^2 = \frac{\alpha_0 Q_0^2 f_0}{\alpha_0 - 1} h_0 = \frac{h_0}{2(\alpha_0 - 1)}, \\ h_1 &= h_0 \frac{\sin \varphi}{\sin 2\varphi}. \end{aligned} \quad (39.16)$$

PROOFS OF THEOREMS 38.1–38.3

6. In the hypotheses of Theorem 38.1, the conditions of Lemmas 38.1 and 39.1 are obviously satisfied as $T_1 = S_0 \subset C_0$, so that relations (38.27)–(38.29), (39.1)–(39.7) are valid.

We are now going to prove that in the hypotheses of Theorem (38.3), $S_0 \subset B$. Indeed, if S_0 were not contained in B , there would exist a t' such that

$$0 < t' \leq 1, \quad \xi_0 + t'h_0 \notin B.$$

Denote the infimum of all such t' by t_0 . Obviously $t_0 > 0$, since a neighborhood of ξ_0 lies in B .

For any τ with $0 < \tau < t_0$, the hypotheses of Lemma 38.1 are satisfied, and therefore by (38.29)

$$\|f(\xi_0 + \tau h_0)\| \leq \|f_0\| \left(1 - \frac{\tau}{2}\right)^2.$$

Therefore

$$\overline{\lim} \|f(\xi_0 + \tau h_0)\| \leq \|f_0\| \left(1 - \frac{\tau}{2}\right)^2 < \|f_0\| \quad (\tau \uparrow t_0),$$

so that by (38.26) $\xi_0 + t_0 h_0$ does not lie on the boundary of B . Since, therefore, in the notation of Lemma 38.1, the whole interval T_{t_0} has no points in common with ∂B , it consists completely of inner points of B , which contradicts the definition of t_0 . We see that S_0 lies completely in K_0^* and therefore all formulas (38.27)–(38.29) and (39.1)–(39.7) are valid in this case, too.

7. Put

$$\rho_1 := e^{-2\varphi} \|h_1\|, \quad \xi_2 := \xi_1 + h_1, \quad (39.17)$$

$$S_1 := \langle \xi_1, \xi_2 \rangle, \quad K_1 := U_{\rho_1}(\xi_2), \quad C_1 := S_1 \cup K_1.$$

We are now going to prove that $C_1 \subset K_0$ and *a fortiori* $C_1 \subset C_0$.

Observe that all points ξ of C_1 can be written in the form

$$\xi = \xi_1 + \theta_1 h_1 + \theta_2 \rho_1, \quad \|\theta_1\| \leq 1, \quad \|\theta_2\| \leq 1;$$

it is sufficient to prove that $\|\xi - \xi_1\| \leq \rho_0$, and this will follow from

$$\|h_1\| + \rho_1 \leq \rho_0. \quad (39.18)$$

But the left-hand expression in (39.18) is, by (39.5) and (39.17),

$$\leq \|h_0\| \left(\frac{1}{e^\varphi + e^{-\varphi}} + \frac{e^{-2\varphi}}{e^\varphi + e^{-\varphi}} \right) = e^{-\varphi} \|h_0\| = \rho_0,$$

and we have proved that $C_1 \subset K_0$.

8. We see now that (38.18) holds also, in the hypotheses of our Theorem 38.1, on C_1 and it follows that unless $f(\xi_1) = 0$, all hypotheses of our Theorem 38.1 remain satisfied if in them we replace

$$\xi_0, \alpha_0, f_0, Q_0, \rho_0, h_0, S_0, K_0, \sigma_0$$

respectively with

$$\xi_1, \alpha_1, f_1, Q_1, \rho_1, h_1, S_1, K_1, \sigma_1.$$

Further, in the hypotheses of Theorem 38.3, again, all assumptions remain satisfied if we replace as above the index 0 with the index 1 and K_0^* with $K_1^* := B \cap K_1$.

We can therefore apply the same argument, starting from ξ_1 and the geometric configuration C_1 or—in the hypotheses of Theorem 38.3— $S_1 \cup K_1^*$, and can go on in this way indefinitely.

If in the sequence of the ξ_μ there happens to be one for which $f(\xi_\mu) = 0$, we have from there on $\xi_\mu = \xi_{\mu+1} = \dots = \zeta, f(\zeta) = 0$, so that we can assume in the following proof that the ξ_μ with which we are dealing are not zeros of f .

9. Since, as we have seen, $C_1 \subset K_0, K_1^* \subset K_0 \cap B$, it follows now more generally that

$$C_{v+1} \subset K_v, \quad K_{v+1}^* \subset K_v \cap B,$$

and therefore

$$C_\mu \subset K_n, \quad K_\mu^* \subset K_n \cap B \quad (\mu > n),$$

$$\xi_{\mu+1} \in K_n, \quad \|\xi_{\mu+1} - \xi_{n+1}\| \leq \rho_n := \exp(-2^n \varphi) \|h_n\| \quad (\mu > n). \quad (39.19)$$

As, by virtue of (39.6), generally

$$\|h_{v+1}\| \leq \|h_v\|/2,$$

we obtain

$$\rho_n \leq \exp(-2^n \varphi) 2^{-n} \|h_0\| \rightarrow 0 \quad (n \rightarrow \infty).$$

It follows therefore from (39.19) that the ξ_μ form a Cauchy sequence and, since X is assumed as complete, we have indeed $\xi_v \rightarrow \zeta$ ($v \rightarrow \infty$).

Since by (39.19) any K_n contains all ξ_v , save a finite number of them, we see that ζ lies in any of the balls K_n ($n = 0, 1, \dots$).

10. Letting μ in (39.19) go to ∞ , we obtain

$$\|\xi_{n+1} - \zeta\| \leq \rho_n. \quad (39.20)$$

As to the value of ρ_n , it follows, since by (39.5) $\|h_{v+1}\| \leq \|h_v\|/(2 \cos 2^v \varphi)$, that

$$\|h_n\| \leq 2^{-n} \|h_0\| \prod_{v=0}^{n-1} \frac{1}{\cos 2^v \varphi},$$

and using (38.11) with $p = 0, q = n$, we obtain

$$\|h_n\| \leq \|h_0\| \frac{\sin \varphi}{\sin 2^n \varphi}. \quad (39.21)$$

We now obtain from the definition of ρ_n in (39.19)

$$\rho_n \leq \exp(-2^n \varphi) \|h_0\| \frac{\sin \varphi}{\sin 2^n \varphi}. \quad (39.22)$$

Introducing this into (39.20) and replacing there n with $v - 1$, the inequality (38.19) is proved.

11. As to the inequality (38.19a), it follows obviously at once if we prove that generally

$$\frac{\sin a}{\sin b} < \frac{a}{b} \quad (0 < a < b), \quad (39.23)$$

that is, that $(\sin x)/x$ strictly monotonically increases for $x > 0$. But the derivative of this expression is

$$\frac{\cos x}{x} - \frac{\sin x}{x^2} = \frac{\cos x}{x^2}(x - \operatorname{Tg} x).$$

Putting $g(x) := x - \operatorname{Tg} x$, it follows that $g(0) = 0$,

$$g'(x) = 1 - \frac{1}{\cos^2 x} = \frac{\cos^2 x - 1}{\cos^2 x} = \operatorname{Tg}^2 x > 0 \quad (x > 0).$$

We see that $g(x) > 0$ as $x > 0$ and the strict monotony of $(\sin x)/x$ is proved for $x > 0$. Formula (38.19a) is proved. Finally, (38.20) follows from (38.19) and (39.20).

12. It remains to prove that ζ is a zero of f . This proof is slightly shortened if we use formula (38.21) of Theorem 38.2, which we are now going to prove.

It follows from (39.4) that generally

$$\frac{\|f(\xi_{v+1})\|}{\|f(\xi_v)\|} \leq \frac{1}{2\alpha_v} \quad (39.24)$$

and therefore

$$\frac{\|f(\xi_{n+v})\|}{\|f(\xi_n)\|} \leq \prod_{\kappa=n}^{n+v-1} \frac{1}{2\alpha_\kappa}.$$

Using (38.12) and (39.23), formula (38.21) follows immediately.

13. Putting in (38.21) $n = 0$ and letting v go to ∞ , it follows that

$$f(\xi_v) \rightarrow 0. \quad (39.25)$$

On the other hand, by the definition of the gradient,

$$\lim_{v \rightarrow \infty} \frac{\|f(\xi_v) - f(\zeta)\|}{\|\xi_v - \zeta\|} = \|P(\zeta)\| \quad (v \rightarrow \infty).$$

But then $|f(\xi_v) - f(\zeta)| \rightarrow 0$ and therefore

$$f(\xi_v) - f(\zeta) \rightarrow 0.$$

By virtue of (39.25), we obtain $f(\zeta) = 0$. Theorem 38.1 and formula (38.21) of Theorem 38.2 are now proved.

14. Replacing in (39.1), (39.2), and (39.5) the indices 0, 1 with v , $v+1$, (38.22) and (38.23) follow immediately. By virtue of (38.9) the bounds in (38.23) can be written as

$$\frac{\alpha_v}{\alpha_v - 1} = \frac{\operatorname{Tg} 2^v \varphi}{\operatorname{Tg} 2^{v-1} \varphi}, \quad \frac{\alpha_v - 1}{\alpha_v} = \frac{\operatorname{Tg} 2^{v-1} \varphi}{\operatorname{Tg} 2^v \varphi}. \quad (39.26)$$

Therefore, we obtain for $\mu > v$

$$\frac{\operatorname{Tg} 2^{v-1} \varphi}{\operatorname{Tg} 2^{\mu-1} \varphi} \leq \frac{\|Q_\mu\|}{\|Q_v\|} \leq \prod_{k=v}^{\mu-1} \frac{\alpha_k}{\alpha_k - 1} = \frac{\operatorname{Tg} 2^{\mu-1} \varphi}{\operatorname{Tg} 2^{v-1} \varphi},$$

and this is (38.24).

15. It follows from Section 9 that the segment $\langle \xi_v, \xi_v + h_v \rangle$ lies in K_0 . Formula (38.18) holds therefore along this segment with the denominator σ_0 . Writing formula (38.27) with $t = 1$ for ξ_v and h_v , it follows that

$$f_{v+1} = \theta \frac{\|h_v\|^2}{2\sigma_0}, \quad \|\theta\| \leq 1.$$

Since $h_{v+1} = -Q_{v+1} f_{v+1}$, it follows further that

$$\|h_{v+1}\| \leq \|Q_{v+1}\| \|h_v\|^2 / (2\sigma_0),$$

which is the first part of (38.25).

On the other hand, it follows from (38.24), if we replace v with 0 and μ with $v+1$, that

$$\|Q_{v+1}\| / \|Q_0\| \leq \operatorname{Tg} 2^v \varphi / \operatorname{Tg} (\varphi/2),$$

and further, since $\operatorname{Tg} u$ is positive and < 1 ($u > 0$), that

$$\|Q_{v+1}\| < \|Q_0\| / \operatorname{Tg} (\varphi/2).$$

The second part of (38.25) now follows immediately and Theorems 38.1, 38.2, 38.3 are completely proved.

40

Complements to the Newton–Raphson Method

EQUALITIES IN ESTIMATES FOR QUADRATIC POLYNOMIALS

1. In formulas (38.19a) and (38.21) the equality sign for the bounds 2^{1-v} , 2^{-2v} can be obtained only if $\varphi = 0$, since otherwise the inequality (39.23) can be used.

On the other hand, (38.19), the first inequality (38.21) and the second inequality (38.24) cannot be improved for any positive φ . Indeed, the real quadratic polynomial with real roots,

$$f_{x_0}(\xi) := (\xi - 1 - e^\varphi)(\xi - 1 - e^{-\varphi}) = \xi^2 - 2\alpha_0 \xi + 2\alpha_0, \quad (40.1)$$

satisfies the conditions of Theorems 38.1 and 38.2 for all $\varphi > 0$. On the other hand, if we start with $\xi_0 = 0$, we have the equality sign in all three of the inequalities mentioned, as shall be shown in Sections 8–10 of Appendix F.

2. The restriction $\varphi > 0$ in (38.25) is essential. Indeed, for the polynomial (40.1) with $\varphi = 0$, $\alpha_0 = 2$ we obtain easily

$$h_v = 2^{-v}, \quad \xi_v = 2 - 2^{1-v},$$

so that in this case the convergence is linear and not quadratic.

Observe that 2 is here a double root of $f_2(\xi)$.

MULTIPLE SOLUTIONS

3. In the case of analytic functions, the zero ζ is only in exceptional cases a multiple root. The corresponding result in the general case of Theorems 38.1–38.3 is the

Theorem 40.1. *Under the hypotheses of Theorem 38.1 or of Theorem 38.3 assume that the linear operator $P(\zeta)$ has no inverse. Then $\alpha_0 = 2$, $\varphi = 0$, ζ lies on the boundary of all balls K_v , and the equality sign holds in the relations (38.20), (38.21) and in the first relation (38.23). Further, we have*

$$\|h_{v+1}\| = \|h_v\|/2 \quad (v \geq 0). \quad (40.2)$$

4. Proof of Theorem 40.1. We can assume without loss of generality that none of the $h_v = 0$. It follows from (38.18) that for all ξ in the ball K_0

$$\|P(\xi) - P(\xi_1)\| \leq \frac{\|\xi - \xi_1\|}{\sigma_0} \leq \frac{\rho_0}{\sigma_0} = e^{-\varphi} \|Q_0\| \|f_0\|/\sigma_0, \quad (40.3)$$

thence

$$\|Q_1 P(\zeta) - 1\| \leq e^{-\varphi} \|Q_1\| \|Q_0\| \|f_0\|/\sigma_0$$

and therefore, using (38.23) and (38.14), we obtain

$$\|Q_1 P(\zeta) - 1\| \leq e^{-\varphi}/(\alpha_0 - 1). \quad (40.4)$$

Obviously, as soon as the right-hand expression in (40.4) is < 1 , $Q_1 P(\zeta)$ has an inverse and the same holds for $P(\zeta)$. Since ζ lies in the ball K_1 , it follows that we must have $\alpha_0 = 2$, $\varphi = 0$. On the other hand, if ζ does not lie on the boundary of K_0 , we have the strict inequality in (40.3) and therefore also in (40.4), where now the right-hand bound is 1. But then again $P(\zeta)$ has an inverse. We see that ζ must lie on the boundary of K_0 . Since the same argument holds for each K_v , it follows that ζ must lie on the boundary of each ball K_v , so that we have $|\zeta - \xi_{v+1}| = \rho_v = |h_v|$ and $\alpha_v = 2$ for every v .

5. On the other hand, if we had the strict inequality in (38.22) or in the first relation (38.23), we would have, for a certain $v > 0$, the strict inequality $\sigma_v < \sigma_0$. But this signifies that, starting anew with this value of v , we could choose our $\alpha_v > 2$.

Further, we have $\|h_v\| = \|\xi_{v+1} - \xi_v\| = |(\xi_v - \zeta) - (\xi_{v+1} - \zeta)| \geq \rho_{v-1} - \rho_v$. But (39.18), rewritten for h_v , gives $\|h_v\| \leq \rho_{v-1} - \rho_v$. It follows that

$$\|h_v\| = \rho_{v-1} - \rho_v = \|h_{v-1}\| - \|h_v\|, \quad \|h_v\| = \|h_{v-1}\|/2, \quad \|h_v\| = \|h_0\|/2^v.$$

Theorem 40.1 is proved. We pursue this analysis further in Appendix U.

UNICITY THEOREM

6. It will now be shown that in the hypotheses of Theorem 38.1, C_0 does not contain any zero of f different from ζ . On the other hand, consider the ball around ξ_0 with the radius $\|h_0\| + \rho_0 = (1 + e^{-\varphi}) \|h_0\|$, which will be denoted in the following by \bar{K} . Obviously \bar{K} contains K_0 and the boundaries of \bar{K} and K_0 have the point $\xi_1 + \rho_0$ in common. We will prove for \bar{K} the property that if the conditions of Theorem 38.1 are satisfied on \bar{K} , there is no zero of f different from ζ on \bar{K} .

This statement about \bar{K} is, of course, more precise than the corresponding statement about C_0 ; however, in the case of \bar{K} , the assumptions of Theorem 38.1 must also be assumed as satisfied on \bar{K} , that is, on a larger set than C_0 .

It is therefore obvious that it is of interest to find “unicity sets of $f = 0$ with respect to ξ_0 ,” intermediate between C_0 and \bar{K} . More precisely, we will call a set U containing C_0 a “unicity set of $f = 0$ with respect to ξ_0 ” if, whenever $f(\xi)$ with given f_0 , P_0 , and Q_0 exists, together with $P(\xi)$, on U and $P(\xi)$ satisfies on U the Lipschitz condition (38.18), U contains ζ but does not contain any zero of f different from ζ .

To define our class of unicity sets, we will denote as a “star set with respect to ξ_0 ” each set S containing ξ_0 and such that whenever a point $\xi \in S$, then the whole interval $\langle \xi_0, \xi \rangle$ lies in S .

7. Theorem 40.2. *Consider in the conditions of Theorem 38.1 a star set with respect to ξ_0 , S , which is contained in \bar{K} and contains the interval $S_0 := \langle \xi_0, \xi_1 \rangle$.*

Then the set

$$U := S \cup K_0 \quad (40.5)$$

is a unicity set of $f = 0$ with respect to ξ_0 . Under the hypotheses of Theorem 38.3, the above statement remains true if we replace in it U with

$$U^* := (S \cup K_0) \cap B. \quad (40.6)$$

8. Proof. Assume that a zero $\zeta' \neq \zeta$ of $f(\xi)$ lies in U .

We will first prove that if, for a $v \geq 0$, the whole interval $\langle \zeta', \xi_v \rangle$ lies in U , then the inequality holds:

$$\|\zeta' - \xi_{v+1}\| \leq \frac{\|Q_v\|}{2\sigma_0} \|\zeta' - \xi_v\|^2 \quad (\langle \zeta', \xi_v \rangle \subset U). \quad (40.7)$$

Indeed, proceeding as in Section 7 of Chapter 38, in the proof of Lemma 38.1 we obtain the development of $f(\zeta')$ around ξ_v ,

$$f(\zeta') = f(\xi_v) + P(\xi_v)(\zeta' - \xi_v) - \frac{\theta^*}{2\sigma_0} \|\zeta' - \xi_v\|^2, \quad \|\theta^*\| \leq 1,$$

and therefore, as $f(\zeta') = 0$, multiplying with $Q_v := P^{-1}(\xi_v)$, we obtain

$$Q_v f(\xi_v) + \zeta' - \xi_v = \frac{\|\zeta' - \xi_v\|^2}{2\sigma_0} Q_v \theta^*,$$

since $Q_v P(\xi_v) = 1$. But we have $Q_v f(\xi_v) = -h_v$, $\xi_v + h_v = \xi_{v+1}$ and we obtain therefore

$$\zeta' - \xi_{v+1} = \frac{\|\zeta' - \xi_v\|^2}{2\sigma_0} Q_v \theta^*.$$

Formula (40.7) follows immediately.

9. We will now show that in the hypotheses of Theorem 38.1 ζ' lies in K_0 . Indeed, otherwise $\zeta' \in S$, $\langle \xi_0, \zeta' \rangle \subset S$ and (40.7) can be applied for $v = 0$. But then it follows, using (38.14), that

$$\|\zeta' - \xi_1\| \leq \frac{\|Q_0\|}{2\sigma_0} \|\zeta' - \xi_0\|^2 \leq \frac{\|Q_0\| \|h_0\|^2}{2\alpha_0 \|f_0\| \|Q_0\|^2} (1 + e^{-\varphi})^2 \leq \frac{(1 + e^{-\varphi})^2}{2\alpha_0} \|h_0\|,$$

since $\|h_0\| \leq \|f_0\| \|Q_0\|$. Here

$$\frac{(1 + e^{-\varphi})^2}{2\alpha_0} = \frac{e^{-\varphi}(e^{\varphi/2} + e^{-\varphi/2})^2}{(e^{\varphi/2} + e^{-\varphi/2})^2} = e^{-\varphi},$$

and it follows from (38.13) that

$$\|\zeta' - \xi_1\| \leq e^{-\varphi} \|h_0\| = \rho_0,$$

that is, $\zeta' \in K_0$. In the hypotheses of Theorem 38.3, it follows in the same way, since B is convex, that $\zeta' \in K_0 \cap B$.

10. Since we now know that $\zeta' \in K_0$ (or, in the hypotheses of Theorem 38.3, $\zeta' \in K_0 \cap B$), the condition for (40.7) is satisfied for any $v \geq 0$. Using now (38.24) with $v = 1$ and replacing there μ with v , we obtain, by (39.2),

$$\|\zeta' - \xi_{v+1}\| \leq \frac{\operatorname{Tg} 2^{v-1} \varphi}{\operatorname{Tg} \varphi} \frac{\|\zeta' - \xi_v\|^2}{2\sigma_0} \frac{\alpha_0}{\alpha_0 - 1} \|Q_0\|,$$

Put here

$$\operatorname{Tg} 2^{v-1} \varphi = \frac{(\exp(2^v \varphi) - 1)^2}{\exp(2^{v+1} \varphi) - 1}$$

and express $\operatorname{Tg} \varphi$, α_0 , $\alpha_0 - 1$ in terms of φ ; we obtain further

$$\|\zeta' - \xi_{v+1}\| \leq \frac{(\exp(2^v \varphi) - 1)^2 (e^{2\varphi} + 1) (e^{\varphi/2} + e^{-\varphi/2})^2 \|Q_0\|}{(\exp(2^{v+1} \varphi) - 1) (e^{2\varphi} - 1) (e^\varphi + e^{-\varphi}) 2\sigma_0} \|\zeta' - \xi_v\|^2,$$

or, simplifying,

$$\|\zeta' - \xi_{v+1}\| \leq \frac{(\exp(2^v \varphi) - 1)^2 (e^\varphi + 1) \|Q_0\|}{(\exp(2^{v+1} \varphi) - 1) (e^\varphi - 1) 2\sigma_0} \|\zeta' - \xi_v\|^2 \quad (v \geq 1). \quad (40.8)$$

11. If we put now

$$\varepsilon_v := \frac{\|Q_0\| (e^\varphi + 1)}{2\sigma_0 (e^\varphi - 1)} (\exp(2^v \varphi) - 1) \|\zeta' - \xi_v\| \quad (v \geq 1), \quad (40.9)$$

relation (40.8) becomes $\varepsilon_{v+1} \leq \varepsilon_v^2$ and we obtain finally

$$\varepsilon_{v+1} \leq \varepsilon_1^{2^v} \quad (v \geq 1). \quad (40.10)$$

As to ε_1 , we obtain, as $\|\zeta' - \xi_1\| \leq \rho_0 = e^{-\varphi} \|h_0\|$,

$$\begin{aligned}\varepsilon_1 &= \frac{\|Q_0\| e^\varphi + 1}{2\sigma_0 e^\varphi - 1} (e^{2\varphi} - 1) |\zeta' - \xi_1| \leq \frac{\|Q_0\|}{2\sigma_0} (e^\varphi + 1)^2 e^{-\varphi} \|h_0\| \\ &\leq \frac{\|f_0\| \|Q_0\|^2}{2\sigma_0} (e^{\varphi/2} + e^{-\varphi/2})^2 = \frac{\|f_0\| \|Q_0\|^2 2\alpha_0}{2\sigma_0} = 1,\end{aligned}$$

and it follows from (40.10) that $\varepsilon_v \leq 1$ ($v \geq 1$) and therefore by (40.9)

$$\|\zeta' - \xi_v\| \leq \frac{2\sigma_0}{\|Q_0\| (e^\varphi + 1)} \frac{e^\varphi - 1}{\exp(2^v \varphi) - 1}, \quad (40.11)$$

where for $\varphi = 0$ the last right-hand fraction is to be replaced with 2^{-v} . But here the right-hand expression tends to 0 as $v \rightarrow \infty$ and we obtain $\|\zeta' - \zeta\| = 0$. We see that ζ' cannot be different from ζ .

Theorem 40.2 is proved.

41

Central Existence Theorem for Finite Systems of Equations

FORMULATION OF THE CENTRAL EXISTENCE THEOREM

1. We are now going to specialize the result of Chapter 37 to the case where $X \wedge Y$ are n -dimensional real vector spaces \mathbb{R}^n .[†] A mapping $\Omega(X \rightarrow Y)$ is here a mapping of the vector $\xi = (x_1, \dots, x_n) \in X$ upon the vector $\eta = (y_1, \dots, y_n) \in Y$,

$$y_\mu = f_\mu(\xi) = f_\mu(x_1, \dots, x_n) \quad (\mu = 1, \dots, n). \quad (41.1)$$

The F -gradient of Ω is then the Jacobian matrix

$$J(\xi) := \text{grad } \Omega = \left(\frac{\partial(f_1, \dots, f_n)}{\partial(x_1, \dots, x_n)} \right)(\xi). \quad (41.2)$$

The inverse operator is then $\Lambda(\xi) = J(\xi)^{-1}$ and we obtain therefore from Theorem 37.1:

Theorem 41.1. (Central Existence Theorem) *Let $f_v(\xi) = f_v(x_1, \dots, x_n)$ ($v = 1, \dots, n$) be n real functions of the real point ξ , defined and continuous with their first derivatives in a neighborhood of a point ξ_0 . Denote by $J = (\partial(f_1, \dots, f_n)/\partial(x_1, \dots, x_n))$ the Jacobian matrix of the f_v with respect to the x_μ and put*

$$f_v(\xi) = y_v, \quad \eta = (y_1, \dots, y_n)', \quad f_v(\xi_0) = y_v^{(0)}, \quad \eta_0 = (y_1^{(0)}, \dots, y_n^{(0)})'. \quad (41.3)$$

Assume in the spaces X and Y arbitrary norms and in the spaces of linear transformations $X \rightarrow Y$ and $Y \rightarrow X$ the corresponding operator norms. Assume further that $f_\mu \in C^1$ throughout the whole neighborhood $U_R(\xi_0)$, $\det J(\xi) \neq 0$, and

$$\|J^{-1}(\xi)\| \leq \frac{R}{\|\eta_0\|} \quad (\|\xi - \xi_0\| \leq R). \quad (41.4)$$

Then $U_R(\xi_0)$ contains at least one point ζ at which all $f_\mu(\zeta)$ ($\mu = 1, \dots, n$) vanish.

[†] If we assume $X \wedge Y$ as \mathbb{C}^n , the formal discussions remain the same, but the $f_\mu(x_1, \dots, x_n)$ have to be assumed as analytic.

THE CHOICE OF NORMS

2. In (41.4) the norms $\|\eta_0\|$, $\|\xi - \xi_0\|$ are the vector norms introduced in Y and X while the norm $\|J^{-1}\|$ is the corresponding matrix norm for the transformations $Y \rightarrow X$. In practice we have to use estimates of these norms, and this may present, for the estimate of $\|J^{-1}\|$, an intricate algebraic problem.

If we want to avoid computing J^{-1} , we can use the Euclidean lengths as the norm in X and Y and correspondingly the Euclidean matricial norm for J^{-1} . Then an upper limit for $|J^{-1}(\xi)|_2$ is obtained from (23.16) as

$$|J^{-1}(\xi)|_2 \leq \frac{1}{(\sqrt{n-1})^{n-1}} \frac{|J(\xi)|_F}{|\det J(\xi)|}. \quad (41.5)$$

Of course, in this case we have to compute the Jacobian determinant or at least a lower limit for its modulus throughout $U_R(\xi_0)$.

3. If the inverse matrix $J^{-1}(\xi)$ can be computed, probably the simplest norms to use are the Minkowski norms, and we will therefore assume in X and in Y the norms

$$|\xi|_p, \quad |\eta|_{p'} \quad (1 \leq p \wedge p' \leq \infty).$$

Here again arises the problem of the computation of the corresponding matricial norm $|A|_{p',p}$. From what has been said in Chapter 23 we can use in practice either

$$p = p' = 1 \vee 2 \vee \infty \quad \text{or} \quad p = 1, \quad p' = q = \infty,$$

since in all these cases the corresponding matricial norms can be easily determined (or, for $p = p' = 2$, at least estimated by $|J^{-1}|_F$). The corresponding formulas are given in (19.11), (19.17), and (24.16).

If we want to use other systems of Minkowski norms, the corresponding matricial norms $|A|_{p',p}$ can be estimated for $p' = q$ by $\Delta_p(A)$, using (24.11), or, if $p' \neq p$, by $\Delta_{p',p}(A)$ where we can again use either (24.9) or (24.15) or (24.12).

A UNIQUENESS THEOREM

4. In practice it may be worthwhile to try out different choices of norms, each time choosing the corresponding R as small as is allowed by (41.4). If we want, however, to use these different neighborhoods simultaneously, we must be sure that we deal each time with the same solution point of the equations $f_v(\xi) = 0$. This can be secured in many cases by using

Theorem 41.2. Consider a convex set, S , of points in the n -dimensional space R_n on which all functions

$$f_v(\xi) = f_v(x_1, \dots, x_n) \quad (v = 1, \dots, n)$$

are continuous with their first derivatives. Assume that the determinant

$$K(\xi_1, \dots, \xi_n) = \begin{vmatrix} \frac{\partial f_v(\xi_v)}{\partial x_\mu} \end{vmatrix} = \begin{vmatrix} \frac{\partial f_1(\xi_1)}{\partial x_1} \dots \frac{\partial f_1(\xi_1)}{\partial x_n} \\ \vdots \\ \frac{\partial f_n(\xi_n)}{\partial x_1} \dots \frac{\partial f_n(\xi_n)}{\partial x_n} \end{vmatrix} \quad (41.6)$$

remains $\neq 0$ if ξ_1, \dots, ξ_n run independently through S . Then if for two points, ξ', ξ'' , of S we have

$$f_v(\xi') - f_v(\xi'') = 0 \quad (v = 1, \dots, n), \quad (41.7)$$

we must have $\xi' = \xi''$.

Proof. For $\xi' = (x_1', \dots, x_n')$, $\xi'' = (x_1'', \dots, x_n'')$, we have by the mean value theorem from (41.7)

$$\sum_{\mu=1}^n (x_\mu' - x_\mu'') \frac{\partial f_v}{\partial x_\mu}(\xi_v) = 0 \quad (v = 1, \dots, n),$$

where for each v the point ξ_v lies on the segment connecting ξ' and ξ'' , that is, on S . But here we have a set of n homogeneous equations with n unknowns $x_\mu' - x_\mu''$, the determinant of which is $K(\xi_1, \dots, \xi_n) \neq 0$. Therefore we have for each μ , $x_\mu' - x_\mu'' = 0$, and Theorem 41.2 is proved.

In practice, if the set S is sufficiently small, the values of the derivative $f'_v(\xi_\mu)$ in all points of S can be considered as coinciding, taking the variation of these values into the rounding-off and similar errors; and then the use of Theorem 41.2 is immediate.

EXAMPLE

5. We will illustrate the use of our theorems on an example. Take

$$f_1 = 4x_1 - x_2 - x_1 \sin x_2, \quad f_2 = x_1 + x_2 - 4 \cot x_2,$$

$$\xi_0 = (0.828, 3.849), \quad \eta_0 = (0.021089, -0.021195).$$

For $h = 10^{-3}$ we consider the neighborhood $U: |\xi - \xi_0|_\infty \leq h$, of ξ_0 , the Jacobian matrix $J(\xi)$, the matrix $K(\xi_1, \xi_2)$, and $J^{-1}(\xi)$. We obtain throughout

the whole neighborhood U

$$\begin{aligned} f'_{1x_1} &= 4.650(\pm 1), & f'_{1x_2} &= -0.370(\pm 2), \\ f'_{2x_1} &= 1, & f'_{2x_2} &= 10.47(\pm 2.5). \end{aligned}$$

Then obviously

$$K(\xi_1, \xi_2) = \begin{vmatrix} 4.650(\pm 1) & -0.370(\pm 2) \\ 1 & 10.47(\pm 2.5) \end{vmatrix} > 0,$$

so that there cannot be more than one solution in U .

J^{-1} is easily computed throughout the whole neighborhood U to be

$$J^{-1} = \frac{1}{100} \begin{pmatrix} 21.67 \pm 0.11 & -2.07 \pm 0.01 \\ -0.77 \pm 0.01 & 9.62 \pm 0.04 \end{pmatrix}.$$

6. We now obtain the following values of the norms of J^{-1} in U , choosing either $p = p' = 1 \vee \infty$ or $p' = q, p = 1 \vee 2 \vee \infty$, with an absolute error ≤ 0.002 :

$$\begin{aligned} p = p' = 1, \quad |J^{-1}|_1 &= 0.224; & p = p' = \infty, \quad |J^{-1}|_\infty &= 0.238; \\ p = p' = 2, \quad \Delta_2(J^{-1}) &= 0.238; & p = \infty, \quad p' = 1, \quad |J^{-1}|_{1,\infty} &= 0.217. \end{aligned}$$

On the other hand, we have

$$|\eta_0|_1 = 0.02228, \quad |\eta_0|_\infty = 0.02120, \quad |\eta_0|_2 = 0.02162.$$

With these values we obtain the smallest values of R compatible with (41.4), putting $\delta = 10^{-4}$:

$$\begin{aligned} p = p' = 1, \quad R &= 5.153\delta; & p = p' = \infty, \quad R &= 2.880\delta, \\ p = p' = 2, \quad R &= 3.888\delta; & p = \infty, \quad p' = 1, \quad R &= 2.628\delta. \end{aligned}$$

Omitting the case where $p = p' = \infty$, we shall consider the three neighborhoods corresponding to $R_1 := 5.153\delta, R_2 := 3.888\delta, R_\infty := 2.628\delta$:

$$U_p: |\xi - \xi_0|_p \leq R_p \quad (p = 1, 2, \infty). \quad (41.8)$$

These neighborhoods are contained in U . Therefore, our values of the norms hold in all neighborhoods (41.8) and Theorem 41.1 can be applied to any of these three neighborhoods. Since U_1 and U_∞ are also contained in $|\xi - \xi_0|_\infty \leq h/2$, we see that the values of the components of ξ_0 can be considered as the correctly rounded-off values of the coordinates of the solution in question. On the other hand, it follows that the solution of the equations $f_1 = 0, f_2 = 0$ lies in the product of the three neighborhoods (41.8). We represent the upper right quarters of these three neighborhoods (41.8) in Fig. 12. We see that the solution in question lies in the product $U_1 \cap U_\infty$. This is the pentagon

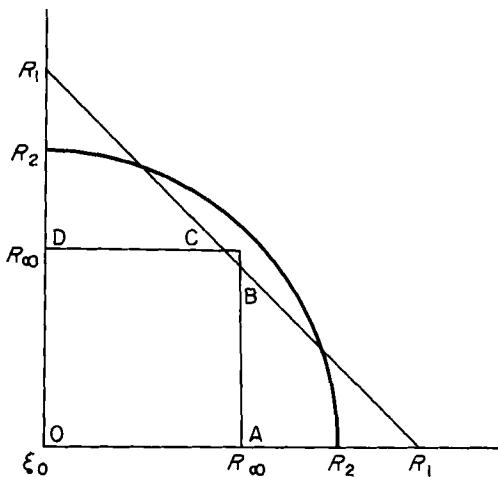


FIGURE 12

OABCDO together with the three pentagons obtained from it by symmetry with respect to OA and OD.

7. Another way to reduce the computation of J^{-1} to its computation *only in one point* consists in the use of formula (24.27). In our example, if J^{-1} is computed at the point ξ_0 , we see easily that if we denote the corresponding J by J_0 , for another point of U the difference $K - J_0$ is majorized by $h \begin{pmatrix} 1 & 1.4 \\ 0 & 0.5 \end{pmatrix}$. Therefore we have

$$|K - J_0|_\infty \leq 20.05h, \quad |K - J_0|_1 \leq 22.9h, \quad |K - J_0|_2 \leq 20.25h,$$

and it follows again, using Theorem 24.1, that K does not vanish and, using the formula (24.29), that

$$|J^{-1} - J_0^{-1}|_\infty < 1.172h, \quad |J^{-1} - J_0^{-1}|_1 < 2.43h, \quad |J^{-1} - J_0^{-1}|_2 < 1.092h.$$

42

Newton–Raphson Iteration for Finite Systems of Equations

FORMULATION OF THE THEOREM

1. In order to specialize Theorem 38.1 to the case of a finite system of equations we proceed as in Chapter 41. We consider two normed n -dimensional vector spaces X, Y and a mapping $\Omega(X \rightarrow Y)$ given by

$$\eta = \Omega(\xi), \quad \eta := (y_1, \dots, y_n), \quad \xi := (x_1, \dots, x_n)$$

where

$$y_\mu = f_\mu(\xi) = f_\mu(x_1, \dots, x_n) \quad (\mu = 1, \dots, n). \quad (42.1)$$

The functions f_μ are assumed to be continuously differentiable at a point $\xi_0 \in X$, and such that their Jacobian matrix at ξ_0 , J_0 , has a nonvanishing determinant so that J_0^{-1} exists. Put $q_0 := \|J_0^{-1}\|$,

$$\eta_0 := \Omega(\xi_0) = (f_1(\xi_0), \dots, f_n(\xi_0))$$

and assume that $\eta_0 \neq 0$. Put further

$$h_0 := -J_0^{-1}\eta_0, \quad \xi_1 := \xi_0 + h_0.$$

Consider now a fixed $\alpha_0 \geq 2$ and the corresponding $\varphi \geq 0$ defined by (38.4). Put

$$\rho_0 := e^{-\varphi} \|h_0\|$$

and consider the ball K_0 around ξ_1 , $\|\xi - \xi_1\| \leq \rho_0$.

We denote now by C_0 the set consisting of the interval $S_0 := \langle \xi_0, \xi_1 \rangle$ and of the ball K_0 , and assume that the functions $f_\mu(\xi)$ are continuously differentiable on the set C_0 .

2. Then we have the theorem:

Theorem 42.1. Denote under the above hypotheses the Jacobian matrix of the f_μ at the general point ξ of C_0 by $J(\xi)$ and assume that $J(\xi)$ satisfies the following Lipschitz condition on S_0 and on K_0 :

$$\|J(\xi') - J(\xi)\| \leq \frac{\|\xi' - \xi\|}{\alpha_0 \|\eta_0\| q_0^2} \quad (\xi' \wedge \xi \in S_0 \vee K_0). \quad (42.2)$$

Then if we replace in our hypotheses ξ_0 by ξ_1 , all these hypotheses remain satisfied and the whole construction can be repeated replacing $\xi_0, h_0 \alpha_0, \varphi, \rho_0, \eta_0, \xi_1, q_0$ resp. with $\xi_1, h_1, \alpha_1, \varphi/2, \rho_1, \eta_1, \xi_2, q_1$ where α_1 is deduced from α_0 by (38.6).

The above construction can be repeated indefinitely, unless a certain η_v becomes = 0, and the sequence ξ_v obtained in this way tends to a point $\zeta \in K_0$ at which all $f_\mu(\zeta)$ ($\mu = 1, \dots, n$) vanish. Then the assertions of Theorems 38.1 and 38.2 hold, replacing P_v with J_v and Q_v with J_v^{-1} .

THE CHOICE OF THE NORMS

3. In verifying inequality (42.2) the convenient choice of norms is very essential. Theoretically it would be sufficient to consider in the space of linear operators ($X \rightarrow Y$) the gradient of $J(\xi)$ with respect to the parameter ξ and use again Theorem 35.1. However, then the convenient representation of this second-order gradient is in most cases very cumbersome if at all possible. In practice one will probably use in the most cases the Minkowski norms $|\xi|_p$, $|\eta|_{p'}$ in X, Y . The induced operator norm is then $|A|_{p, p'}$, which cannot be expressed algebraically except in some few cases. In all these cases, however, this expression can be written as $\Delta_{p, p'}$ for A or A' , so that our problem consists in obtaining convenient limits for $\Delta_{p, p'}(J(\xi') - J(\xi))$, where, in a certain case, $J(\xi') - J(\xi)$ is to be replaced by the transpose matrix.

If we now assume that the functions f_μ have continuous second derivatives in C_0 , a solution of our problem is contained in the following theorem.

Theorem 42.2. Consider an $n \times n$ matrix $A = (a_{\mu\nu}(\xi))$, where $a_{\mu\nu}$ have continuous second derivatives with respect to all x_v in C_0 . Put

$$\hat{\Delta}_{p, p'}(\xi) := \left(\sum_{\mu=1}^n \left[\left(\sum_{v, \lambda=1}^n \left| \frac{\partial a_{\mu v}}{\partial x_\lambda} \right|^q \right)^{1/q} \right]^{p'} \right)^{1/p'}, \quad (42.3)$$

$$\max_{\xi \in C_0} \hat{\Delta}_{p, p'}(\xi) =: \Delta_{p, p'}^*. \quad (42.4)$$

Then we have if $\xi \wedge \xi'$ lie either on S_0 or on K_0 :

$$\Delta_{p, p'}(A(\xi') - A(\xi)) \leq \Delta_{p, p'}^* |\xi' - \xi|_p \quad (\xi' \wedge \xi \in S_0 \vee K_0). \quad (42.5)$$

In order to prove this theorem we first prove two lemmas.

4. Lemma 42.1. Assume that the components x_v ($v = 1, \dots, n$) of the vector ξ are functions of t , differentiable at $t = t_0$. If $1 < p < \infty$, $|\xi|_p$ is differentiable at $t = t_0$ and we have there

$$\left| \frac{d}{dt} |\xi|_p \right| \leq |\xi'_t|_p. \quad (42.6)$$

Relation (42.6) holds also for $p = 1$ and $p = \infty$ if the derivatives in the assumptions and assertions are all replaced with the right-handed or all with the left-handed derivatives.

Proof. Let the prime denote the derivative at t_0 . Assume first $1 < p < \infty$. Then, if $x_v \neq 0$,

$$\begin{aligned} (|x_v|^p)' &= ((x_v \bar{x}_v)^{p/2})' = \frac{p}{2} |x_v|^{p-2} (x_v \bar{x}_v' + x_v' \bar{x}_v), \\ |(|x_v|^p)'| &\leq p |x_v|^{p-1} |x_v'|, \end{aligned} \quad (42.7)$$

and this formula remains valid if $x_v(t_0) = 0$. Therefore

$$\left| \frac{d}{dt} |\xi|_p \right| \leq \frac{1}{p} \left(\sum_{v=1}^n |x_v|^p \right)^{(1/p)-1} p \sum_{v=1}^n |x_v|^{p-1} |x_v'|. \quad (42.8)$$

But, by Hölder's inequality,

$$\sum_{v=1}^n |x_v|^{p-1} |x_v'| \leq \left(\sum_{v=1}^n |x_v|^{(p-1)q} \right)^{1/q} \left(\sum_{v=1}^n |x_v'|^p \right)^{1/p}$$

and introducing this into (42.8), we obtain, since $(p-1)q = p$, the assertion if $1 < p < \infty$.

On the other hand, the above deduction of (42.7) remains valid also for $p = 1$ if $x_v(t_0) \neq 0$. And if $x_v(t_0) = 0$ we have, denoting by D_ϵ the right-handed or left-handed derivative at t_0 , according as $\epsilon = +$ or $-$, with $t \downarrow t_0$ or $t \uparrow t_0$:

$$\begin{aligned} D_\epsilon |x_v| &= \lim \frac{|x_v(t)|}{t - t_0} = \epsilon |D_\epsilon x_v|, \\ |D_\epsilon |x_v|| &= |D_\epsilon x_v|. \end{aligned} \quad (42.9)$$

From (42.7) for $p = 1$ and (42.9), relation (42.6) follows immediately for $p = 1$. And for $p = \infty$ we have, if $\text{Max}_v |x_v(t_0)| = |x_k(t_0)|$,

$$|D_\epsilon |\xi|_\infty| = |D_\epsilon \text{Max} |x_v|| = |D_\epsilon |x_k|| = |D_\epsilon x_k| \leq \text{Max}_v |D_\epsilon x_v| = |D_\epsilon \xi|_\infty.$$

Observe that in relation (42.6) the equality sign holds for every p if we put $\xi = \gamma e^t$ with a vector γ independent of t .

5. Lemma 42.2. *Assume that the elements $a_{\mu\nu}$ of the matrix A are functions of t differentiable at t_0 . Assume $0 \leq p \wedge p' \leq 1$. Then $(d/dt) \Delta_{p, p'}(A)$ exists at t_0 and we have there*

$$\left| \frac{d}{dt} \Delta_{p, p'}(A) \right| \leq \Delta_{p, p'} \left(\frac{d}{dt} A \right), \quad (42.10)$$

where, however, if one of the numbers p, p' is 1 or ∞ , in the assumptions and assertions the derivatives at t_0 are to be all replaced with the right-handed derivatives or all with the left-handed derivatives.

Proof. Denote by R_μ ($\mu = 1, \dots, n$) the vector consisting of the elements of the μ th row of A , $(a_{\mu 1}, \dots, a_{\mu n})$, and by R the vector $(|R_1|_q, \dots, |R_n|_q)$, so that $\Delta_{p, p'}(A) = |R|_p$. Then it follows by (24.10), applying Lemma 42.1 repeatedly and denoting by primes the derivation at t_0 , that

$$\left| \frac{d}{dt} \Delta_{p, p'}(A) \right| = \left| |R|'_{p'} \right| \leqslant \left| (|R_1|'_q, \dots, |R_n|'_q)' \right|_{p'} \leqslant \left| (|R_1|'_q, \dots, |R_n|'_q) \right|_{p'},$$

and this is $\Delta_{p, p'} \frac{d}{dt} A$.

6. In the hypotheses of Theorem 42.1, put

$$\beta := \xi' - \xi$$

where $\xi \wedge \xi'$ lie either on S_0 or on K_0 ; then, for $0 \leq t \leq 1$, $\xi + t\beta$ also lies either on S_0 or on K_0 . Further, put

$$W(\xi, t) := A(\xi + t\beta) - A(\xi) \quad (0 \leq t \leq 1). \quad (42.11)$$

Then, if we apply to $W(\xi, t)$ Lemma 42.2, it follows that

$$\left| \frac{d}{dt} \Delta_{p, p'}(W(\xi, t)) \right| \leq \Delta_{p, p'} \left(\frac{d}{dt} W(\xi, t) \right). \quad (42.12)$$

But for the general element of the matrix $(d/dt)W(\xi, t)$ we have, putting $\beta = (b_1, \dots, b_n)$,

$$\frac{\partial a_{\mu\nu}}{\partial t}(\xi + t\beta) = \sum_{\lambda=1}^n b_\lambda \frac{\partial a_{\mu\nu}}{\partial x_\lambda}(\xi + t\beta)$$

and therefore, applying Hölder's inequality, if $q < \infty$,

$$\left| \frac{\partial a_{\mu\nu}}{\partial t}(\xi + t\beta) \right|^q \leq \sum_{\lambda=1}^n \left| \frac{\partial a_{\mu\nu}}{\partial x_\lambda} \right|^q |\beta|_p^q.$$

Introducing this into the right-hand expression in (42.12), it follows that

$$\left| \frac{d}{dt} \Delta_{p, p'}(W(\xi, t)) \right| \leq \left(\sum_{\mu=1}^n \left[\left(\sum_{\nu, \lambda=1}^n \left| \frac{\partial a_{\mu\nu}}{\partial x_\lambda} \right|^q \right)^{1/q} \right]^{p'} \right)^{1/p'} |\xi' - \xi|_p.$$

This formula also holds for $q = \infty$, as is immediately verified by the argument which was used in Section 4.

It follows now that

$$\left| \frac{d}{dt} \Delta_{p, p'}(W(\xi, t)) \right| \leq \hat{\Delta}_{p, p'}(\xi + t\beta) |\xi' - \xi|_p$$

and therefore, by the mean value theorem of differential calculus, since $W(\xi, 0) = 0$,

$$\Delta_{p, p'}(W(\xi, 1)) \equiv \Delta_{p, p'}(A(\xi') - A(\xi)) \leq \hat{\Delta}_{p, p'}(\xi + \theta\beta) |\xi' - \xi| \quad (0 \leq \theta \leq 1).$$

Using (42.4) the assertion (42.5) of Theorem 42.2 follows.

7. If we now replace A with $J(\xi)$ in Theorem 42.2, we obtain

$$|J(\xi') - J(\xi)|_{p, p'} \leq \max_{\xi \in C_0} \left(\sum_{\mu=1}^n \left[\left(\sum_{v, \lambda=1}^n \left| \frac{\partial^2 f_\mu(\xi)}{\partial x_v \partial x_\lambda} \right|^q \right)^{1/q} \right]^{p'} \right)^{1/p'} |\xi' - \xi|_p \quad (42.13)$$

and in the case of $|A|_1 = |A'|_\infty$, replacing J by the transpose matrix,

$$|J(\xi') - J(\xi)|_1 \leq \max_{\xi \in C_0} \max_v \sum_{\mu, \lambda=1}^n \left| \frac{\partial^2 f_\mu(\xi)}{\partial x_v \partial x_\lambda} \right| |\xi' - \xi|_1. \quad (42.14)$$

Then condition (42.2) is satisfied if

$$\max_{\xi \in C_0} \left(\sum_{\mu=1}^n \left[\left(\sum_{v, \lambda=1}^n \left| \frac{\partial^2 f_\mu(\xi)}{\partial x_v \partial x_\lambda} \right|^q \right)^{1/q} \right]^{p'} \right)^{1/p'} \leq \frac{1}{\alpha_0 |\eta_0|_p q_0^2} \quad (42.15)$$

and, in the case of the norm $|A|_1$,

$$\max_{\xi \in C_0} \max_v \sum_{\mu, \lambda=1}^n \left| \frac{\partial^2 f_\mu(\xi)}{\partial x_v \partial x_\lambda} \right| \leq \frac{1}{\alpha_0 |\eta_0|_1 q_0^2}. \quad (42.16)$$

APPLICATION TO COMPLEX FUNCTIONS OF A COMPLEX VARIABLE

8. From Theorems 38.1 and 38.2, Theorem 7.2 follows immediately. Indeed, under the conditions of Theorem 7.2, Q_v is to be replaced with $1/f'(z_v)$ and M with $1/\sigma_0$. Then formula (7.15) follows immediately from the first inequality (38.25). As to (7.16), it follows from (38.20) if for $\|h_v\|$ there we introduce its bound from (7.15) and drop the factor $\exp(-2^v \varphi)$.

In particular, if a domain is known which certainly does not contain any zeros of $f(z)$, in many cases Theorem 38.3 can be used.

Appendices

A

Continuity of the Roots of Algebraic Equations

1. In solving equations containing numerical parameters, the values of these parameters must usually be rounded off, i.e., replaced by approximate values. About the influence of this procedure on the computed values of the roots little can be said in the general case. If the equation with the unknown z and parameter t is $f(z, t) = 0$, we have of course $dz/dt = -f'_t/f'_z$. This relation can, however, be used only after the values of z and f'_z have been obtained or at least estimated with sufficient precision.

General results are obtained only in the case of algebraic equations. These results are very weak, just because they are very general.

2. We consider two polynomials

$$\begin{aligned} f(x) &= a_0 x^n + \cdots + a_n, & a_0 &= 1, \\ g(x) &= b_0 x^n + \cdots + b_n, & b_0 &= 1. \end{aligned} \tag{A.1}$$

Let the n roots of $f(x)$ be x_1, \dots, x_n , those of $g(x)$, y_1, \dots, y_n . Our problem is to obtain estimates for the differences between x_v and y_v in terms of the expressions $|b_v - a_v|$. Put

$$\gamma = 2\Gamma, \quad \Gamma := \operatorname{Max}_{v>0}(|a_v|^{1/v}, |b_v|^{1/v}). \tag{A.2}$$

Introduce now the expression

$$\varepsilon = \sqrt[n]{\sum_{v=1}^n |b_v - a_v|^{\gamma^{n-v}}}. \tag{A.3}$$

Theorem. *The roots x_v and y_v can be ordered in such a way that we have*

$$|x_v - y_v| \leq (2n-1) \varepsilon \quad (v = 1, \dots, n). \tag{A.4}$$

3. Proof. We can obviously assume $n > 1$.

There is one particular case in which (A.4) follows immediately. Assume that there exists a zero y of $g(x)$ such that we have for all x_v ,

$$|x_v - y| \leq \varepsilon,$$

and further that there exists a zero x of $f(x)$ such that we have for all y_v

$$|y_v - x| \leq \varepsilon.$$

Then obviously, since for all y_v

$$|y - y_v| < 2\varepsilon,$$

it follows that for all v

$$|x_v - y_v| \leq 3\varepsilon \leq (2n-1)\varepsilon.$$

We see that in this case (A.4) holds independently of the ordering of the zeros.

Since, on the other hand, we can always interchange $f(x)$ with $g(x)$, we can from now on assume that there does not exist any zero y of $g(x)$ such that all x_v lie in the closed ε -neighborhood of this zero.

4. It is proved in elementary algebra that all $|x_v|$ are $\leq 2 \operatorname{Max}_{v>0} |a_v|^{1/v}$. It follows therefore that in our case all $|x_v|$, $|y_v|$ are $\leq \gamma$.

Consider now the family of polynomials

$$g_t(x) := f(x) + t(g(x) - f(x)) \quad (0 \leq t \leq 1). \quad (\text{A.5})$$

Obviously $g_0(x) = f(x)$, $g_1(x) = g(x)$. Denote the zeros of $g_t(x)$ by $y_v^{(t)}$ ($v = 1, \dots, n$).

I say now that we have

$$|y_v^{(t)}| \leq \gamma \quad (0 \leq t \leq 1, \quad v = 1, \dots, n). \quad (\text{A.6})$$

Indeed,

$$\begin{aligned} |(1-t)a_v + tb_v|^{1/v} &\leq ((1-t)|a_v| + t|b_v|)^{1/v} \\ &\leq (\operatorname{Max}(|a_v|, |b_v|))^{1/v} = \frac{\gamma}{2} \quad (v = 1, \dots, n). \end{aligned}$$

5. We call two zeros x_v and x_μ of $f(x)$ *neighbors* if their distance is $\leq 2\varepsilon$. Two zeros x_v , x_μ of $f(x)$ will be called *connected* if there is a sequence of zeros of $f(x)$ beginning with x_v and ending with x_μ : $x_v, x_{\lambda_1}, x_{\lambda_2}, \dots, x_{\lambda_k}, x_\mu$, such that each zero of the sequence except x_v is a neighbor of the immediately preceding one.

Thus all n zeros of $f(x)$ can be decomposed into a number of groups g_1, \dots, g_k such that all zeros of the same group are connected, while two zeros belonging to different groups are never connected. The sum of the open ε -neighborhoods of all zeros contained in g_κ may be denoted by G_κ and the boundary of the open set G_κ by B_κ ; B_κ consists of a finite number of circular arcs. G_κ and G_λ have no zeros x_v in common if $\kappa \neq \lambda$.

6. We will now prove that none of the $y_v^{(t)}$ ($0 < t \leq 1$) lies on $B_1 \cup \dots \cup B_k$. Indeed, otherwise there would exist a t , $0 < t \leq 1$, a $y_v^{(t)} =: y_0^{(t)}$, and an $x_\mu =: x_0$ such that

$$|y_0^{(t)} - x_0| = \varepsilon.$$

On the other hand, we have

$$\prod_{v=1}^n (y_0^{(t)} - x_v) = f(y_0^{(t)}) - g_t(y_0^{(t)}) = t \sum_{v=0}^n (a_v - b_v) y_0^{(t)^{n-v}}.$$

By (A.6) and (A.3) it follows that

$$\prod_{v=1}^n |y_0^{(t)} - x_v| \leq t \varepsilon^n. \quad (\text{A.7})$$

From this formula, however, it follows that for $t < 1$ we have for one of the x_v : $|y_0^{(t)} - x_v| < \varepsilon$, so that $y_0^{(t)}$ lies in the corresponding G_κ and cannot lie on one of the B_κ . For $t = 1$ we have

$$\prod_{v=1}^n |y_0^{(1)} - x_v| \leq \varepsilon^n$$

and therefore, since $y_0^{(1)}$ is assumed not to lie in one of the G_κ , $|y_0^{(1)} - x_v| = \varepsilon$ ($v = 1, \dots, n$). This contradicts, however, the assumption made at the end of Section 3.

7. Before giving the proof of our theorem, we prove a lemma from the theory of functions of a complex variable.

Lemma. *Let B be a closed region in the x plane, the boundary of which consists of a finite number of regular arcs; let the functions $f(x), h(x)$ be regular on B . Assume that for no value of the real parameter t , running through the interval $a \leq t \leq b$, the function $f(x) + th(x)$ becomes $=0$ on the boundary of B . Then the number $N(t)$ of the zeros of $f(x) + th(x)$ inside B is independent of t for $a \leq t \leq b$.*

Proof of the Lemma. Let t_0 be a value from $\langle a, b \rangle$. The modulus of $u(x) := f(x) + t_0 h(x)$ has a positive lower limit p on the boundary of B . Then if $|\delta|$ is sufficiently small, the function $v(x) := \delta h(x)$ has its modulus $< p$ everywhere on the boundary of B . Therefore, everywhere on the boundary of B we have $|u| > |v|$ and by an important theorem due to Rouché, $u + v$ has the same numbers of roots inside B as the function $u(x)$.

It follows now that $N(t)$ is a continuous function for any t_0 from $\langle a, b \rangle$, since it is constant in a neighborhood of any t_0 belonging to $\langle a, b \rangle$; and since $N(t)$ has only integer values, it must be constant throughout $\langle a, b \rangle$. Our lemma is proved.

8. We consider now the group g_1 and assume that it contains exactly ρ zeros of $f(x)$, counted, of course, according to their multiplicity.

But it follows by our lemma that the function (A.5) has a constant number of zeros inside G_1 for $0 \leq t \leq 1$. Since for $t = 0$ this number is ρ , we see that $g(x)$ also has exactly ρ zeros y_v inside G_1 . But then, since the distance of any of these y_v from any of the x_v in g_1 is $\leq (2\rho - 1)\varepsilon \leq (2n - 1)\varepsilon$, relation (A.4) holds for the x_v and the y_v in G_1 . The theorem now follows immediately by applying this argument to each group g_k .

9. In order to use (A.4), we must of course find a convenient estimate of ε . Putting

$$\delta = \operatorname{Max}_v |b_v - a_v|, \quad (\text{A.8})$$

we have from (A.3),

$$\varepsilon^n \leq \delta \sum_{v=0}^{n-1} \gamma^v.$$

On the other hand, we have for any $u \geq 0$ the relation[†]

$$\sum_{v=0}^{n-1} u^v \leq \operatorname{Max}(1, u^n) \operatorname{Min}\left(n, \frac{1}{|1-u|}\right). \quad (\text{A.9})$$

We obtain therefore

$$\varepsilon \leq \delta^{1/n} \operatorname{Max}(1, \gamma) \sqrt[n]{\operatorname{Min}\left(n, \frac{1}{|1-\gamma|}\right)}. \quad (\text{A.10})$$

Inequality (A.10) may be inconvenient because all differences $a_v - b_v$ are used with the same *weight*. Another estimate is obtained by assuming

$$|b_v - a_v| \leq \sigma \Gamma^v \quad (v = 1, \dots, n); \quad (\text{A.11})$$

then we have from (A.3)

$$\begin{aligned} \varepsilon^n &\leq \sigma \sum_{v=1}^n \Gamma^v (2\Gamma)^{n-v} = \sigma \Gamma^n (1 + 2 + \dots + 2^{n-1}), \\ \varepsilon &\leq \gamma \sigma^{1/n}. \end{aligned} \quad (\text{A.12})$$

10. A very instructive example is given by the equation

$$(z-1)^4 - (7 \cdot 10^{-4}z)^2 = z^4 - 4z^3 + (6 - 49 \cdot 10^{-8})z^2 - 4z + 1 = 0, \quad (\text{A.13})$$

the four roots of which are

$$y_1 = 1.02681, \quad y_2 = 0.97389, \quad y_{3,4} = 0.99965 \pm i0.026455.$$

[†] Equation (A.9) is immediately verified if we treat the cases $u \leq 1$ and $u > 1$ separately and prove the inequality with n directly, and that with $1/|1-u|$ by replacing the left-hand expression by $(1-u^n)/(1-u)$.

Take the polynomial (A.13) as $g(z)$, and $(z-1)^4$ as $f(z)$. Then all x_v are 1 and $\text{Max}_v |y_v - x_v| = 0.02681 = y_1 - 1$, while $\varepsilon^4 = 49 \cdot 10^{-8} y_1^2 = (y_1 - 1)^4$, $\varepsilon = y_1 - 1$.

11. We show finally in an example that it is impossible to deduce for $n = 3$ the relation (A.4) with the coefficient 1 instead of 5. Take

$$f(z) = z^3 + \frac{3}{2}\sqrt[3]{2}z^2 - 1, \quad g(z) = z^3 + \frac{3}{2}\sqrt[3]{2}z^2.$$

Here we have

$$x_1 = -\sqrt[3]{2}, \quad x_2 = -\sqrt[3]{2}, \quad x_3 = \frac{1}{\sqrt[3]{4}} = 0.6296,$$

$$y_1 = 0, \quad y_2 = 0, \quad y_3 = -\frac{3}{2}\sqrt[3]{2} = -1.89,$$

and it is obviously impossible to reorder y_1, y_2, y_3 in such a way that we get $|y_v - x_v| \leq \varepsilon = 1$.

B

Relative Continuity of the Roots of Algebraic Equations

1. In Appendix A, the continuity problem was discussed from the point of view of absolute errors, while very often in a computation in which numbers of very different order of magnitude are used the restriction to *relative errors* cannot be avoided. This is in particular the case when, for all numbers occurring in the computation, only a fixed number of significant digits is given. It is of some importance to discuss the question whether, if the coefficients of an algebraic equation are given with a certain number of significant digits, then irrespective of the absolute magnitude of the coefficients, a certain number of significant digits of the roots can be guaranteed. The answer is indeed in the affirmative.

2. We shall prove the following theorem.

Theorem. Consider two polynomials

$$f(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n, \quad g(z) = b_0 z^n + b_1 z^{n-1} + \cdots + b_n \quad (\text{B.1})$$

and assume that $a_0 a_n \neq 0$, i.e., that $f(z)$ has n finite zeros x_v ($v = 1, \dots, n$), all $\neq 0$. Assume further that for a certain positive τ with

$$4n\tau^{1/n} \leq 1 \quad (\text{B.2})$$

we have

$$|b_v - a_v| \leq \tau |a_v| \quad (v = 0, 1, \dots, n). \quad (\text{B.3})$$

Then the n zeros y_1, \dots, y_n of $g(z)$ can be ordered in such a way that we have

$$\left| \frac{y_v}{x_v} - 1 \right| < 8n\tau^{1/n} \quad (v = 1, \dots, n). \quad (\text{B.4})$$

3. Proof. Put $|x_v| = r_v$ ($v = 1, \dots, n$). We can assume that

$$r_1 \geq r_2 \geq \cdots \geq r_n, \quad |y_1| \geq |y_2| \geq \cdots \geq |y_n|.$$

Consider the polynomial

$$F(z) = |a_0| \prod_{v=1}^n (z + r_v) = S_0 z^n + \cdots + S_n. \quad (\text{B.5})$$

We have obviously

$$S_0 = |a_0|, \quad S_v \geq |a_v| \quad (v = 1, \dots, n). \quad (\text{B.6})$$

It follows now from (B.3) and (B.6) that

$$|b_v - a_v| \leq \tau S_v \quad (v = 0, 1, \dots, n). \quad (\text{B.7})$$

It is important for later discussion that in the proof of the above theorem we use the relations (B.7) instead of the relations (B.3).

4. Let y_0 be an arbitrary zero of $g(z)$ and denote by x_0 one of the zeros x_v , for which we have

$$\min_v \left| 1 - \frac{y_0}{x_v} \right| = \left| 1 - \frac{y_0}{x_0} \right|. \quad (\text{B.8})$$

Then we have by (B.7)

$$|f(y_0)| = |f(y_0) - g(y_0)| \leq \tau F(|y_0|),$$

and therefore if we put $|y_0| = r$, since $S_0 = |a_0|$,

$$\prod_{v=1}^n |y_0 - x_v| \leq \tau \prod_{v=1}^n (r + r_v). \quad (\text{B.9})$$

5. Consider now a constant q satisfying the inequality

$$0 < q < \frac{1 - \tau^{1/n}}{1 + \tau^{1/n}}. \quad (\text{B.10})$$

The right-hand inequality is equivalent to

$$\frac{1+q}{1-q} \tau^{1/n} < 1. \quad (\text{B.11})$$

We decompose now the set of the $r_v = |x_v|$ into three classes.

In the *first class*, we take all r_λ , $\lambda = 1, \dots, l$, which are $\geq r/q$. If there are no such r_λ , we take $l = 0$. For an r_λ from the first class and the corresponding x_λ , we have obviously

$$\left| \frac{r + r_\lambda}{y_0 - x_\lambda} \right| \leq \frac{r + r_\lambda}{r_\lambda - r} = \frac{1 + r/r_\lambda}{1 - r/r_\lambda} \leq \frac{1+q}{1-q} \quad (\lambda = 1, \dots, l). \quad (\text{B.12})$$

In the *second class* we take all r_κ ($\kappa = k+1, \dots, n$), which are $\leq rq$. If there are no such r_κ , then we take $k = n$. For an r_κ from the second class, we have obviously

$$\left| \frac{r + r_\kappa}{y_0 - x_\kappa} \right| \leq \frac{r + r_\kappa}{r - r_\kappa} = \frac{1 + r_\kappa/r}{1 - r_\kappa/r} \leq \frac{1+q}{1-q} \quad (\kappa = k+1, \dots, n). \quad (\text{B.13})$$

In the *third class*, we take finally all r_σ ($\sigma = l+1, \dots, k$) which lie *strictly between* r/q and qr . If the number $s = k-l$ of r_ν from the third class is 0, we must have $k = l$.

6. We divide both sides of (B.9) by the product of those $|y_0 - x_\nu|$ which correspond to the r_ν of the first two classes. Then we obtain, using (B.12) and (B.13),

$$\prod_{\sigma=l+1}^k |y_0 - x_\sigma| \leq \tau \prod_{\sigma=l+1}^k (r + r_\sigma) \left(\frac{1+q}{1-q} \right)^{n-s}$$

and dividing this by the product of the r_σ , $\sigma = l+1, \dots, k$, we have

$$\sum_{\sigma=l+1}^k \left| 1 - \frac{y_0}{x_\sigma} \right| \leq \tau \prod_{\sigma=l+1}^k \left(1 + \frac{r}{r_\sigma} \right) \left(\frac{1+q}{1-q} \right)^{n-s} \quad (\text{B.14})$$

But here we have for any r_σ from the third class

$$1 + \frac{r}{r_\sigma} < 1 + \frac{1}{q} = \frac{1+q}{q}.$$

Using (B.8) it follows now from (B.14) that

$$\left| 1 - \frac{y_0}{x_0} \right|^s \leq \left(\frac{1-q}{q} \right)^s \tau \left(\frac{1+q}{1-q} \right)^n. \quad (\text{B.15})$$

Inequality (B.15) would give us for $s = 0$: $\tau[(1+q)/(1-q)]^n \geq 1$, which contradicts (B.11). We see therefore that $s > 0$ and it follows now from (B.15) and (B.11) that

$$\left| 1 - \frac{y_0}{x_0} \right|^s \leq \left(\frac{1-q}{q} \right)^s, \quad \left| 1 - \frac{y_0}{x_0} \right| \leq \frac{1-q}{q}.$$

7. The last relation holds for all q satisfying (B.10). If now q tends to $(1-\tau^{1/n})/(1+\tau^{1/n})$, we obtain finally by (B.2)

$$\left| 1 - \frac{y_0}{x_0} \right| \leq \frac{2\tau^{1/n}}{1-\tau^{1/n}} < 1. \quad (\text{B.16})$$

We denote now by ε an arbitrary number which satisfies

$$\frac{2\tau^{1/n}}{1-\tau^{1/n}} < \varepsilon < 1 \quad (\text{B.17})$$

and put

$$\varepsilon r_\nu = \varepsilon_\nu. \quad (\text{B.18})$$

Let U_v be the inside of the circle around x_v with the radius ε_v ($v = 1, \dots, n$). Then it follows from (B.16) that each zero y_μ of $g(z)$ is contained in one of the *open* neighborhoods U_v of the x_v .

8. We proceed now as in Appendix A. We call two zeros x_v, x_μ *neighbors* if we have

$$|x_v - x_\mu| < \varepsilon_v + \varepsilon_\mu. \quad (\text{B.19})$$

From (B.19) and (B.18) we have

$$x_\mu = \frac{1 + \theta\varepsilon}{1 + \theta'\varepsilon} x_v, \quad |\theta|, |\theta'| < 1. \quad (\text{B.20})$$

Two zeros x_{v_1}, x_{v_l} will be called *connected* if there exists a sequence $x_{v_1}, x_{v_2}, \dots, x_{v_l}$ of zeros of $f(z)$, such that each zero of the sequence is a neighbor of the adjoining zeros. Without loss of generality, we may assume $l \leq n$. Using (B.20) repeatedly, we have obviously

$$x_{v_l} = x_{v_1} \sum_{\lambda=1}^{l-1} \frac{1 + \theta_\lambda \varepsilon}{1 + \theta' \varepsilon}, \quad |\theta_\lambda| \wedge |\theta'_\lambda| < 1, \quad (\text{B.21})$$

and therefore *a fortiori*

$$x_{v_l} = x_{v_1} \sum_{\lambda=1}^{n-1} \frac{1 + \theta_\lambda \varepsilon}{1 + \theta' \varepsilon}, \quad |\theta_\lambda| \wedge |\theta'_\lambda| < 1, \quad (\text{B.22})$$

since the $\theta_\lambda, \theta'_\lambda$ with $\lambda \geq l$ can be taken as 0.

9. We decompose all x_v into groups g_1, \dots, g_k , such that two x_v from the same group are always *connected*, while two x_v of different groups are never connected. For each group we form the sum set of the corresponding neighborhoods U_v , $G_\kappa = \bigcup_{x_v \in g_\kappa} U_v$. The boundary B_κ of each G_κ consists of a finite number of circular arcs, and each B_κ has a *positive* distance of all G_λ with $\lambda \neq \kappa$. It follows now from (B.16) and (B.17) that each y_v lies in one of the open sets G_κ and that therefore $g(z)$ is different from zero on all B_κ .

10. Consider now the set G_1 . If in our result (B.16) we replace the polynomial $g(z)$ by the polynomial

$$f(z) + t[g(z) - f(z)] \quad (0 \leq t \leq 1), \quad (\text{B.23})$$

then condition (B.7) remains satisfied, replacing τ by $t\tau$. It follows therefore that no polynomial of the set (B.23) has a zero on B_1 . By the lemma in Appendix A, Section 5 we see therefore that the number of zeros of (B.23) in G_1 is independent of t for $0 \leq t \leq 1$. Applying this to $t = 0$ and $t = 1$, we see that the number of the y_v contained in G_1 is the same as the number of the x_v inside G_1 , and the analogous result holds of course for every G_κ .

11. We reorder the y_v in such a way that all y_v contained in a set G_κ have the same indices as the x_v from the same G_κ . This obviously can be done in many ways if a G_κ contains more than one x_v .

Consider now a zero y_v of $g(z)$ which is contained in G_κ . Then it follows from (B.16) that for a certain x_μ from G_κ we have $|y_v - x_\mu| \leq r_\mu \varepsilon$, $y_v = x_\mu(1 + \theta\varepsilon)$, $|\theta| \leq 1$. On the other hand x_μ is connected with x_v . It follows therefore from (B.22) that

$$y_v = x_v(1 + \theta\varepsilon) \prod_{\mu=1}^{n-1} \frac{1 + \theta_\mu \varepsilon}{1 + \theta'_\mu \varepsilon},$$

$$\frac{y_v}{x_v} - 1 = (1 + \theta\varepsilon) \prod_{\mu=1}^{n-1} \frac{1 + \theta_\mu \varepsilon}{1 + \theta'_\mu \varepsilon} - 1.$$

The right-hand expression is majorized by $[(1 + \varepsilon)^n / (1 - \varepsilon)^{n-1}] - 1$ and we have

$$\left| \frac{y_v}{x_v} - 1 \right| \leq \varphi(\varepsilon), \quad \varphi(\varepsilon) = \frac{(1 + \varepsilon)^n}{(1 - \varepsilon)^{n-1}} - 1. \quad (\text{B.24})$$

12. We now put

$$\delta = \frac{2\tau'}{1 - \tau'}, \quad \tau' = \tau^{1/n}, \quad (\text{B.25})$$

and discuss first $\varphi(\delta)$. Using (B.2), $\tau' \leq 1/4n$, we have by (B.25)

$$\varphi(\delta) = \frac{(1 + \tau')^n}{(1 - \tau')(1 - 3\tau')^{n-1}} - 1 = \tau' \psi(\tau'),$$

where $\psi(\tau')$ is a power series with *positive* coefficients. We have therefore

$$\psi(\tau') \leq \psi\left(\frac{1}{4n}\right), \quad \varphi(\delta) \leq \psi\left(\frac{1}{4n}\right)\tau'.$$

On the other hand,

$$\psi\left(\frac{1}{4n}\right) = 4n \left[\frac{\left(1 + \frac{1}{4n}\right)^n}{\left(1 - \frac{3}{4n}\right)^{n-1} \left(1 - \frac{1}{4n}\right)} - 1 \right].$$

The first term in the square brackets is here

$$\begin{aligned} & \frac{4n}{4n-1} \left(1 + \frac{1}{4n}\right)^n \left(1 + \frac{3}{4n-3}\right)^{n-1} \\ & < \left(1 + \frac{1}{4n-1}\right) \left(1 + \frac{3/4}{n-1}\right)^{n-1} \left(1 + \frac{1/4}{n}\right)^n < \left(1 + \frac{1}{4n-1}\right) e^{3/4} e^{1/4}, \end{aligned}$$

since we have for all positive integers n and all positive x

$$\frac{x}{n} > \ln\left(1 + \frac{x}{n}\right), \quad e^x > \left(1 + \frac{x}{n}\right)^n.$$

For $n \geq 3$ this is majorized by $(12/11)e < 2.9654$, and we see that the expression in the square brackets is < 1.9654 for $n \geq 3$. For $n = 2$, this expression has the value

$$\left(\frac{9}{8}\right)^2 \cdot \frac{8}{5} \cdot \frac{8}{7} - 1 = \frac{46}{35} < 1.9,$$

and we see finally that $\psi\left(\frac{1}{4n}\right) < 7.9n$. We have therefore

$$\varphi(\delta) < 7.9n\tau'. \quad (\text{B.26})$$

Observe now that by (B.17) ε can be chosen arbitrarily between δ and 1. Since $\varphi(\varepsilon)$ is continuous for $\varepsilon < 1$, we can therefore from the beginning choose ε so near δ that we have $\varphi(\varepsilon) < 8n\tau'$ and inequality (B.4) is proved.

13. It may be added that if in relation (B.4) the x_v, y_v are replaced by $|x_v|, |y_v|$, the bound on the right-hand side can be replaced by a smaller one, which is $\sim 2n\tau^{1/4}$ as $\tau \downarrow 0$.[†]

14. Applying our theorem, we obtain from (B.3) for a v with $a_v = 0$: $b_v = a_v$; and if a_v is very small, we have from (B.3) only a very small range of values for b_v . It is therefore very important that our theorem hold even if the assumption (B.3) is replaced by (B.7). Consider, for instance, the equation $x^n - 1 = 0$. Here we obtain from (B.4), using (B.3), $|b_0 - 1| \leq \tau$, $|b_n + 1| \leq \tau$, $b_v = 0$ ($0 < v < n$). On the other hand, in this case we have $S_v = \binom{n}{v}$ and using (B.7), our conditions for b_v become

$$|b_0 - 1| \leq \tau, \quad |b_n + 1| \leq \tau, \quad |b_v| \leq \binom{n}{v} \tau \quad (0 < v < n-1).$$

However, in order to apply (B.7) directly in this way, we must know all $|x_v|$, which is usually not the case.

15. We will now treat the problem of how *positive* constants T_v can be formed directly from the $|a_v|$ in such a way that we have

$$|a_v| \leq T_v \leq S_v \quad (v = 1, \dots, n-1), \quad (\text{B.27})$$

[†] Compare A. Ostrowski, Mathematische Miszellen XXIV, "Zur relativen Stetigkeit von Wurzeln algebraischer Gleichungen," *Jahresber. Deut. Math.-Ver.* **58**, 98–102 (1956).

and that therefore the conditions

$$|b_v - a_v| \leq \tau T_v \quad (\text{B.28})$$

can be used instead of (B.3).

For this purpose we need an inequality which goes back to Newton. If the polynomial

$$\varphi(z) = q_0 z^n + \binom{n}{1} q_1 z^{n-1} + \cdots + \binom{n}{v} q_v z^{n-v} + \cdots + q_n, \quad q_0 \neq 0, \quad (\text{B.29})$$

has real coefficients and n real roots, then we have

$$q_v^2 \geq q_{v-1} q_{v+1} \quad (v = 1, \dots, n-1). \quad (\text{B.30})$$

This is obvious for $n = 2$ and is proved in the general case by induction using Rollés theorem.[†]

16. We assume from now on that

$$a_0 = 1, \quad (\text{B.31})$$

write $F(z)$ from (B.5) in the form

$$F(z) = \sum_{v=0}^n \binom{n}{v} s_v z^{n-v}, \quad \binom{n}{v} s_v = S_v, \quad s_0 = S_0 = 1, \quad (\text{B.32})$$

and put

$$|a_v| = \binom{n}{v} k_v \quad (v = 0, \dots, n); \quad (\text{B.33})$$

then we have from (B.30)

$$s_v^2 \geq s_{v-1} s_{v+1} \quad (v = 1, \dots, n-1). \quad (\text{B.34})$$

Suppose now that for a certain index v we have $a_v = 0$, $a_{v-1} a_{v+1} \neq 0$. Then it follows from (B.34) and (B.6) that $\sqrt{k_{v-1} k_{v+1}} \leq s_v$ and therefore, if we use (B.33),

$$\sqrt{\left(1 + \frac{1}{v}\right)\left(1 + \frac{1}{n-v}\right)} |a_{v-1} a_{v+1}| \leq S_v.$$

Therefore, in the corresponding relation (B.3) we can replace $|a_v|$ in the right-hand expression by

$$\sqrt{\left(1 + \frac{1}{v}\right)\left(1 + \frac{1}{n-v}\right)} |a_{v-1} a_{v+1}|.$$

[†] Compare, for instance, G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, pp. 53, 54, Cambridge Univ. Press, London and New York, 1934.

If a sequence of a_v vanishes, e.g.,

$$a_{v_1} \neq 0, \quad a_{v_1+1} = a_{v_1+2} = \cdots = a_{v_2-1} = 0, \quad a_{v_2} \neq 0, \quad (\text{B.35})$$

then it is easy to show that the corresponding $|a_v|$ in the inequality (B.3) can be replaced by the expressions

$$\binom{n}{v} \binom{n}{v_2}^{(v_1-v)/(v_2-v_1)} \binom{n}{v_1}^{(v-v_2)/(v_2-v_1)} a_{v_1}^{(v-v_1)/(v_2-v_1)} a_{v_1}^{(v_2-v)/(v_1-v_1)}. \quad (\text{B.36})$$

However, in this case it will be preferable to use a systematic geometric approach to our problem of building up the expressions for the T_v .

17. Put

$$\sigma_v = \ln s_v, \quad \kappa_v = \ln k_v \quad (v = 0, 1, \dots, n), \quad (\text{B.37})$$

where, if $k_v = 0$, the corresponding κ_v is $-\infty$. Then we have from (B.34) and (B.6)

$$\sigma_v \geq \frac{\sigma_{v-1} + \sigma_{v+1}}{2}, \quad (\text{B.38})$$

$$\kappa_v \leq \sigma_v. \quad (\text{B.39})$$

Inequality (B.38) has a very simple geometric interpretation. If we mark in the (v, σ) -plane the points with the coordinates (v, σ_v) , $v = 0, 1, \dots, n$, and connect them by rectilinear segments (see Fig. 13), we obtain a polygonal line P_S , which is *convex from above*. On the other hand, if we mark the points (v, κ_v) ($v = 0, 1, \dots, n$), these points lie by (B.39) *below* P_S or *on* this polygonal line.

We now draw a polygonal line P_T connecting the points $(0, \kappa_0)$ and (n, κ_n) , *convex from above* and such that all vertices of P_T belong to the points (v, κ_v) ,

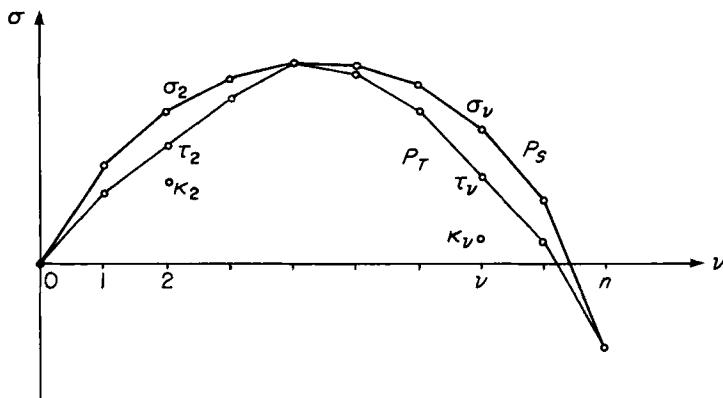


FIGURE 13

while all other points (v, κ_v) lie either on P_T or below. It follows from (B.6) that no point of P_T is situated above P_S . Denote for each $v, v = 0, 1, \dots, n$, by τ_v the ordinate of the point of P_T corresponding to the abscissa v ; then we have obviously

$$\kappa_v \leq \tau_v \leq \sigma_v, \quad (\text{B.40})$$

and therefore, if we put

$$\binom{n}{v} e^{\tau_v} = T_v, \quad (\text{B.41})$$

(B.27) is certainly satisfied. In this way we can replace the conditions (B.3) by the conditions (B.28), where the constants T_v are obtained by constructing the polygonal line P_T .

C

An Explicit Formula for the n th Derivative of the Inverse Function

1. Suppose $y = f(x)$ differentiable a sufficient number of times and denote $f^{(v)}(x)$ by y_v . Let $x = \varphi(y)$ be the inverse function of $y = f(x)$. Then we have seen in Chapter 2 that we can write

$$\varphi^{(n)}(y) = y_1^{-(2n-1)} X(y_1, \dots, y_n), \quad (\text{C.1})$$

where X_n is a polynomial in y_1, \dots, y_n :

$$X_n = \sum a_{\alpha_1 \dots \alpha_n} y_1^{\alpha_1} \cdots y_n^{\alpha_n}, \quad (\text{C.2})$$

the exponents $\alpha_1, \dots, \alpha_n$ being subjected to the conditions

$$\alpha_v \geq 0, \quad \sum_{v=1}^n \alpha_v = n-1, \quad \sum_{v=1}^n v\alpha_v = 2n-2. \quad (\text{C.3})$$

We will prove in this appendix that each term satisfying the conditions (C.3) has in X_n a nonvanishing coefficient and that this coefficient is given by the formula

$$a_{\alpha_1 \dots \alpha_n} = (-1)^{n-\alpha_1-1} \frac{(2n-\alpha_1-2)!}{\alpha_2!(2!)^{\alpha_2}\alpha_3!(3!)^{\alpha_3} \cdots \alpha_n!(n!)^{\alpha_n}}. \quad (\text{C.4})$$

2. If we consider in particular the term of X_n containing y_n , it follows at once from the conditions (C.3) that we have for $n > 1$:

$$\alpha_n = 1, \quad \alpha_1 = n-2,$$

while all α_v with $1 < v < n$ vanish. But then the coefficient (C.4) becomes -1 in accordance with (2.12). Therefore, our assertion is certainly true for $\alpha_n > 0$, $n > 1$.

Instead of proving (C.4) directly, we prove the corresponding expression for the coefficient of $\varphi^{(n)}(y)$. Indeed, dividing (C.2) by y_1^{2n-1} and putting

$$2n - \alpha_1 - 1 = \beta_1, \quad \alpha_v = \beta_v \quad (v > 1),$$

we obtain from (C.2) and (C.4) the expression

$$\varphi^{(n)}(y) = \sum (-1)^{n+\beta_1} \frac{(\beta_1-1)!}{\beta_2!(2!)^{\beta_2} \cdots \beta_n!(n!)^{\beta_n}} y_1^{-\beta_1} y_2^{\beta_2} \cdots y_n^{\beta_n}, \quad (\text{C.5})$$

where the β_v are subject to the conditions

$$\beta_1 \leq 2n - 1, \quad \beta_v \geq 0 \quad (v > 1), \quad \sum_{v>1} \beta_v = \beta_1 - n, \quad (\text{C.6})$$

$$\sum_{v>1} v\beta_v = \beta_1 - 1. \quad (\text{C.7})$$

3. We know already that the term of (C.5) with $\beta_n > 0$, $n > 1$ is correct. Observe now that for $n = 1$ the conditions (C.6) and (C.7) give $\beta_1 = n = 1$, and we obtain from (C.5) $\varphi'(y) = 1/y_1$.

For $n = 2$ again, the only combination of β_1, β_2 satisfying (C.6) and (C.7) is given by

$$\beta_1 = 3, \quad \beta_2 = 1,$$

as we have $\beta_2 = \beta_1 - 2$, $2\beta_2 = \beta_1 - 1$. According to (C.5), $\varphi^{(n)}(y)$ becomes here $-y_2/y_1^3$, which is true.

We shall therefore assume that (C.5) is true for a value of $n \geq 2$ and prove the corresponding formulas

$$T_{\gamma_1 \dots \gamma_{n+1}} := T := (-1)^{n+1+\gamma_1} \frac{(\gamma_1 - 1)!}{\gamma_2! (2!)^{\gamma_2} \dots \gamma_n! (n!)^{\gamma_n} \gamma_{n+1}! [(n+1)!]^{\gamma_{n+1}}}, \quad (\text{C.8})$$

where we put

$$\varphi^{(n+1)}(y) = \sum T_{\gamma_1 \dots \gamma_{n+1}} y_1^{-\gamma_1} y_2^{\gamma_2} \dots y_{n+1}^{\gamma_{n+1}}. \quad (\text{C.9})$$

In this proof we can assume without loss of generality that $\gamma_{n+1} = 0$.

Our γ_v satisfy the conditions which are analogous to (C.6) and (C.7), and the condition corresponding to (C.7) is

$$\sum_{v \geq 2} v\gamma_v = \gamma_1 - 1. \quad (\text{C.10})$$

It is easily seen that $\gamma_1 - 1$ is > 0 . For otherwise from (C.10) it would follow that

$$\gamma_1 = 1, \quad \gamma_2 = \gamma_3 = \dots = \gamma_n = 0.$$

On the other hand we have, as in (C.6),

$$\gamma_2 + \dots + \gamma_{n-1} = \gamma_1 - (n+1).$$

and it would follow that $n+1 = 1$, while we have $n > 1$.

4. Now the term of (C.9) with $y_1^{-\gamma_1} y_2^{\gamma_2} \dots y_{n+1}^{\gamma_{n+1}}$ is obtained from different terms of (C.5) by the process

$$\varphi^{(n+1)}(y) = \frac{1}{y_1} \left(y_2 \frac{\partial}{\partial y_1} + y_3 \frac{\partial}{\partial y_2} + \dots + y_n \frac{\partial}{\partial y_{n-1}} + y_{n+1} \frac{\partial}{\partial y_n} \right) \varphi^{(n)}(y), \quad (\text{C.11})$$

and we will compute the contributions of the different terms of (C.11) to $T_{\gamma_1 \dots \gamma_{n+1}}$. Consider first the contribution of $(y_2/y_1) \partial/\partial y_1$. This obviously must be applied to the monomial

$$y_1^{-\beta_1} y_2^{\beta_2} \cdots y_n^{\beta_n}, \quad (\text{C.12})$$

where

$$\gamma_1 = \beta_1 + 2, \quad \gamma_2 = \beta_2 + 1, \quad \beta_\mu = \gamma_\mu \quad (\mu = 3, \dots, n).$$

But then the contribution of this term to $T_{\gamma_1 \dots \gamma_{n+1}}$ is

$$-(\gamma_1 - 2)(-1)^{n+\gamma_1-2} \frac{(\gamma_1 - 3)!}{(\gamma_2 - 1)! (2!)^{\gamma_2-1} \gamma_3! (3!)^{\gamma_3} \cdots \gamma_n! (n!)^{\gamma_n}},$$

and this is $[2\gamma_2/(\gamma_1 - 1)] T$.

5. Consider now for a $v > 1$ the contribution arising from $(y_{v+1}/y_1) \partial/\partial y_v$. This operation must be applied to the term of (C.5) corresponding to the monomial (C.12) with

$$\begin{aligned} \beta_1 &= \gamma_1 - 1, & \beta_v &= \gamma_v + 1, & \beta_{v+1} &= \gamma_{v+1} - 1, \\ \beta_\mu &= \gamma_\mu, & (\mu \neq 1 \wedge v \wedge (v+1)). \end{aligned}$$

The corresponding contribution to $T_{\gamma_1 \dots \gamma_{n+1}}$ is then

$$\frac{(\gamma_v + 1)(-1)^{n+\gamma_1-1} (\gamma_1 - 2)!}{\cdots (\gamma_v + 1)! (v!)^{\gamma_v+1} (\gamma_{v+1} - 1)! [(v+1)!]^{\gamma_{v+1}-1} \cdots},$$

where the factors in the denominator corresponding to $\gamma_2, \dots, \gamma_{v-1}, \gamma_{v+2}, \dots, \gamma_n$ are the same as in (C.8) and are not explicitly written down; this expression is obviously

$$\frac{(v+1)! \gamma_{v+1}}{v! (\gamma_1 - 1)} T = \frac{(v+1) \gamma_{v+1}}{\gamma_1 - 1} T.$$

Forming now the sum of all contributions, we get

$$\frac{T}{\gamma_1 - 1} \left(\sum_{v=2}^n (v+1) \gamma_{v+1} + 2\gamma_2 \right). \quad (\text{C.13})$$

But now it follows at once from (C.10) that the expression in parentheses is $= \gamma_1 - 1$, and we see that the expression (C.13) is $= T$. Formula (C.8) is proved.

6. If $f(x)$ is a quadratic polynomial, we have $y_3 = y_4 = \cdots = 0$ and in (C.5) $\beta_2 = n-1$, $\beta_1 = 2n-1$,

$$\varphi^{(n)}(y) = (-1)^{n-1} 3 \cdot 5 \cdots (2n-3) \frac{y_2^{n-1}}{y_1^{2n-1}}. \quad (\text{C.14})$$

If $f(x)$ is a *cubic polynomial*, we have $y_4 = y_5 = \dots = 0$ and $\varphi^{(n)}(y)$ becomes

$$\varphi^{(n)}(y) = \sum (-1)^{n+\beta_1} \frac{(\beta_1 - 1)!}{\beta_2! 2^{\beta_2} \beta_3! 6^{\beta_3}} y_1^{-\beta_1} y_2^{\beta_2} y_3^{\beta_3}$$

with

$$\begin{aligned} \beta_2 &\geq 0, & \beta_3 &\geq 0, & \beta_1 &\leq 2n - 1, & \beta_2 + \beta_3 &= \beta_1 - n, \\ 2\beta_2 + 3\beta_3 &= \beta_1 - 1. \end{aligned}$$

We put $\beta_3 = v$ and obtain

$$\beta_2 = n - v - 1, \quad \beta_1 = 2n - v - 1.$$

v is ≥ 0 but must remain $\leq (n-1)/2$, since β_2 is ≥ 0 . Our expression for $\varphi^{(n)}(y)$ now becomes

$$\varphi^{(n)}(y) = y_2^{n-1} y_1^{1-2n} \sum_{v=0}^{(n-1)/2} (-1)^{n-v-1} \frac{(2n-v-2)!}{(n-2v-1)! v! 2^{n-v-1} 3^v} \left(\frac{y_1 y_3}{y_2^2} \right)^v \quad (\text{C.15})$$

D

Analog of the Regula Falsi for Two Equations with Two Unknowns

1. In order to solve approximately the equations

$$F(P) \equiv F(x, y) = 0, \quad G(P) \equiv G(x, y) = 0,$$

assume with C. F. Gauss[†] that for three points P_1, P_2, P_3 the values of F and G are known:

$$F(P_v), \quad G(P_v) \quad (v = 1, 2, 3).$$

Then we define two linear functions in x, y ,

$$L_1 \equiv ax + by + c, \quad L_2 \equiv ex + fy + g,$$

by the six conditions

$$L_1(P_v) = F(P_v), \quad L_2(P_v) = G(P_v) \quad (v = 1, 2, 3) \quad (\text{D.1})$$

and have to solve with respect to x, y the two equations

$$L_1(x, y) = 0, \quad L_2(x, y) = 0. \quad (\text{D.2})$$

In order to eliminate a, b, c, e, f, g from the system (D.1), (D.2), we first eliminate c and g , subtracting for each v from any of the equations (D.2) the corresponding equation (D.1); thus we obtain

$$\begin{aligned} a(x - x_v) + b(y - y_v) &= -F(P_v) \\ e(x - x_v) + f(y - y_v) &= -G(P_v) \end{aligned} \quad (v = 1, 2, 3). \quad (\text{D.3})$$

2. The result of the elimination of a, b, e, f from (D.3) amounts to the statement that the (3×4) matrix

$$(x - x_v \quad y - y_v \quad F(P_v) \quad G(P_v)) \quad (v = 1, 2, 3)$$

has rank ≤ 2 . We obtain the two equations

$$\left| \begin{array}{c} x - x_v \\ F(P_v) \\ G(P_v) \end{array} \right| = 0, \quad \left| \begin{array}{c} y - y_v \\ F(P_v) \\ G(P_v) \end{array} \right| = 0, \quad (\text{D.4})$$

[†] *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, No. 120, pp. 136–138, 1809; C. F. Gauss, *Werke*, Vol. VII, pp. 150–152.

which were given by Gauss in order to obtain from the three given approximations P_v to the solution in question a fourth improved one.

If we assume that the determinant

$$\Delta = \begin{vmatrix} 1 \\ F(P_v) \\ G(P_v) \end{vmatrix} \quad (v = 1, 2, 3) \quad (\text{D.5})$$

is $\neq 0$, the solution of (D.4) is given by

$$x = \frac{1}{\Delta} \begin{vmatrix} x_v \\ F(P_v) \\ G(P_v) \end{vmatrix}, \quad y = \frac{1}{\Delta} \begin{vmatrix} y_v \\ F(P_v) \\ G(P_v) \end{vmatrix} \quad (v = 1, 2, 3). \quad (\text{D.6})$$

3. The condition $\Delta \neq 0$ is equivalent to the condition that the three points $[F(P_v), G(P_v)]$ ($v = 1, 2, 3$) are not collinear. On the other hand, we obtain from (D.5), using (D.3), the identical relation

$$\Delta = (af - be) \begin{vmatrix} 1 \\ x_v \\ y_v \end{vmatrix} \quad (v = 1, 2, 3). \quad (\text{D.7})$$

We see that the necessary condition for $\Delta \neq 0$ is that the three points P_1, P_2, P_3 are not collinear. Gauss proposed therefore to choose the triplet P_1, P_2, P_3 in such a way that $x_2 = x_1, y_3 = y_1$; i.e., P_1, P_2, P_3 form a right-angled triangle.

After the fourth approximating point P_4 has been found from (D.6), one of the points P_1, P_2, P_3 is dropped and the procedure is repeated, starting from the remaining triplet.

Apparently this method remained practically unknown and has not been used much in computational work, while the discussion of the convergence, which implies some rather subtle points, has never been carried through. However, in the last years several methods have been developed that can be considered as more or less complete generalizations of the *regula falsi* to n -dimensional spaces and even to functional spaces.

E

Steffensen's Improved Iteration Rule

1. In order to improve the iteration by $\psi(x)$ Steffensen[†] proposed in 1933 to obtain, starting from an x_0 , the values $x_1 = \psi(x_0)$, $x_2 = \psi(x_1)$ and then to apply the *regula falsi* to the equation

$$F(x) \equiv x - \psi(x) = 0$$

and to the two points x_0, x_1 . In this way we obtain for the next approximation

$$y_1 = \frac{x_0 F(x_1) - x_1 F(x_0)}{F(x_1) - F(x_0)}$$

and, since $F(x_0) = x_0 - x_1$, $F(x_1) = x_1 - x_2$,

$$y_1 = \frac{x_0 x_2 - x_1^2}{x_0 - 2x_1 + x_2}.$$

If we now put $y_0 = x_0$, we obtain finally the iteration $y_1 = \Psi(y_0)$ with

$$\Psi(y) = \frac{y\psi[\psi(y)] - \psi(y)^2}{y - 2\psi(y) + \psi[\psi(y)]}. \quad (\text{E.1})$$

2. Steffensen showed by examples that this iteration converges very quickly to a zero of $F(x)$, even if $|\psi'| > 1$. Willers says in his textbook ([7], p. 259) that this method "always works" and gives in an article in *Zeitschrift für angewandte Mathematik und Mechanik*[§] an elegant geometric illustration for the working of this iteration. Householder ([1]), pp. 126–128 shows under somewhat special assumptions that a fixed point of the iteration by $\psi(x)$ becomes a point of attraction for $\Psi(y)$. In what follows we shall give very

† J. F. Steffensen, "Remarks on iteration," *Skand. Aktuar Tidskr.* **16**, 64–72 (1933).

‡ The same formula is used in an entirely different connection in the so-called "Aitken's method." In this method the use of expression (E.1) serves to derive from a given complete iteration sequence a better convergent one. However, by this "Aitken's transformation" a linearly convergent sequence remains, as a rule, linearly convergent after transformation.

§ F. A. Willers, *Z. Angew. Math. Mech.* **22**, 125–126 (1948).

general conditions for a zero ζ of $F(x)$ to be a point of attraction for the iteration with $\Psi(y)$.[†]

3. In proving our theorems, the computations are considerably simplified if we assume $\zeta = 0$. This is always possible, since, if we put

$$x = z + \zeta, \quad F(x) = F^*(z) \quad \psi^*(z) = \psi(z + \zeta) - \zeta,$$

then the numbers $z_v = x_v - \zeta$ are obtained from $z_0 = x_0 - \zeta$ by iteration with the iterating function $\psi^*(z)$.

We assume from now on that $\psi'(\zeta)$ exists and has the value α . We prove first:

I. If $\alpha \neq 1$, then for the iterating function $\Psi(x)$ the point ζ always becomes a point of attraction and we have $\Psi'(\zeta) = 0$.

4. Proof. Without loss of generality assume $\zeta = 0$. We have then for $x \rightarrow 0$

$$\psi(x) = \alpha x + o(x),$$

$$\psi[\psi(x)] = \alpha[\alpha x + o(x)] + o[\alpha x + o(x)] = \alpha^2 x + o(x),$$

$$\psi[\psi(x)] - 2\psi(x) + x = (\alpha - 1)^2 x + o(x),$$

$$\psi(x)^2 = \alpha^2 x^2 + o(x^2),$$

$$x\psi[\psi(x)] - \psi(x)^2 = o(x^2),$$

and it follows from (E.1) that $\Psi(x) = o(x)$, i.e., $\Psi'(0) = 0$. I is proved.

The result of I can be improved if we know more about the order of vanishing of $\psi(x) - \zeta - \alpha(x - \zeta)$ as $x \rightarrow \zeta$.

5. II. Suppose that for a $\lambda > 1$ with $x \rightarrow \zeta$ the expression

$$E(x) = \frac{\psi(x) - \zeta - \alpha(x - \zeta)}{|x - \zeta|^\lambda} \quad (\text{E.2})$$

either (a) remains bounded, i.e., is $O(1)$, or (b) tends to 0, i.e., is $o(1)$. Then we have respectively, if $\alpha(\alpha - 1) \neq 0$,

$$\Psi(x) - \zeta = O(|x - \zeta|^\lambda), \quad (\text{E.3a})$$

$$\Psi(x) - \zeta = o(|x - \zeta|^\lambda). \quad (\text{E.3b})$$

If $\alpha > 0$ and one of the assumptions (a) or (b) holds for a one-sided convergence to ζ , then the corresponding relation (E.3a) or (E.3b) holds for the same one-sided convergence.

[†] In the discussion of this situation the value of $\Psi(\zeta)$ cannot be obtained from (E.1) since this expression becomes indeterminate at ζ . We define therefore once and for all $\Psi(\zeta)$ as ζ . The continuity of $\Psi(y)$ at ζ follows then from the results obtained in this appendix under the corresponding conditions.

6. Proof. Without loss of generality assume $\zeta = 0$. Then we have from (E.2)

$$\begin{aligned}\psi(x) &= \alpha x + E(x)|x|^\lambda, \\ \psi[\psi(x)] &= \alpha[\alpha x + E(x)|x|^\lambda] + E[\psi(x)]|\alpha x + E(x)|x|^\lambda|^{\lambda} \\ &= \alpha^2 x + \{\alpha E(x) + |\alpha|^\lambda E[\psi(x)]\}|x|^\lambda + o(|x|^\lambda).\end{aligned}$$

It now follows that

$$\psi[\psi(x)] - 2\psi(x) + x = (\alpha - 1)^2 x + o(x), \quad (\text{E.4})$$

$$\psi(x)^2 = \alpha^2 x^2 + 2\alpha E(x)|x|^\lambda x + o(|x|^{\lambda+1}),$$

$$x\psi[\psi(x)] - \psi(x)^2 = \{|\alpha|^\lambda E[\psi(x)] - \alpha E(x)\}|x|^\lambda x + o(|x|^{\lambda+1}). \quad (\text{E.5})$$

From (E.4) and (E.5) we obtain by division

$$\Psi(x) = \frac{|x|^\lambda}{(\alpha - 1)^2} \{|\alpha|^\lambda E[\psi(x)] - \alpha E(x)\} + o(|x|^\lambda). \quad (\text{E.6})$$

Observe now that as $x \rightarrow 0$, then $\psi(x) \rightarrow 0$, and we see that in the cases (a) and (b) the assertions (E.3a) and (E.3b) follow from (E.6) immediately.

7. If $\alpha > 0$ and, for instance, $E(x) = O(1)$ for $x \downarrow 0$, then $\psi(x)$ remains > 0 for sufficiently small $x > 0$ and $E(\psi(x))$ is also $O(1)$. But then (E.3a) follows again from (E.6) for $x \downarrow 0$. In the same way we obtain the assertion corresponding to $x \uparrow 0$. II is proved

8. If $\alpha = 0$, the assertions of I and II would not give any improvement of the convergence. However, in this case, II can be sharpened to give an improvement:

III. Suppose that $\alpha = 0$ and for a $\lambda > 1$

$$E(x) = \frac{\psi(x) - \zeta}{|x - \zeta|^\lambda} \quad (\text{E.7})$$

as $x \rightarrow \zeta$ either (a) remains bounded, or (b) tends to 0. Then we have accordingly[†]

[†] Householder (*loc. cit.*) obtains the results corresponding to those of II and III assuming that $\psi(x)$ can be developed in (apparently) integer powers of $x - \zeta$. Further, Householder gives a generalization of Steffensen's procedure by showing that from any two iterating functions ψ_1, ψ_2 , a new iterating function $\Psi(x)$ can be formed which usually gives a faster convergence than ψ_1 and ψ_2 . For $\psi_1 = \psi_2$, Householder's procedure becomes that of Steffensen. On the other hand, a more detailed study of Householder's generalization shows that the combination of ψ_1 with ψ_2 can be particularly useful if ψ_2 is obtained by combining ψ_1 with ψ_1 , and if $\psi_1'(\zeta) = 1$. See our notes: "Über Verfahren von Steffensen und Householder zur Verbesserung der Konvergenz von Iterationen," *Z. Angew. Math. Phys.* 7, 218–219 (1956); "On the Convergence of the Rayleigh Quotient Iteration for the Computation of Characteristic Roots and Vectors VI. (Usual Rayleigh Quotient for Nonlinear Elementary Divisors)," *Arch. Rational Mech. and Anal.* 4, 152–165 (1959).

$$\Psi(x) - \zeta = O(|x - \zeta|^{2\lambda-1}), \quad (\text{E.8a})$$

$$\Psi(x) - \zeta = o(|x - \zeta|^{2\lambda-1}). \quad (\text{E.8b})$$

9. Proof. Assume without loss of generality $\zeta = 0$. We then have

$$\psi(x) = E(x)|x|^\lambda, \quad \psi(x)^2 = E(x)^2|x|^{2\lambda},$$

$$\psi[\psi(x)] = E[\psi(x)]|E(x)|x|^\lambda|^\lambda = E[\psi(x)]|E(x)|^\lambda|x|^{\lambda^2};$$

but now it follows, since $E(\psi(x)) = O(1)$, that

$$\psi[\psi(x)] - 2\psi(x) + x = x + O(|x|^\lambda)$$

and

$$|x\psi[\psi(x)]| = O(|x|^{\lambda^2+1}) = o(|x|^{2\lambda}),$$

since $\lambda^2 + 1 > 2\lambda$. Further,

$$x\psi[\psi(x)] - \psi(x)^2 = -E(x)^2|x|^{2\lambda} + o(|x|^{2\lambda}).$$

From (E.1) we obtain finally

$$\Psi(x) = -E(x)^2x|x|^{2\lambda-2} + o(|x|^{2\lambda-1}), \quad (\text{E.9})$$

and the assertions (E.8a) and (E.8b) follow immediately.

10. We consider finally the case $\psi'(\zeta) = \alpha = 1$.

IV. Assume that $\alpha = 1$, $\psi'(x)$ is continuous in the neighborhood of ζ , and for a constant $\lambda > 1$

$$\psi'(x) - 1 = T(x)|x - \zeta|^{\lambda-1}, \quad (\text{E.10})$$

where either as $x \uparrow \zeta$ or $x \downarrow \zeta$, $T(x)$ tends to a limit $\gamma \neq 0$. Then we have for the corresponding one-sided derivative

$$\Psi'(\zeta) = 1 - \frac{1}{\lambda}, \quad (\text{E.11})$$

and ζ is a point of attraction (from the corresponding side) for the iterating function $\Psi(x)$.

11. Proof. Without loss of generality assume $\zeta = 0$. Consider the function

$$g(x) = \psi(x) - x; \quad (\text{E.12})$$

we have by virtue of (E.10)

$$g(x) = \int_0^x T(x)|x|^{\lambda-1} dx = \frac{\gamma}{\lambda}x|x|^{\lambda-1} + o(|x^\lambda|) \quad (\text{E.13})$$

as follows at once, considering the cases $x \gtrless 0$ separately.

12. If we replace y by x in (E.1) and subtract x on both sides, we have

$$\Psi(x) - x = \frac{-(\psi(x) - x)^2}{\psi[\psi(x)] - 2\psi(x) + x}.$$

The denominator here is by virtue of (E.12) $\psi[\psi(x)] - g(x)$, and we obtain therefore

$$\Psi(x) - x = -\frac{g(x)^2}{g[\psi(x)] - g(x)}. \quad (\text{E.14})$$

It follows by the mean value theorem that

$$g[\psi(x)] - g(x) = [\psi(x) - x]g'(\xi) = g(x)g'(\xi), \quad (\text{E.15})$$

where ξ lies between x and $\psi(x)$, and therefore by (E.13) for $0 < \theta < 1$

$$\xi = x + \theta g(x) = x + o(x),$$

and by (E.10)

$$g'(\xi) = \psi'(\xi) - 1 = T(\xi)|\xi|^{\lambda-1} = \gamma|x|^{\lambda-1} + o(|x|^{\lambda-1}).$$

It follows from (E.14), (E.13), and (E.15) by division

$$\Psi(x) - x = -\frac{g(x)}{g'(\xi)} = -\frac{(\gamma/\lambda)x + o(x)}{\gamma + o(1)} \sim -\frac{x}{\lambda}. \quad (\text{E.16})$$

Equation (E.11) is an immediate consequence of (E.19).

13. The assumption $\alpha = 1$ of IV is obviously equivalent to $F'(\zeta) = 0$. We usually have then at ζ a multiple zero of $F(x)$ and by Theorem 5.2 the convergence in the case of the iterating function $\psi(x)$ is quite particularly slow, if there is convergence at all. On the other hand, it follows from IV that the convergence using the iterating function $\Psi(x)$ becomes linear and, if we apply Steffensen's procedure once more, even becomes quadratic by I.

If, as usually will be the case, λ is known, formula (4.8) can be used; and we obtain quadratic convergence, even without using Steffensen's procedure once more, replacing $\Psi(x)$ by

$$\Psi^*(x) = \lambda(\Psi(x) - (1 - 1/\lambda)x) = \lambda\Psi(x) - (\lambda - 1)x.$$

It may be remarked finally that from formula (E.14) it follows immediately that if the iteration with the iterating function $\Psi(x)$ is convergent to a limit ζ , ζ is in any case a zero of $\psi(x) - x$.

F

The Newton–Raphson Algorithm for Quadratic Polynomials

1. Theorem. Suppose that in the quadratic equation

$$f(x) := (x - \xi)(x - \eta) = 0, \quad (\text{F.1})$$

we have

$$|\xi - x_0| < |\eta - x_0|. \quad (\text{F.2})$$

Then the Newton–Raphson algorithm starting with x_0 is convergent to the value ξ . If $\xi = \eta$, then x_v is convergent to ξ for every $x_0 \neq \xi$.

On the other hand, if we have

$$|\xi - x_0| = |\eta - x_0|, \quad \xi \neq \eta, \quad x_0 \neq \frac{\xi + \eta}{2}, \quad (\text{F.3})$$

the Newton–Raphson algorithm starting with x_0 is divergent.

2. Proof. Put

$$\xi - x_v = a_v, \quad \eta - x_v = b_v, \quad a_0 = a, \quad b_0 = b. \quad (\text{F.4})$$

We have by definition

$$\begin{aligned} x_{v+1} &= x_v - \frac{(x_v - \xi)(x_v - \eta)}{2x_v - \xi - \eta} = x_v + \frac{a_v b_v}{a_v + b_v}, \\ a_{v+1} &= a_v - \frac{a_v b_v}{a_v + b_v} = \frac{a_v^2}{a_v + b_v}, \end{aligned}$$

and, if we use the symmetry,

$$a_{v+1} = \frac{a_v^2}{a_v + b_v}, \quad b_{v+1} = \frac{b_v^2}{a_v + b_v}. \quad (\text{F.5})$$

Now we have generally, if $\xi \neq \eta$,

$$a_v = \frac{a^{2^v}(a-b)}{a^{2^v}-b^{2^v}}, \quad b_v = \frac{b^{2^v}(a-b)}{a^{2^v}-b^{2^v}} \quad (\xi \neq \eta) \quad (\text{F.6})$$

and, if $\xi = \eta$, $a = b$,

$$a_v = b_v = \frac{a}{2^v} \quad (\xi = \eta). \quad (\text{F.6a})$$

3. Indeed, (F.6) is obvious for $\xi \neq \eta$ and $v = 0$. Suppose that (F.6) is true for a v ; we have then from (F.5)

$$a_{v+1} = \frac{a^{2^{v+1}}(a-b)^2/(a^{2^v}-b^{2^v})^2}{(a^{2^v}+b^{2^v})(a-b)/(a^{2^v}-b^{2^v})} = \frac{a^{2^{v+1}}(a-b)}{a^{2^{v+1}}-b^{2^{v+1}}}.$$

Since the value of b_{v+1} is obtained by symmetry, (F.6) is proved by induction.

In the case $\xi = \eta$, $a_v = b_v$, (F.5) becomes

$$a_{v+1} = b_{v+1} = a_v/2,$$

and (F.6a) follows at once.

4. Now in the case $\xi = \eta$ the assertion of our theorem follows immediately from (F.6a), since we then have

$$a_v = \xi - x_v \rightarrow 0 \quad (v \rightarrow \infty).$$

For $\xi \neq \eta$ we have under the hypothesis (F.2) $|a| < |b|$,

$$a_v \sim (b-a)(a/b)^{2^v} \rightarrow 0 \quad (v \rightarrow \infty).$$

If, on the other hand, $|a| = |b|$, $a \neq b$, we see from (F.6) that $|a_v| = |b_v|$, and in the case of convergence both a_v and b_v must tend to 0. But, on the other hand, it follows from (F.6) that $a_v - b_v = a - b$, and if a_v and b_v were both convergent to 0, we would have $a = b$, contrary to the hypothesis. The theorem is proved.

5. We shall further discuss to what extent, in the case of convergence, the sufficient conditions of Theorem 7.2 are satisfied.

Keeping the notation (F.4), put $p = b/a$; then we have either $p = 1$ or $|p| \neq 1$. From Section 2 we obtain

$$h_v = x_{v+1} - x_v = \frac{a_v b_v}{a_v + b_v} = a^{2^v} b^{2^v} \frac{(a-b)}{a^{2^{v+1}} - b^{2^{v+1}}} \quad (|p| \neq 1),$$

$$h_v = \frac{a}{2^{v+1}} \quad (p = 1).$$

Therefore, if we write $p := b/a$,

$$h_v = \begin{cases} \frac{a}{2^{v+1}} & (p = 1) \\ ap^{2^v} \frac{p-1}{p^{2^{v+1}} - 1} & (|p| \neq 1). \end{cases} \quad (\text{F.7})$$

On the other hand, by (F.4)

$$\begin{aligned} f'(x_v) &= 2x_v - \xi - \eta = -(a_v + b_v) = -(a^{2^v} + b^{2^v}) \frac{a-b}{a^{2^v} - b^{2^v}} \\ &= -a(p-1) \frac{p^{2^v}+1}{p^{2^v}-1} \quad (|p| \neq 1), \end{aligned}$$

and for $p = 1$,

$$f'(x_v) = -\frac{a}{2^{v-1}} \quad (p = 1).$$

Therefore, since in our case the number M of Theorem 7.2 is 2,

$$\begin{aligned} \frac{2Mh_v}{f'(x_v)} &= -4 \frac{p^{2^v}(p^{2^v}-1)}{(p^{2^{v+1}}-1)(p^{2^v}+1)} = -4 \frac{p^{2^v}}{(p^{2^v}+1)^2} \\ &= -\frac{4}{(p^{2^{v-1}}+p^{-2^{v-1}})^2} \quad (|p| \neq 1), \end{aligned} \quad (\text{F.8})$$

while for $p = 1$ we have

$$\frac{2Mh_v}{f'(x_v)} = -1 \quad (p = 1). \quad (\text{F.8a})$$

6. We see that in the case $p = 1$ we have for each v the limiting case $|2Mh_v/f'(x_v)| = 1$. What happens in the case $p \neq 1$?

From (F.8) it follows that the modulus of the left-hand expression tends to 0 as $v \rightarrow \infty$. Therefore, the conditions of Theorem 7.2 are satisfied from a certain v on. On the other hand, choosing p conveniently, we can ensure that the conditions of Theorem 7.2 do not hold for $v = 0, 1, \dots, N-2$, where N can be chosen as large as we like. Indeed, if we take $p = \rho e^{i\alpha}$, we have

$$|p^{2^{v-1}} + p^{-2^{v-1}}|^2 = \rho^{2^v} + \rho^{-2^v} + 2 \cos 2^v \alpha.$$

Take here $\alpha = \pi/2^N$, then we obtain for $v = 0, 1, \dots, N$

$$|p^{2^{v-1}} + p^{-2^{v-1}}|^2 = \rho^{2^v} + \rho^{-2^v} + 2 \cos(\pi/2^{N-v}).$$

But if we take $\rho > 1$ sufficiently near to 1, we can ensure that the inequality $\rho^{2^v} + \rho^{-2^v} < 4 - 2 \cos(\pi/4)$ holds for $v = 0, 1, \dots, N$, and the modulus of the right-hand expression in (F.8) remains > 1 for all these values of v .

7. We further ask whether the modulus of expression (F.8) can be 1 for one v or two consecutive values of v . Here we see at once from the relation $q^2 + 1/q^2 = (q + 1/q)^2 - 2$ that if both moduli $|q^2 + 1/q^2|$, $|q + 1/q|$ have the value 2, this is possible only if we have $q^2 + 1/q^2 = 2$, $q^2 = 1$, $q = \pm 1$. Therefore, since in (F.8) $|p| \neq 1$, the modulus of (F.8) certainly cannot be 1 for two consecutive values of v .

We see in particular that if for a value of v the expression $2Mh_v/f'(x_v)$ is $= 1$, this expression becomes < 1 for all greater v unless our quadratic polynomial has a double zero.

8. In order to apply our formulas to an important special case we develop them further.

Introducing again

$$p := \frac{b}{a}, \quad (\text{F.9})$$

we obtain at once from (F.6)

$$a_v = \frac{b-a}{p^{2^v}-1}, \quad b_v = \frac{b-a}{p^{2^v}-1} p^{2^v}. \quad (\text{F.10})$$

It follows then further that

$$f(x_v) = a_v b_v = (b-a)^2 \frac{p^{2^v}}{(p^{2^v}-1)^2}, \quad (\text{F.11})$$

$$-f'(x_v) = a_v + b_v = (b-a) \frac{p^{2^v}+1}{p^{2^v}-1}. \quad (\text{F.12})$$

9. We apply the above discussion to the polynomial

$$F(x) = A(x-\xi)(x-\eta) \quad (\text{F.13})$$

where, for a $\varphi > 0$,

$$\xi = 1 + e^{-\varphi}, \quad \eta = 1 + e^{\varphi}. \quad (\text{F.14})$$

The factor A is usually canceled out and can therefore be assumed, without loss of generality, as 1.

If we start with $x_0 = 0$, it follows from (F.4) and (F.9) that

$$a = \xi, \quad b = \eta, \quad a < b, \quad p := b/a = e^\varphi, \quad b - a = 2 \sin \varphi. \quad (\text{F.15})$$

(F.10) gives here

$$a_v = \frac{2 \sin \varphi}{\exp(2^v \varphi) - 1} = \frac{\sin \varphi}{\sin 2^{v-1} \varphi} \exp(-2^{v-1} \varphi), \quad b_v = \frac{\sin \varphi}{\sin 2^{v-1} \varphi} \exp(2^{v-1} \varphi). \quad (\text{F.16})$$

Since in our case, as easily follows from (F.7), $h_0 = 1$, we obtain finally

$$\xi - x_{v+1} = \frac{\sin \varphi}{\sin 2^v \varphi} \exp(-2^v \varphi) h_0.$$

The sufficient conditions of Theorem 7.1 are satisfied here; indeed, they are reduced by (F.8) to the inequality

$$p^{1/2} + p^{-1/2} \geq 2,$$

which is satisfied for any positive p .

10. Expression (F.16) for a_v shows that in our case, as $h_0 = 1$, (38.19) holds with the equality sign for every v . Further, introducing p and $(b-a)$ from (F.15) into (F.11), we obtain

$$f(x_v) = 4 \sin^2 \varphi \left(\frac{1}{\exp(2^{v-1}\varphi) - \exp(-2^{v-1}\varphi)} \right)^2 = \frac{\sin \varphi}{\sin^2 2^{v-1}\varphi}. \quad (\text{F.17})$$

It follows that in this case the first relation (38.21) holds with the equality sign for every v . On the other hand, we obtain from (F.12)

$$-f'(x_v) = 2 \sin \varphi \operatorname{Ctg} 2^{v-1}\varphi \quad (\text{F.18})$$

and, since in our case $Q_v = 1/f'(x_v)$, the first relation in formula (38.24) holds with the equality sign.

G

Some Modifications and Improvements of the Newton–Raphson Method

1. In some textbooks we find the remark that in the formula (6.3) the denominator $f'(x_v)$ can be replaced by $f'(x_1)$ as soon as x_1 is sufficiently near to ζ . This is obviously wrong, since we have then simply the iteration formula

$$x_{v+1} = x_v - cf(x_v)$$

and this formula provides, unless $c = 1/f'(\zeta)$, only a *linear convergence* and not the superlinear convergence characteristic of the Newton–Raphson method.

$$x^3 - 2x - 5 = 0, \quad \zeta = 2.094\ 551\ 481\ 542\ 326\ 591\ 482\ 386\ 54, \quad x_0 = 2$$

I	II
$x_1 = 2.1$	$x_1 = 2.1$
$x_2 = 2.094\ 568\ 1$	$x_2 = 2.093\ 9$
$x_3 = 2.094\ 551\ 481\ 72$	$x_3 = 2.094\ 627$
	$x_4 = 2.094\ 542\ 7$
	$x_5 = 2.094\ 552\ 5$
	$x_6 = 2.094\ 551\ 363$
III	IV
$y_0 = 2.1$	$y_0 = 2.1$
$x_1 = 2.093\ 9$	$x_1 = 2.094\ 563\ 28$
$y_1 = 2.094\ 551\ 72$	$y_1 = 2.094\ 551\ 481\ 162\ 069\ 454\ 26$
$x_2 = 2.094\ 551\ 481\ 367\ 28$	$x_2 = 2.094\ 551\ 481\ 542\ 326\ 591\ 9$

In the accompanying table for Newton's equation, $x^3 - 2x - 5 = 0$, already treated in Chapter 3, we give in column I the three values x_1, x_2, x_3 obtained by the Newton–Raphson formula and in column II the six values of the x_v obtained by using the simplified formula and by replacing $f'(x_v)$ by $f'(x_0)$. In both cases x_0 is 2. Comparing the values obtained with the value of ζ , we see that in column II at each step the error is only about 1/10 of the preceding error.

2. On the other hand, it may still present an advantage to compute $f'(x_v)$ not at every step but *only at every second step*. This rule can be interpreted in the following way: from the v th approximation x_v of ζ we obtain the next approximation x_{v+1} by taking

$$y_v = x_v - \frac{f(x_v)}{f'(x_v)}, \quad x_{v+1} = y_v - \frac{f(y_v)}{f'(x_v)}. \quad (\text{G.1})$$

To discuss the rapidity of the convergence of the sequence x_v we assume that x_v (and therefore y_v) tend to ζ and $f'(\zeta) \neq 0$. We have then from (6.9), replacing there x_0 by x_v and x_1 by y_v ,

$$\frac{y_v - \zeta}{(\zeta - x_v)^2} \rightarrow \frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)}. \quad (\text{G.2})$$

On the other hand, since for $y_v \rightarrow \zeta$

$$f(y_v) = (y_v - \zeta) f'(\zeta) + O[(y_v - \zeta)^2],$$

we have from (G.1) and (G.2)

$$f'(x_v)(x_{v+1} - \zeta) = f'(x_v)(y_v - \zeta) - (y_v - \zeta) f'(\zeta) + O[(y_v - \zeta)^2],$$

$$\begin{aligned} f'(x_v) \frac{x_{v+1} - \zeta}{y_v - \zeta} &= f'(x_v) - f'(\zeta) + O(y_v - \zeta) \\ &= f''(\zeta)(x_v - \zeta) + O[(x_v - \zeta)^2], \quad \zeta \in \langle x_v, \zeta \rangle, \\ f'(x_v) \frac{x_{v+1} - \zeta}{(y_v - \zeta)(x_v - \zeta)} &\rightarrow f''(\zeta); \end{aligned}$$

therefore, since $f'(\zeta) \neq 0$,

$$\frac{x_{v+1} - \zeta}{(y_v - \zeta)(x_v - \zeta)} \rightarrow \frac{f''(\zeta)}{f'(\zeta)}, \quad (\text{G.3})$$

and finally from (G.2)

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^3} \rightarrow \frac{1}{2} \left(\frac{f''(\zeta)}{f'(\zeta)} \right)^2. \quad (\text{G.4})$$

3. Now observe that the computation by (G.1) requires *three horners*. Therefore, “at the price” of six horners, we move from x_v to x_{v+2} , where

$$\frac{x_{v+2} - \zeta}{(x_v - \zeta)^9} \rightarrow \frac{1}{16} \left(\frac{f''(\zeta)}{f'(\zeta)} \right)^8. \quad (\text{G.5})$$

On the other hand, if we use the formula (6.3) three times, we obtain “at the

price" of six horners x_{v+3} , where

$$\frac{(x_{v+3} - \zeta)}{(x_v - \zeta)^8} \rightarrow \left(\frac{f''(\zeta)}{2f'(\zeta)} \right)^7. \quad (\text{G.6})$$

We see that our new rule (G.1) is indeed an improvement of the rule (6.3), although of course the necessity of computing $f'(x_v)$ with more decimals than in the classical Newton-Raphson method is a drawback.

In column III of the table on p. 306, we give the values y_0, y_1 and x_1, x_2 , obtained by this method, starting from $x_0 = 2$. Here the error in x_2 is of the same order of magnitude as that of x_3 in column I, although from formulas (G.5) and (G.6) one could expect a smaller error in column III. In this case, however, the expression $f''(\zeta)/f'(\zeta)$ is about 1.2 and the factor $8f''(\zeta)/f'(\zeta)$ just makes up in the formula (G.6) for the one factor $x_v - \zeta$ missing in (G.6), as long as $x_v - \zeta$ is not much smaller.

4. On the other hand, we can try to reduce the number of horners in the Newton-Raphson formula by replacing at every second step the denominator $f'(x_v)$ by a convenient combination of $f(x_v)$ and $f(x_{v-1})$.

We shall now discuss the following rule, which can be used in this direction. Starting with x_0 , put

$$y_v = x_v - \frac{f(x_v)}{f'(x_v)}, \quad x_{v+1} = y_v - \frac{f(y_v)(y_v - x_v)}{2f(y_v) - f(x_v)}. \quad (\text{G.7})$$

We see that we pass from x_v to x_{v+1} using three horners. On the other hand, we shall show that if x_v tends to a zero ζ of $f(x)$, we have

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^4} \rightarrow \frac{1}{24} \frac{f''(\zeta)}{f'(\zeta)^3} [3f''(\zeta)^2 - 2f'(\zeta)f'''(\zeta)], \quad (\text{G.8})$$

assuming again $f'(\zeta) \neq 0$. Here we obtain with three horners the same order of improvement which is obtained by two consecutive Newton-Raphson steps, i.e., by using four horners.

5. Rule (G.7) can be obtained from (11.16) if we choose there x_1 as given by the formula $x_1 = x_0 - f_0/f'_0$. Then the expression Δ^* in (11.7) becomes

$$\frac{-1}{f'_0} \frac{f_0 - f_1}{(-f_0/f'_0)} = \frac{f_0 - f_1}{f_0}$$

and we get from (11.16)

$$x_2 - x_1 = \frac{f_1(x_1 - x_0)}{f_0 \Delta^* - f_1} = -\frac{f_1(x_1 - x_0)}{2f_1 - f_0}.$$

Replacing here x_0 by x_v , x_1 by y_v , and x_2 by x_{v+1} , we get (G.7).

6. Applying then formula (11.30), we have

$$\frac{\zeta - x_{v+1}}{(x_v - \zeta)^2(y_v - \zeta)} \rightarrow \frac{1}{4}f''(\zeta)^2f'(\zeta)^{-2} - \frac{1}{6}f^{(3)}(\zeta)f'(\zeta)^{-1}. \quad (\text{G.9})$$

But here we have by (6.9) $y_v - \zeta \sim (x_v - \zeta)^2f''(\zeta)/2f'(\zeta)$ and from (G.9) follows immediately the relation (G.8).

7. In column IV of the table on p. 306, we give the values of the y_0, y_1 and x_1, x_2 computed by formula (G.7) starting from $x_0 = 2$. We see that, using six horners, we get in column IV a very much better result than at the same "price" in column I. On the other hand, we must not forget that when computing by formula (G.7) we must use the double number of decimals much sooner, and furthermore the self-correcting property of the Newton-Raphson formula does not come into play in the same way.

8. It may be remarked finally that we did not discuss in this chapter the question of convergence criteria for the modified forms of the Newton-Raphson method. Our asymptotic formulas make it clear that these methods converge if we start in a sufficiently close neighborhood of ζ . But this neighborhood need not be the same as in the original Newton-Raphson method.

H

Rounding Off in Inverse Interpolation

1. If we use repeated inverse interpolation with a fixed number n of points at every step, as discussed in Chapter 13, Section 2, the error of x_{n+1} is $O(\prod_{v=1}^n |\zeta - x_v|)$, by (13.6). The precision of the value of $y_v = f(x_v)$ therefore must be such that this theoretical error of x_{n+1} will not be worsened. Now, denote the expression to the right in (13.1) as function of $x_1, \dots, x_n; y_1, \dots, y_n$ by $X = X(x_1, \dots, x_n; y_1, \dots, y_n)$. Denote the error made in computing y_v by δ_v ; then the resulting error of x_{n+1} is

$$\sum_{v=1}^n \delta_v \frac{\partial X}{\partial y_v},$$

where the second factors are to be taken at the point $(x_1, \dots, x_n; \theta_1 y_1, \dots, \theta_n y_n)$, $|\theta_v| \leq 1$ ($v = 1, \dots, n$). Since the x_v tend to ζ and y_v to 0, it is essential to know the order of magnitude of $\partial X / \partial y_v$ for $y_v \rightarrow 0$, $x_v \rightarrow \zeta$ ($v = 1, \dots, n$).

2. In this discussion we can without loss of generality assume from the beginning $\zeta = 0$ and $f'(0) = 1$, since we can always subtract ζ from x and multiply $f(x)$ by a given constant. We then have $y_v/x_v \rightarrow 1$. On the other hand, since each x_{v+1} is an essentially better approximation than x_v , it is reasonable to make the hypothesis $y_{v+1}/y_v \rightarrow 0$. Then a partial solution of our problem is given by the following lemma.

3. Lemma 1. Consider for an integer $n \geq 2$, $2n$ quantities x_v, y_v ($v = 1, \dots, n$) which tend to zero in such a way that we have

$$\frac{x_v}{y_v} \rightarrow 1 \quad (v = 1, \dots, n), \quad \frac{y_{v+1}}{y_v} \rightarrow 0 \quad (v = 1, \dots, n-1). \quad (\text{H.1})$$

Put $F(y) \equiv \prod_{v=1}^n (y - y_v)$ and form

$$X := (-1)^{n-1} y_1 \cdots y_n \sum_{v=1}^n \frac{x_v}{y_v} \frac{1}{F'(y_v)} \quad (\text{H.2})$$

and for $1 \leq k \leq n$

$$D_k := \frac{1}{y_1 \cdots y_n} \frac{\partial X}{\partial y_k} \quad (1 \leq k \leq n). \quad (\text{H.3})$$

Then we have

$$D_n \sim \frac{-1}{y_1 \cdots y_n}, \quad (\text{H.4})$$

$$D_{n-1} \sim \frac{1}{y_1 \cdots y_{n-2} y_{n-1}^2},$$

$$D_{n-2} \sim \frac{-1}{y_1 \cdots y_{n-3} y_{n-2}^3}, \quad (n \geq 3), \quad (\text{H.5})$$

$$D_k = o\left(\frac{1}{y_1 \cdots y_{n-2} y_k^2}\right) \quad (1 \leq k < n-2, \quad n \geq 4). \quad (\text{H.6})$$

4. Proof. For a fixed k , $1 \leq k \leq n$, put

$$F(y) := (y - y_k) G(y), \quad G(y) = \prod_{\substack{v=1 \\ v \neq k}}^n (y - y_v). \quad (\text{H.7})$$

Then we have from (H.1)

$$G(y_k) \sim (-1)^{k-1} y_1 \cdots y_{k-1} y_k^{n-k}. \quad (\text{H.8})$$

If we now form the logarithmic derivative, we have for $k < n$

$$\begin{aligned} \frac{G'(y_k)}{G(y_k)} &= \sum_{v=1}^{k-1} \frac{1}{y_k - y_v} + \sum_{v=k+1}^n \frac{1}{y_k - y_v} \\ &= - \sum_{v=1}^{k-1} \frac{1 + o(1)}{y_v} + \frac{1}{y_k} \sum_{v=k+1}^n [1 + o(1)], \\ y_k \frac{G'(y_k)}{G(y_k)} &\rightarrow n - k, \end{aligned}$$

and for $k = n$

$$\frac{G'(y_n)}{G(y_n)} = - \sum_{v=1}^{n-1} \frac{1}{y_v} \frac{1}{1 - y_n/y_v} \sim \frac{-1}{y_{n-1}};$$

therefore

$$\frac{G'(y_k)}{G(y_k)} \sim \begin{cases} \frac{n-k}{y_k} & (k < n) \\ \frac{-1}{y_{n-1}} & (k = n), \end{cases}$$

and by (H.8)

$$\frac{G'(y_k)}{G(y_k)^2} \sim \begin{cases} (-1)^{k-1} \frac{n-k}{y_1 \cdots y_{k-1} y_k^{n-k+1}} & (k < n) \\ \frac{(-1)^n}{y_1 y_2 \cdots y_{n-2} y_{n-1}^2} & (k = n). \end{cases} \quad (\text{H.9})$$

5. On the other hand, for a $v \neq k$ we have

$$G'(y_v) = \prod_{\substack{\sigma=1 \\ \sigma \neq k, v}}^n (y_v - y_\sigma),$$

and therefore

$$G'(y_\mu) \sim (-1)^{\mu-1} y_1 \cdots y_{\mu-1} y_\mu^{n-\mu-1} \quad (\mu < k), \quad (\text{H.10})$$

$$G'(y_\lambda) \sim (-1)^\lambda \frac{y_1 \cdots y_{\lambda-1} y_\lambda^{n-\lambda}}{y_k} \quad (\lambda > k). \quad (\text{H.11})$$

We have further from (H.2)

$$X = \sum_{v=1}^n U_v, \quad U_v = (-1)^{n-1} y_1 \cdots y_n \frac{x_v}{y_v} \frac{1}{F'(y_v)},$$

where in particular, by virtue of (H.7),

$$\begin{aligned} U_k &= (-1)^{n-1} y_1 \cdots y_n \frac{x_k}{y_k} \frac{1}{G(y_k)}, \\ U_v &= (-1)^{n-1} y_1 \cdots y_n \frac{x_v}{y_v} \frac{1}{G'(y_v)(y_v - y_k)} \quad (v \neq k). \end{aligned}$$

We have, therefore, differentiating with respect to y_k ,

$$\frac{\partial U_k}{\partial y_k} = y_1 \cdots y_n \frac{x_k}{y_k} \frac{(-1)^n G'(y_k)}{G(y_k)^2}$$

and for $v \neq k$

$$\frac{\partial U_v}{\partial y_k} = \frac{y_1 \cdots y_n}{y_k} \frac{x_v}{y_v} \frac{(-1)^{n-1}}{G'(y_v)} \frac{\partial}{\partial y_k} \frac{y_k}{y_k - y_v} = \frac{y_1 \cdots y_n}{y_k} \frac{x_v}{y_v} \frac{(-1)^{n-1}}{G'(y_v)} \frac{y_v}{(y_v - y_k)^2};$$

therefore

$$\begin{aligned} D_k &= \frac{1}{y_1 \cdots y_n} \frac{\partial X}{\partial y_k} = \sum_{v=1}^n T_v, \\ T_k &:= \frac{(-1)^n x_k G'(y_k)}{y_k G(y_k)^2}, \quad T_v := \frac{(-1)^{n-1} x_v}{y_k G'(y_v)(y_v - y_k)^2} \quad (v \neq k). \end{aligned} \quad (\text{H.12})$$

6. But now it follows from (H.9) that

$$\begin{aligned} T_k &\sim \frac{(-1)^{n-k+1}(n-k)}{y_1 \cdots y_{k-1} y_k^{n-k+1}} \quad (k < n), \\ T_k &\sim \frac{1}{y_1 y_2 \cdots y_{n-2} y_{n-1}^2} \quad (k = n), \end{aligned} \tag{H.13}$$

and from (H.10) and (H.11) for $\mu < k$ and $\lambda > k$:

$$\begin{aligned} T_\mu &\sim \frac{(-1)^{n-\mu-1} y_\mu}{y_k (-1)^{\mu-1} y_1 \cdots y_{\mu-1} y_\mu^{n-\mu-1} y_\mu^2} \\ &= \frac{(-1)^{n-\mu}}{y_1 \cdots y_{\mu-1} y_\mu^{n-\mu} y_k} \quad (\mu < k), \end{aligned} \tag{H.14}$$

$$\begin{aligned} T_\lambda &\sim \frac{(-1)^{n-\lambda-1} y_\lambda y_k}{y_k (-1)^\lambda y_1 \cdots y_{\lambda-1} y_\lambda^{n-\lambda-1} y_k^2} \\ &= \frac{(-1)^{n-\lambda-1}}{y_1 \cdots y_{\lambda-1} y_\lambda^{n-\lambda-1} y_k^2} \quad (\lambda > k). \end{aligned} \tag{H.15}$$

From (H.14) and (H.13) we have for $k < n$

$$\frac{T_\mu}{T_k} \sim \frac{(-1)^{k-\mu-1}}{n-k} \left(\frac{y_\mu}{y_\mu} \right) \left(\frac{y_{\mu+1}}{y_\mu} \right) \cdots \left(\frac{y_{k-1}}{y_\mu} \right) \left(\frac{y_k}{y_\mu} \right)^{n-k},$$

and this tends to 0; therefore,

$$\frac{T_\mu}{T_k} \rightarrow 0 \quad (\mu < k < n). \tag{H.16}$$

Further, we have from (H.13), (H.14), and (H.15), whether $k = n$, or $k = n-1$, or $k \leq n-2$,

$$\frac{T_n}{T_{n-1}} \sim -\frac{y_n}{y_{n-1}} \rightarrow 0. \tag{H.17}$$

7. We have on the other hand from (H.14) and (H.13), for $k = n$ and $\mu < n-1$,

$$\frac{T_\mu}{T_{n-1}} \sim (-1)^{n-\mu-1} \left(\frac{y_\mu}{y_\mu} \right) \left(\frac{y_{\mu+1}}{y_\mu} \right) \cdots \left(\frac{y_{n-1}}{y_\mu} \right) \rightarrow 0 \quad (\mu < n-1 < k).$$

We see now from this relation and from (H.17) that for $k = n$ the term T_{n-1}

in (H.12) dominates and therefore by (H.14)

$$D_n \sim T_{n-1} \sim \frac{-1}{y_1 \cdots y_n};$$

the first relation (H.4) is proved.

8. For $k = n-1$, we see by (H.16) and (H.17) that again the term T_{n-1} dominates and therefore from (H.13) with $k = n-1$

$$D_{n-1} \sim T_{n-1} \sim \frac{1}{y_1 \cdots y_{n-2} y_{n-1}^2};$$

the second relation (H.4) is proved.

9. For $k = n-2$ we have now from (H.15) with $\lambda = n-1$ and from (H.13) with $k = n-2$

$$\frac{T_{n-2}}{T_{n-1}} \sim \frac{(-1)^3 2 y_1 \cdots y_{n-2} y_{n-2}^2}{y_1 \cdots y_{n-3} y_{n-2}^3} = -2,$$

$T_k/T_{n-1} \rightarrow -2$, and it follows now from (H.17) and (H.16) that

$$D_{n-2} \sim -T_{n-1} \sim \frac{-1}{y_1 \cdots y_{n-3} y_{n-2}^3},$$

which proves (H.5).

10. We assume now $k < n-2$. Then we have from (H.13) and (H.15) with $\lambda = n-1$

$$\frac{T_k}{T_{n-1}} \sim (-1)^{n-k+1} (n-k) \frac{y_k}{y_k} \frac{y_{k+1}}{y_k} \cdots \frac{y_{n-2}}{y_k} \rightarrow 0, \quad (\text{H.18})$$

and from (H.15) for $\lambda < n-1$

$$\begin{aligned} \frac{T_\lambda}{T_{n-1}} &\sim \frac{(-1)^{n-\lambda+1} y_1 \cdots y_{n-2} y_k^2}{y_1 \cdots y_{\lambda-1} y_\lambda^{n-\lambda-1} y_k^2} \\ &= (-1)^{n-\lambda+1} \frac{y_\lambda}{y_\lambda} \frac{y_{\lambda+1}}{y_\lambda} \cdots \frac{y_{n-2}}{y_\lambda} \quad (\lambda < n-1). \end{aligned}$$

But this tends to 0 for $\lambda < n-2$ and is -1 for $\lambda = n-2$. It follows now from (H.16), (H.17), and (H.18) that

$$T_v = o(T_{n-1}) \quad (v < n-2 \text{ or } v = n)$$

and

$$T_{n-2} + T_{n-1} = o(T_{n-1}),$$

and by (H.12)

$$D_k = o(T_{n-1}) = o\left(\frac{1}{y_1 \cdots y_{n-2} y_k^2}\right) \quad (k < n-2);$$

our lemma is proved.

11. Now, using formulas (H.3)–(H.6), we see that in order to make

$$\sum_{v=1}^n \delta_v \frac{\partial X}{\partial y_v} = O(y_1 \cdots y_n)$$

it is sufficient to make (a)

$$\delta_n = O(y_1 \cdots y_n),$$

$$\delta_{n-1} = O(y_1 \cdots y_{n-2} y_{n-1}^2),$$

$$\delta_{n-2} = O(y_1 \cdots y_{n-3} y_{n-2}^3),$$

and (b)

$$\delta_k = O(y_1 \cdots y_{n-2} y_k^2) \quad (k = 1, \dots, n-3).$$

Here, the results contained in (a) are very satisfactory. Indeed, the y_v used in these estimates can be considered as already known at the moment the decision about the corresponding δ_v is made. On the contrary, the estimates (b) do not give the “true” order of magnitude necessary for δ_k and require the knowledge of y_v with $v > k$.

However, using the further hypothesis that $f^{(n-2)}(x)$ exists and is continuous in J_x , we can obtain for δ_k with $k < n-2$ estimates which give the “best” order of magnitude and depend only on y_1, \dots, y_k . Of course, the need for these estimates arises only for $n \geq 4$. Since under our new hypothesis the $(n-2)$ nd derivative of the inverse function of $w = f(x)$, $x = \varphi(w)$ exists and is continuous in the neighborhood of the origin, we have, in using $\varphi'(0) = 1$,

$$\varphi(w) = w + \sum_{v=2}^{n-3} a_v w^v + O(w^{n-2}),$$

and therefore for $w = y_\lambda$

$$x_\lambda = y_\lambda + \sum_{v=2}^{n-3} a_v y_\lambda^v + O(y_\lambda^{n-2}) \quad (1 < \lambda \leq n-3).$$

But then the following lemma gives a solution of our problem.

12. Lemma 2. Under the hypotheses and in the notation of Lemma 1, suppose that we have $1 \leq k < n-2$, $n \geq 4$; further, for each λ with $k < \lambda \leq n$

$$x_\lambda = y_\lambda + \sum_{v=2}^{n-k-2} a_v y_\lambda^v + O(y_\lambda^{n-k-1}) \quad (k < \lambda \leq n) \quad (\text{H.19})$$

with certain constants a_v independent of λ . Then we have instead of (H.6)

$$D_k \sim \frac{(-1)^{n-k+1}}{y_1 \cdots y_{k-1} y_k^{n-k+1}} \quad (1 \leq k \leq n-3). \quad (\text{H.20})$$

13. Proof. We begin by establishing certain relations following from the hypotheses of Lemma 1. Putting $\eta = y_k$, we have from (H.11) for $\lambda > k$ for any integer v

$$\frac{y_\lambda^v}{\eta(\eta-y_\lambda)^2 G'(y_\lambda)} \sim \frac{(-1)^\lambda y_\lambda^{v-n+\lambda}}{\eta^2 y_1 \cdots y_k y_{k+1} \cdots y_{\lambda-1}}.$$

The modulus of the right-hand expression is not decreased if the factors $y_{k+1}, \dots, y_{\lambda-1}$ are replaced by y_λ , and we obtain

$$\frac{y_\lambda^v}{\eta(\eta-y_\lambda)^2 G'(y_\lambda)} = O\left(\frac{y_\lambda^{v+1-(n-k)}}{y_1 \cdots y_k \eta^2}\right) = O\left(\frac{\eta^{v-1-(n-k)}}{y_1 \cdots y_k}\right),$$

if we assume $v+1 \geq n-k$. Dividing this by T_k and using the first equivalence (H.13), we obtain

$$\frac{y_\lambda^v}{\eta(\eta-y_\lambda)^2 G'(y_\lambda)} = O(\eta^{v-1} T_k) \quad (v+1 \geq n-k, \quad k < \lambda \leq n). \quad (\text{H.21})$$

14. In what follows we shall use the decomposition $G(y) = G_1(y)G_2(y)$, with

$$G_1(y) = \prod_{\mu=1}^{k-1} (y-y_\mu), \quad G_2(y) = \prod_{\lambda=k+1}^n (y-y_\lambda).$$

We have obviously

$$G'(y_\lambda) = G_1(y_\lambda) G_2'(y_\lambda) \quad (k < \lambda \leq n)$$

and

$$G_1(y_\lambda) \sim (-1)^{k-1} y_1 \cdots y_{k-1}, \quad G_2(\eta) \sim \eta^{n-k},$$

$$\eta \frac{G_2'(\eta)}{G_2(\eta)} = \sum_{\lambda=k+1}^n \frac{\eta}{\eta-y_\lambda} \rightarrow n-k. \quad (\text{H.22})$$

Introducing the equivalence for $G_1(y_\lambda)$ into (H.21), we obtain further

$$\frac{y_\lambda^v}{(\eta-y_\lambda)^2 G_2'(y_\lambda)} = O(y_1 \cdots y_k \eta^{v-1} T_k) \quad (v+1 \geq n-k, \quad k < \lambda \leq n). \quad (\text{H.23})$$

15. Consider now the expression

$$K_v := \sum_{\lambda=k+1}^n \frac{y_\lambda^v}{(\eta-y_\lambda)^2 G_2'(y_\lambda)} = -\frac{\partial}{\partial \eta} \frac{1}{G_2(\eta)} \sum_{\lambda=k+1}^n \frac{y_\lambda^v G_2(\eta)}{(\eta-y_\lambda) G_2'(y_\lambda)}. \quad (\text{H.24})$$

For $1 \leq v < n-k$ the last sum to the right is by Lagrange's interpolation formula identical to η^v , and we have therefore by (H.22)

$$K_v = -\frac{\partial}{\partial \eta} \frac{\eta^v}{G_2(\eta)} = \frac{\eta^{v-1}}{G_2(\eta)} \left[\eta \frac{G_2'(\eta)}{G_2(\eta)} - v \right] \sim \frac{n-k-v}{\eta^{n-k+1-v}},$$

and in particular

$$K_1 \sim (n-k-1) \eta^{k-n}, \quad K_v = O(\eta^{v-(n-k)-1}) \quad (v > 1).$$

Dividing by T_k and using (H.13), we obtain finally

$$\frac{K_1}{T_k} \sim (-1)^{n-k-1} \frac{n-k-1}{n-k} y_1 \cdots y_k, \quad (\text{H.25})$$

$$\frac{K_v}{T_k} = O(y_1 \cdots y_k \eta^{v-1}) \quad (v > 1). \quad (\text{H.26})$$

The last estimate has been deduced for $v < n-k$. It also remains valid, however, for $v \geq n-k$ by virtue of (H.23) and (H.24).

16. We now introduce the expressions

$$S_v := (-1)^{n-1} \sum_{\lambda=k+1}^n \frac{y_\lambda^v}{\eta(\eta-y_\lambda)^2 G'(\eta)} \quad (v = 1, \dots, n-k-2). \quad (\text{H.27})$$

We have from (H.12) and (H.16)

$$D_k = T_k + \sum_{\lambda=k+1}^n T_\lambda + o(T_k).$$

Introducing in the expressions for T_λ from (H.12) the values of x_λ from (H.19), we have further

$$D_k = T_k + S_1 + \sum_{v=2}^{n-k-2} a_v S_v + \sum_{\lambda=k+1}^n O\left(\frac{y_\lambda^{n-k-1}}{\eta G'(\eta)(\eta-y_\lambda)^2}\right) + o(T_k).$$

where the first right-hand sum is to be left out for $n=4$, $k=1$. Each term of the second right-hand sum is $o(T_k)$, by (H.21) applied for $v=n-k-1$, since $n-k-1 > 1$. We have therefore

$$D_k = T_k + S_1 + \sum_{v=2}^{n-k-2} a_v S_v + o(T_k). \quad (\text{H.28})$$

17. We consider first the case where $k=1$. In this case we have from (H.27) and (H.24), since $G_2(y) = G(y)$,

$$\frac{S_v}{T_k} = (-1)^{n-1} \frac{K_v}{\eta T_k},$$

and we have therefore from (H.25) and (H.26) for $k = 1$ the relations

$$S_v = o(T_k) \quad (v > 1), \quad (\text{H.29})$$

$$S_1 \sim -\left(1 - \frac{1}{n-k}\right)T_k. \quad (\text{H.30})$$

We are now going to prove that (H.29) and (H.30) also hold for $2 \leq k < n-2$.

18. It follows from (H.27) that

$$(-1)^{n-k} y_1 \cdots y_k S_v = \sum_{\lambda=k+1}^n \prod_{\mu=1}^{k-1} \frac{1}{1-y_\lambda/y_\mu} \frac{y_\lambda^v}{G_2'(y_\lambda)(\eta-y_\lambda)^2}. \quad (\text{H.31})$$

On the other hand, we have for each $\lambda > k$ and each $\mu < k$

$$\frac{1}{1-y_\lambda/y_\mu} = \sum_{\sigma=0}^n y_\mu^{-\sigma} y_\lambda^\sigma + O(y_\lambda^{n+1} y_\mu^{-n-1}),$$

where the coefficient of each power $y_\lambda^\sigma (\sigma > 0)$ is $o(\eta^{-\sigma})$. Therefore, if we multiply over all $\mu < k$,

$$\prod_{\mu=1}^{k-1} \frac{1}{1-y_\lambda/y_\mu} = 1 + \sum_{\sigma=1}^{n+1} F_\sigma y_\lambda^\sigma, \quad (\text{H.32})$$

$$F_\sigma = o(\eta^{-\sigma}) \quad (\sigma = 1, 2, \dots, n+1). \quad (\text{H.33})$$

Introducing (H.32) into (H.31) and using (H.24), we have

$$(-1)^{n-k} y_1 \cdots y_k S_v = K_v + \sum_{\sigma=1}^{n+1} F_\sigma K_{v+\sigma}. \quad (\text{H.34})$$

But by virtue of (H.26), each term of the right-hand sum is equal to

$$o(\eta^{-\sigma} y_1 \cdots y_k \eta^{v+\sigma-1} T_k) = o(y_1 \cdots y_k \eta^{v-1} T_k).$$

19. It follows now from (H.34) and (H.26) that

$$S_v = o(T_k) \quad (v > 1),$$

$$S_1 = (-1)^{n-k} \frac{K_1}{y_1 \cdots y_k} + o(T_k),$$

and by (H.25) the formulas (H.29) and (H.30) now follow in the general case. But now (H.20) follows immediately from (H.28), (H.30), and (H.13), and Lemma 2 is proved.

20. By virtue of Lemma 2, we now obtain for all δ_k ($1 \leq k \leq n$) the conditions (c)

$$\delta_k = O(y_1 \cdots y_{k-1} y_k^{n-k+1}) \quad (1 \leq k \leq n),$$

which can be considered as completely satisfactory in the same sense as the conditions (a) in Section 11.

Accelerating Iterations with Supralinear Convergence

1. We consider in what follows a sequence z_v ($v = 1, 2, \dots$) convergent to ζ and assume that for a constant $s > 1$ we have

$$\frac{|z_{v+1} - \zeta|}{|z_v - \zeta|^s} \rightarrow \alpha \quad (v \rightarrow \infty, \quad \alpha \neq 0, \quad \alpha \neq \infty). \quad (\text{I.1})$$

Under these assumptions we shall show that if we want to stop at z_{n+1} , the approximation to ζ will be improved if z_{n+1} is replaced by

$$Z := z_{n+1} - \frac{|z_n - z_{n+1}|^{s+1}}{|z_{n-1} - z_n|^s} \operatorname{sgn}(z_{n+1} - \zeta). \quad (\text{I.2})$$

If more is known about the rapidity of the convergence in (I.1), the improvement obtained by (I.2) will be specified accordingly.

This result can of course be applied if the sequence z_v is given by an *iteration formula of the first order*,

$$z_{v+1} = \varphi(z_v),$$

or more generally, if z_{v+1} is obtained by an *iteration of the order k*,

$$z_{v+1} = \varphi(z_v, z_{v-1}, \dots, z_{v-k+1}).$$

But this special assumption is not necessary for the validity of our results.

2. We can obviously write

$$|z_{n+1} - \zeta| = \alpha |z_n - \zeta|^s (1 + \varepsilon_n), \quad |z_n - \zeta| = \alpha |z_{n-1} - \zeta|^s (1 + \varepsilon_{n-1}), \quad (\text{I.3})$$

where as $n \rightarrow \infty$ we have $\varepsilon_n \rightarrow 0$, $\varepsilon_{n-1} \rightarrow 0$. Then put

$$\Delta_n := \operatorname{Max}(|\varepsilon_n|, |\varepsilon_{n-1}|, |z_{n-1} - \zeta|^{s-1}). \quad (\text{I.4})$$

We shall prove that, while the approximation of z_{n+1} is characterized by

$$\frac{|z_{n+1} - \zeta|}{\alpha^{s+1} |z_{n-1} - \zeta|^{s^2}} \rightarrow 1 \quad (n \rightarrow \infty), \quad (\text{I.5})$$

we have for Z

$$\frac{|Z - \zeta|}{\alpha^{s+1} |z_{n-1} - \zeta|^{s^2}} = O(\Delta_n) \quad (n \rightarrow \infty). \quad (\text{I.6})$$

3. Putting

$$|z_{n-1} - \zeta| = \delta,$$

we have by (I.3) and (I.4) as $n \rightarrow \infty$

$$\begin{aligned} \frac{|z_{n-1} - z_n|}{\delta} &= 1 + O\left(\frac{z_n - \zeta}{\delta}\right) = 1 + O(\delta^{s-1}) = 1 + O(\Delta_n), \\ |z_{n-1} - z_n|^s &= \delta^s [1 + O(\Delta_n)]. \end{aligned} \quad (\text{I.7})$$

Further, again by (I.3) and (I.4),

$$\begin{aligned} |z_n - z_{n+1}| &= |z_n - \zeta| \left| 1 - \frac{z_{n+1} - \zeta}{z_n - \zeta} \right| = \alpha \delta^s [1 + O(|z_n - \zeta|^{s-1})] (1 + \varepsilon_{n-1}) \\ &= \alpha \delta^s [1 + O(\delta^{s-1})] [1 + O(\Delta_n)], \\ |z_{n+1} - z_n| &= \alpha \delta^s [1 + O(\Delta_n)]. \end{aligned} \quad (\text{I.8})$$

From (I.7) and (I.8) we have

$$\frac{|z_{n+1} - z_n|^{s+1}}{|z_n - z_{n-1}|^s} = \alpha^{s+1} \delta^{s^2} [1 + O(\Delta_n)]. \quad (\text{I.9})$$

4. On the other hand, applying (I.3) twice, we obtain

$$|z_{n+1} - \zeta| = \alpha |z_n - \zeta|^s [1 + O(\Delta_n)] = \alpha^{s+1} \delta^{s^2} [1 + O(\Delta_n)]. \quad (\text{I.10})$$

From (I.10), (I.5) follows immediately. Put

$$s_{n+1} := \operatorname{sgn}(z_{n+1} - \zeta);$$

then we have from (I.9) and (I.10), respectively,

$$\begin{aligned} s_{n+1} \frac{|z_n - z_{n+1}|^{s+1}}{|z_{n-1} - z_n|^s} &= s_{n+1} \alpha^{s+1} \delta^{s^2} [1 + O(\Delta_n)], \\ z_{n+1} - \zeta &= s_{n+1} \alpha^{s+1} \delta^{s^2} [1 + O(\Delta_n)]. \end{aligned}$$

Dividing these formulas by $\alpha^{s+1} \delta^{s^2}$ and subtracting, we get by (1.2)

$$\frac{Z - \zeta}{\alpha^{s+1} \delta^{s^2}} = O(\Delta_n),$$

and this is (I.6).

5. If in particular we have for the $\varepsilon_{n-1}, \varepsilon_n$ defined by (I.3),

$$|\varepsilon_{n+1}| + |\varepsilon_n| = O(\delta^p), \quad p > 1, \quad (\text{I.11})$$

and if we put

$$\text{Min}(p, s-1) = d, \quad (\text{I.12})$$

we have obviously $\Delta_n = O(\delta^d)$ and we can replace (I.6) by

$$Z = \zeta = O(\delta^{s^2+d}). \quad (\text{I.13})$$

Usually we have $d = 1$, and then the use of (I.2) gives an improvement of 25% for $s = 2$ and 11.1% for $s = 3$.

6. In order to use the approximation (I.2), the value of $\text{sgn}(z_{n+1} - \zeta)$ must be known. In many cases it can be considered as known, namely, when in the case of real z , and integer s the limit of

$$\frac{z_{v+1} - \zeta}{(z_v - \zeta)^s}$$

exists and is $\neq 0, \infty$. Consider, for example, in the case of the Newton–Raphson method, the relation (6.9), from which it follows that

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^2} \rightarrow \frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)}.$$

Here, if the x_v are real, the approximation (I.2) can be used indeed, not only at the end of the computation, but also *after each step* of the Newton–Raphson method. In this case we have (I.11)–(I.13) with $d = 1$.

An analogous remark applies also to the Schröder method dealt with in Chapter 8 and to the rules discussed in Appendix G when the formulas (G.4) and (G.8) are used.

In the case of the *regula falsi*, we can use (3.19) in the case of real x , and the signs of the successive differences $x_v - \zeta$ are easily obtained from (3.11). However, as by (12.32) ε_v and ε_{v-1} are only $O(t_2^v) = O(1/\ln \delta_v)$, the improvement is not very considerable. The same is true in the case of the iteration considered in Chapter 5.

It can be said generally that approximation (I.2) can be used in the case of an iteration $z_{v+1} = \varphi(z_v)$ of the first order, since here the determination of $\text{sgn}(z_v - \zeta)$ usually does not present difficulties; in this case we can even use (I.2) at every step of iteration. This is in many cases also true for an iteration of finite order k ,

$$z_{v+k} = \varphi(z_v, z_{v+1}, \dots, z_{v+k-1}).$$

But even in the most general case of an arbitrary sequence z_v , the use of (I.2) at the final step of the computation can still give an appreciable improvement.

In the above discussion it was necessary to use three consecutive approximations z_{n-1}, z_n, z_{n+1} in order to eliminate α , since α is not supposed as known.

In the case where we know α from theoretical discussion, the above formula can be considerably improved. Assume that we have

$$\frac{|z_{v+1} - \zeta|}{|z_v - \zeta|^s} = \alpha + O(|z_v - \zeta|), \quad s > 1, \quad (\text{I.14})$$

where $z_v \rightarrow \zeta$ and $\alpha \neq 0, \alpha \neq \infty$. Then we have from (I.14), as

$$\begin{aligned} \frac{z_v - \zeta}{z_v - z_{v+1}} - 1 &= \frac{z_{v+1} - \zeta}{z_v - \zeta} \left(1 - \frac{z_{v+1} - \zeta}{z_v - \zeta}\right)^{-1} = O(|z_v - \zeta|^{s-1}); \\ |z_{v+1} - \zeta| &= \alpha |z_v - z_{v+1}|^s (1 + O(|z_v - \zeta|)) (1 + O(|z_v - \zeta|^{s-1})), \\ z_{v+1} - \zeta &= \alpha |z_v - z_{v+1}|^s \operatorname{sgn}(z_{v+1} - \zeta) + O(|z_v - \zeta|^{\min(s+1, 2s-1)}), \\ \zeta &= z_{v+1} - \alpha |z_v - z_{v+1}|^s \operatorname{sgn}(z_{v+1} - \zeta) \\ &\quad + O(|z_v - \zeta|^{\min(s+1, 2s-1)}). \end{aligned} \quad (\text{I.15})$$

We see that the expression

$$Z^* = z_{v+1} - \alpha |z_v - z_{v+1}|^s \operatorname{sgn}(z_{v+1} - \zeta) \quad (\text{I.16})$$

gives a better approximation to ζ than z_{v+1} .

J

Roots of $f(z) = 0$ in Terms of the Coefficients of the Development of $1/f(z)$

1. In this appendix we shall discuss the equation

$$f(z) \equiv a_0 + a_1 z + \cdots = 0 \quad (a_0 = 1), \quad (\text{J.1})$$

where the right-hand expression is a power series with a radius of convergence r , $0 < r \leq \infty$; $f(z)$ is a polynomial of degree n if all a_v with $v > n$ are 0.

2. We shall use the development

$$\Phi(z) \equiv \frac{1}{f(z)} = \sum_{v=0}^{\infty} P_v z^v, \quad (\text{J.2})$$

where $k(z)$ has a radius of convergence $\rho_0 > 0$. The coefficients P_v in (J.2) can be computed recursively. Multiplying the right-hand expression in (J.2) by the development (J.1) of $f(z)$, we obtain

$$1 \equiv (a_0 + a_1 z + a_2 z^2 + \cdots)(P_0 + P_1 z + P_2 z^2 + \cdots),$$

and therefore

$$\begin{aligned} a_0 P_0 &= 1 \\ a_1 P_0 + a_0 P_1 &= 0 \\ &\vdots \\ a_v P_0 + a_{v-1} P_1 + \cdots + a_0 P_v &= 0 \\ &\vdots \end{aligned} \quad (\text{J.3})$$

From these formulas the P_v are easily obtained one after another.

If in particular $f(z)$ is a polynomial of degree n , formulas (J.3) with $v \geq n$ show that the P_v satisfy the linear difference equation (12.2) with the characteristic equation (12.3).

3. In order to solve the system (J.3) by determinants, we put

$$D_{1,v} = \begin{vmatrix} a_1 & a_2 & \cdots & a_v \\ a_0 & a_1 & \cdots & a_{v-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{2-v} & a_{3-v} & \cdots & a_1 \end{vmatrix} \quad (v = 1, 2, \dots), \quad D_{1,0} = 1, \quad (\text{J.4})$$

using the convention

$$a_{-\mu} = 0 \quad (\mu = 1, 2, \dots). \quad (\text{J.5})$$

Then we easily obtain from the first $v+1$ equations (J.3), since their determinant has the value 1, the expression for P_v due to Wronski (1811):

$$P_v = (-1)^v D_{1,v}. \quad (\text{J.6})$$

4. Assume now that (J.1) has a *simple* root ξ_1 such that $|\xi_1| = \rho_0$ is $< r$ and less than the moduli of all other roots of (J.1). Then we can write for a suitable constant $\alpha_1 \neq 0$

$$\Phi(z) = \frac{\alpha_1 \xi_1}{\xi_1 - z} + \sum_{v=0}^{\infty} b_v z^v, \quad (\text{J.7})$$

where the right-hand power series has a radius of convergence $\rho_1 > \rho_0$. We have therefore for any positive q with

$$|\xi_1| < \frac{1}{q} < \rho_1, \quad (\text{J.8})$$

$$b_v = o(q^v),$$

$$P_v = \alpha_1 \xi_1^{-v} + o(q^v), \quad (\text{J.9})$$

and from (J.9) it follows that

$$\frac{P_{v-1}}{P_v} \rightarrow \xi_1, \quad \frac{P_{v-1}}{P_v} = \xi_1 + o(|\xi_1|^v q^v). \quad (\text{J.10})$$

5. The first formula (J.10) can be written as

$$\xi_1 = \frac{P_0}{P_1} + \sum_{v=2}^{\infty} \left(\frac{P_{v-1}}{P_v} - \frac{P_{v-2}}{P_{v-1}} \right). \quad (\text{J.11})$$

The v th term of this infinite series can be easily represented (using a Sylvester

determinantal formula) in the form

$$\frac{P_{v-1}}{P_v} - \frac{P_{v-2}}{P_{v-1}} = -\frac{D_{2,v-1}}{D_{1,v-1} D_{1,v}}, \quad \dagger$$

if we put

$$D_{2,v} = \begin{vmatrix} a_2 & a_3 & \cdots & a_{v+1} \\ a_1 & a_2 & \cdots & a_v \\ \vdots & \vdots & \ddots & \vdots \\ a_{3-v} & a_{4-v} & \cdots & a_2 \end{vmatrix} \quad (v = 1, 2, \dots), \quad D_{2,0} := 1.$$

Thus we obtain the following series for the “minimal” root of equation (J.1), discussed by E. T. Whittaker and proved by him under very special assumptions:

$$\xi_1 = -\sum_{v=1}^{\infty} \frac{D_{2,v-1}}{D_{1,v-1} D_{1,v}}.$$

However, obviously the use of this series implies completely unnecessary additional computations, as in order to obtain P_{v-1}/P_v from this series we would have to compute all the determinants

$$\begin{aligned} D_{2,1}, & D_{2,2}, \dots, D_{2,v-1}, \\ D_{1,1}, & D_{1,2}, \dots, D_{1,v}, \end{aligned}$$

while by (J.6) the knowledge of $D_{1,v}$ and $D_{1,v-1}$ is completely sufficient.

†Indeed let us denote generally by

$$\Delta \left(\begin{array}{c} \alpha_1, \alpha_2, \dots \\ \beta_1, \beta_2, \dots \end{array} \right)$$

the determinant obtained from the determinant Δ by dropping the rows with the indices $\alpha_1, \alpha_2, \dots$ and the columns with the indices β_1, β_2, \dots . The Sylvester formula we have to use is then

$$\Delta \left(\begin{array}{c} 1 \\ 1 \end{array} \right) \Delta \left(\begin{array}{c} v \\ v \end{array} \right) - \Delta \left(\begin{array}{c} v \\ 1 \end{array} \right) \Delta \left(\begin{array}{c} 1 \\ v \end{array} \right) = \Delta \Delta \left(\begin{array}{cc} 1 & v \\ 1 & v \end{array} \right).$$

If we take now the determinant $D_{1,v}$ in (J.4) as Δ , we obtain easily

$$\begin{aligned} \Delta \left(\begin{array}{c} 1 \\ 1 \end{array} \right) &= D_{1,v-1}, \quad \Delta \left(\begin{array}{c} v \\ v \end{array} \right) = D_{1,v-1}, \quad \Delta \left(\begin{array}{c} v \\ 1 \end{array} \right) = D_{2,v-1}, \\ \Delta \left(\begin{array}{c} 1 \\ v \end{array} \right) &= a_0^{v-1} = 1, \quad \Delta \left(\begin{array}{cc} 1 & v \\ 1 & v \end{array} \right) = D_{1,v-2} \end{aligned}$$

and therefore from Sylvester's formula

$$D_{1,v-1}^2 - D_{2,v-1} = D_{1,v} D_{1,v-2};$$

this gives, together with (J.6), the assertion.

6. We now proceed to show that the knowledge of the coefficients P_v of (J.2) enables us in many cases to obtain easily *products of roots* of Eq. (J.1).

Instead of the assumption of Section 4, suppose that (J.1) has inside its circle of convergence exactly k roots ξ_1, \dots, ξ_k such that

$$0 < |\xi_1| \leq |\xi_2| \leq \dots \leq |\xi_k|, \quad (\text{J.12})$$

and that the moduli of all other roots of (J.1) are $> |\xi_k|$. Put

$$N(z) := \prod_{\kappa=1}^k (z - \xi_\kappa) = A_0 z^k + A_1 z^{k-1} + \dots + A_k, \quad A_0 = 1. \quad (\text{J.13})$$

Then we can write

$$\Phi(z) = \sum_{v=0}^{\infty} P_v z^v = \frac{S(z)}{N(z)} + \sum_{v=0}^{\infty} c_v z^v, \quad (\text{J.14})$$

where the right-hand power series, with the coefficients c_v , has a radius of convergence $\rho > |\xi_k|$, and $S(z)$ is a polynomial of degree $< k$, relatively prime to $N(z)$.

For any positive q with

$$\rho > \frac{1}{q} > |\xi_k| \quad (\text{J.15})$$

we then have

$$c_v = o(q^v) \quad (v \rightarrow \infty). \quad (\text{J.16})$$

We shall find an expression for the product $\xi_1 \xi_2 \cdots \xi_k$ in terms of the determinants

$$\Delta_v = \begin{vmatrix} P_v & P_{v+1} & \dots & P_{v+k-1} \\ P_{v-1} & P_v & \dots & P_{v+k-2} \\ \vdots & \vdots & & \vdots \\ P_{v-k+1} & P_{v-k+2} & \dots & P_v \end{vmatrix}, \quad v \geq k-1. \quad (\text{J.17})$$

Our proof will use only the assumptions concerning (J.12), (J.13), and (J.14) and will be independent of formula (J.2), that is, of the assumption that $1/\Phi(z)$ is regular for $|z| < r$.

7. Develop $S(z)/N(z)$ in powers of z ; we have

$$\frac{S(z)}{N(z)} = \sum_{v=0}^{\infty} y_v z^v. \quad (\text{J.18})$$

Multiplying both sides of (J.18) by $N(z)$ and comparing the coefficients of equal powers of z on both sides, we get

$$y_v A_k + \dots + y_{v-k} A_0 = 0 \quad (v \geq k). \quad (\text{J.19})$$

Consider now the determinants

$$D_v = \begin{vmatrix} y_v & y_{v+1} & \cdots & y_{v+k-1} \\ y_{v-1} & y_v & \cdots & y_{v+k-2} \\ \vdots & \vdots & & \vdots \\ y_{v-k+1} & y_{v-k+2} & \cdots & y_v \end{vmatrix}, \quad v \geq k-1. \quad (\text{J.20})$$

If in the determinant (J.20) we add to the first row the second row multiplied by A_{k-1}/A_k , the third row multiplied by A_{k-2}/A_k , ..., the last row multiplied by A_1/A_k , we get by (J.19) as the new first row

$$-\frac{y_{v-k}}{A_k}, \quad -\frac{y_{v-k+1}}{A_k}, \dots, -\frac{y_{v-1}}{A_k}.$$

If we bring this row by $k-1$ permutations into the k th row, we have, except for the factor

$$\frac{(-1)^k}{A_k} = \frac{1}{\xi_1 \cdots \xi_k},$$

the determinant D_{v-1} . Thus we obtain

$$D_{v-1} = \xi_1 \cdots \xi_k D_v. \quad (\text{J.21})$$

8. We are now going to prove that

$$D_{k-1} = \begin{vmatrix} y_{k-1} & y_k & \cdots & y_{2k-2} \\ y_{k-2} & y_{k-1} & \cdots & y_{2k-3} \\ \vdots & \vdots & & \vdots \\ y_0 & y_1 & \cdots & y_{k-1} \end{vmatrix}$$

does not vanish. Indeed, otherwise we would have the n relations

$$\begin{aligned} \alpha_{k-1} y_0 + \alpha_{k-2} y_1 + \cdots + \alpha_0 y_{k-1} &= 0, \\ \alpha_{k-1} y_1 + \alpha_{k-2} y_2 + \cdots + \alpha_0 y_k &= 0, \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{k-1} y_{k-1} + \alpha_{k-2} y_k + \cdots + \alpha_0 y_{2k-2} &= 0, \end{aligned} \quad (\text{J.22})$$

where the constants $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ do not all vanish.

Set

$$H(z) = \alpha_{k-1} z^{k-1} + \alpha_{k-2} z^{k-2} + \cdots + \alpha_0$$

and multiply both sides of (J.18) by $H(z)$.

Then in the product, $H(z) \sum_{v=0}^{\infty} y_v z^v$, the coefficients of $z^{k-1}, z^k, \dots, z^{2k-2}$ vanish by virtue of (J.22), and we obtain therefore

$$H(z) \frac{S(z)}{N(z)} = T_{k-2}(z) + z^{2k-1} W(z),$$

where $T_{k-2}(z)$ is a polynomial of degree $k-2$ and $W(z)$ is a power series in z containing only nonnegative powers. But then in the equation

$$\frac{H(z) S(z) - T_{k-2}(z) N(z)}{N(z)} = z^{2k-1} W(z)$$

the numerator on the left-hand side is, at the most, of degree $2k-2$; therefore, since it is divisible by z^{2k-1} , it must vanish identically. We would then have, however,

$$\frac{S(z)}{N(z)} = \frac{T_{k-2}(z)}{H(z)},$$

while $S(z)$ and $N(z)$ are assumed to be relatively prime. We see that the determinant D_{k-1} cannot vanish.[†]

9. From (J.14) and (J.18) we have

$$P_v = y_v + c_v \quad (v = 0, 1, \dots). \quad (\text{J.23})$$

It can therefore be expected that the determinant Δ_v is not very different from D_v . However, to prove the corresponding estimates we shall have to discuss the inverse matrix of (D_v) .[‡]

We consider the matrix

$$X = \begin{bmatrix} -A_1 & 1 & 0 & \cdots & 0 \\ -A_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -A_{k-1} & 0 & \cdots & \cdots & 1 \\ -A_k & 0 & \cdots & \cdots & 0 \end{bmatrix}, \quad (\text{J.24})$$

A_k being the coefficients of $N(z)$ in (J.13), and form the product $(D_v) X$. To obtain the first row in this product, we take the first row of (D_v) :

$$y_v, \quad y_{v+1}, \dots, y_{v+k-1}. \quad (\text{J.25})$$

Multiplying it by the first column of X , we obtain

$$-A_1 y_v - A_2 y_{v+1} - \cdots - A_k y_{v+k-1},$$

[†] This result apparently goes back to L. Kronecker (1881).

[‡] For any determinant D we denote the corresponding matrix by the symbol (D) .

and this is, by (J.19) and (J.13), equal to y_{v-1} . The following elements of the first row of $(D_v)X$ are obtained by multiplying (J.25) by the 2nd, 3rd, ..., k th columns of X and are easily seen to be $y_v, y_{v+1}, \dots, y_{v+k-2}$. Therefore the complete first row of $(D_v)X$ is

$$y_{v-1}, \quad y_v, \quad y_{v+1}, \dots, y_{v+k-2}. \quad (\text{J.26})$$

The following rows of $(D_v)X$ are obtained from (J.26) by diminishing consecutively the indices of the elements of (J.26) by 1, since this is true in (D_v) . We see that our product is just the matrix (D_{v-1}) ,

$$(D_v)X = (D_{v-1}),$$

and it follows that

$$\begin{aligned} (D_v)X^{v-k+1} &= (D_{k-1}), & (D_v)^{-1} &= X^{v-k+1}(D_{k-1})^{-1}, \\ (D_v)^{-1} &= X^v T, & T &:= X^{-k+1}(D_{k-1})^{-1}. \end{aligned} \quad (\text{J.27})$$

10. Set

$$C_v = \begin{pmatrix} c_v & c_{v+1} & \cdots & c_{v+k-1} \\ c_{v-1} & c_v & \cdots & c_{v+k-2} \\ \vdots & \vdots & \vdots & \vdots \\ c_{v-k+1} & c_{v-k+2} & \cdots & c_v \end{pmatrix}. \quad (\text{J.28})$$

We obtain from (J.17), (J.20), (J.23), and (J.27)

$$(\Delta_v) = (D_v) + C_v, \quad (\Delta_v)(D_v)^{-1} = I + C_v X^v T.$$

Taking the determinants, we get

$$\Delta_v/D_v = |I + C_v X^v T|. \quad (\text{J.29})$$

We now have to obtain an estimate for X^v as $v \rightarrow \infty$, and in order to do so, we must obtain the fundamental roots of X .

Now we will show that

$$|zI - X| = \begin{vmatrix} A_1 + z & -1 & 0 & \cdots & 0 & 0 & 0 \\ A_2 & z & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{k-2} & 0 & 0 & \cdots & z & -1 & 0 \\ A_{k-1} & 0 & 0 & \cdots & 0 & z & -1 \\ A_k & 0 & 0 & \cdots & 0 & 0 & z \end{vmatrix} = N(z). \quad (\text{J.30})$$

Indeed, putting for $v \leq k$,

$$N_v(z) = A_0 z^v + \cdots + A_v, \quad N_k(z) = N(z),$$

we have

$$N_k(z) = zN_{k-1}(z) + A_k.$$

Equation (J.30) is true for $k = 1$, $N_1 = z + A_1 = A_0 z + A_1$. Assume the relation corresponding to (J.30) true for $k - 1$ and develop the determinant in (J.30) by the elements of the last row. Then we get

$$|zI - X| = (-1)^{k-1} A_k (-1)^{k-1} + zN_{k-1}(z) = N(z).$$

We see that the fundamental roots of X are just given by[†]

$$\xi_1, \xi_2, \dots, \xi_k.$$

We have in particular (see (19.24)) $\lambda_X = |\xi_k|$.

11. It follows now from Theorem 20.1, replacing there all U_μ by 0, that for any $\varepsilon > 0$ we have with a convenient $\sigma = \sigma(X, \varepsilon) > 0$ in the notation of Chapter 19

$$|X^v|_\infty < \sigma(\lambda_X + \varepsilon)^v = \sigma(|\xi_k| + \varepsilon)^v \quad (v = 1, 2, \dots).$$

On the other hand, from (J.16) it follows that for the matrix (J.28)

$$|C_v|_\infty = o(q^v) \quad (v \rightarrow \infty),$$

and therefore by (19.22)

$$|C_v X^v|_\infty = o[(q|\xi_k| + q\varepsilon)^v],$$

$$|C_v X^v T|_\infty = o[(q|\xi_k| + q\varepsilon)^v] \quad (v \rightarrow \infty).$$

Take now an arbitrary θ with $1 > \theta > |\xi_k|/\rho$ and put

$$\theta - \frac{|\xi_k|}{\rho} = \delta, \quad q = \frac{1}{\rho} + \frac{\delta}{2|\xi_k|}, \quad |\xi_k|q = \frac{|\xi_k|}{\rho} + \frac{\delta}{2}.$$

Then (J.15) is satisfied. If we now take

$$\varepsilon = \delta/2q,$$

we have

$$q(|\xi_k| + \varepsilon) = \frac{|\xi_k|}{\rho} + \frac{\delta}{2} + \frac{\delta}{2} = \theta,$$

and therefore by (J.29)

$$\frac{\Delta_v}{D_v} = 1 + o(\theta^v) \quad \left(1 > \theta > \frac{|\xi_k|}{\rho}, \quad v \rightarrow \infty \right).$$

[†] This goes back to S. Günther (1876).

12. From this we have finally by (J.21), for all θ with $1 > \theta > |\xi_k|/\rho$,

$$\frac{\Delta_{v-1}}{\Delta_v} \rightarrow \xi_1 \cdots \xi_k, \quad \frac{\Delta_{v-1}}{\Delta_v} = \xi_1 \cdots \xi_k + o(\theta^v). \quad (\text{J.31})$$

Relation (J.31) has been deduced under the assumption that $\Phi(z)$ is meromorphic in $|z| < \rho$ and has there ξ_1, \dots, ξ_k as its only poles. If we now obtain $\Phi(z)$ from (J.2), we have to assume that $f(z)$ has inside $|z| < \rho$ the zeros ξ_1, \dots, ξ_k and that all other zeros of $f(z)$ inside $|z| < \rho$, if they exist at all, have moduli $> \text{Max}(|\xi_1|, \dots, |\xi_k|)$. The estimate of θ then has to be altered accordingly. If in particular $f(x)$ is a polynomial, the formula (J.31) gives a generalization of the so-called Bernoullian method in which only $k = 1$ is considered. In the Bernoullian case the linear character of the convergence is well known.

Formula (J.31) can be deduced easily from a famous theorem due to J. Hadamard ("Essai sur l'étude des fonctions données par leur développement de Taylor," *J. Math. Pures Appl.* [4] **8**, 101–186 (1892); partly reprinted in J. Hadamard's *Selecta*, pp. 19–37, Gauthier-Villars, 1935).

13. If we have in the notation of Section 6

$$|\xi_1| < |\xi_2| < |\xi_3| < \cdots < |\xi_k|, \quad (\text{J.32})$$

the determination of the ξ_1, \dots, ξ_k can be carried out by a recursive process without computing the determinants Δ_v . Indeed, multiplying (J.2) by $1 - z/\xi_1$, we obtain

$$\frac{1 - z/\xi_1}{f(z)} = \sum_{v=0}^{\infty} Q_v z^v, \quad Q_0 = 1, \quad Q_v = P_v - \frac{1}{\xi_1} P_{v-1} \quad (v > 0). \quad (\text{J.33})$$

From this and from (J.10), applied to the series (J.33), we obtain

$$\frac{Q_{v-1}}{Q_v} \rightarrow \xi_2. \quad (\text{J.34})$$

Multiplying the series in (J.33) again by $1 - z/\xi_2$, we get the series

$$\sum_{v=0}^{\infty} R_v z^v, \quad R_0 = 1, \quad R_v = Q_v - \frac{1}{\xi_2} Q_{v-1} \quad (v > 0), \quad (\text{J.35})$$

and from this again

$$\frac{R_{v-1}}{R_v} \rightarrow \xi_3, \quad (\text{J.36})$$

and so on. If we have computed from the beginning the coefficients P_v of (J.2) up to P_N , we can replace $1/\xi_1$ by P_N/P_{N-1} in (J.33) and obtain easily as the

approximate value of ξ_2

$$\frac{P_{N-1} P_{N-2} - P_N P_{N-3}}{P_{N-1}^2 - P_N P_{N-2}}. \quad (\text{J.37})$$

Again, using (J.37) as the approximate value of ξ_3 we obtain R_{N-3}/R_{N-2} and so on. However, this method is only applicable if (J.32) is satisfied.

14. We consider as a numerical example the function

$$\cos \sqrt{x} = \sum_{v=0}^{\infty} (-1)^v \frac{x^v}{(2v)!},$$

for which

$$\xi_1 = \frac{\pi^2}{4} = 2.46740, \quad \xi_1 \xi_2 = 9 \cdot \frac{\pi^4}{16} = 54.79.$$

In this case we obtain the P_v for $v = 0, \dots, 5$:

v	0	1	2	3	4	5
P_v	1	0.5	0.208333	0.08472222	0.034350198	0.013922233

and the corresponding approximations P_{v-1}/P_v with the corresponding errors E_v :

v	1	2	3	4	5
P_{v-1}/P_v	2	2.4	2.4590	2.466426	2.46791
E_v	0.4	0.067	0.0084	0.0010	0.00011

Further, we obtain the values of Δ_{v-1}/Δ_v for $k = 2$ and $v = 2, 3, 4$ with the corresponding errors E'_v :

v	2	3	4
Δ_{v-1}/Δ_v	40.01	48.19	52.25
E'_v	14.78	6.60	2.54

In this case the errors E_v indeed decrease by a factor convergent to $1/9 = (\pi^2/4) \cdot (1/(3\pi/2))^2$, while the errors E'_v decrease by a factor going to $9/25 = (3\pi/2)^2/(5\pi/2)^2$.

K

Continuity of the Fundamental Roots as Functions of the Elements of the Matrix

1. We shall prove:

Theorem. Let $A = (a_{\mu\nu})$, $B = (b_{\mu\nu})$ be two $n \times n$ matrices and

$$\varphi(\lambda) := |A - \lambda I| = 0, \quad \psi(\lambda) := |B - \lambda I| = 0 \quad (\text{K.1})$$

the corresponding characteristic polynomials and equations. Denote the zeros of $\varphi(\lambda)$ by λ_v and those of $\psi(\lambda)$ by λ'_v . Put

$$M = \text{Max}(|a_{\mu\nu}|, |b_{\mu\nu}|) \quad (\mu, v = 1, \dots, n), \quad (\text{K.2})$$

$$\frac{1}{nM} \sum_{\lambda, v} |a_{\mu\nu} - b_{\mu\nu}| = \delta. \quad (\text{K.3})$$

Then to every root λ'_v of $\psi(\lambda)$ belongs a certain root λ_v of $\varphi(\lambda)$ such that we have

$$|\lambda'_v - \lambda_v| \leq (n+2) M \delta^{1/n}. \quad (\text{K.4})$$

Further, for a suitable ordering of λ_v and λ'_v we have

$$|\lambda_v - \lambda'_v| \leq 2(n+1)^2 M \delta^{1/n} \quad (v = 1, \dots, n). \quad (\text{K.5})$$

2. We prove first the

Lemma. Under the hypotheses of the theorem, we have for any λ with $|\lambda| \leq nM$

$$|\varphi(\lambda) - \psi(\lambda)| \leq (n+2)^n M^n \delta. \quad (\text{K.6})$$

Proof of the Lemma. Denote the $a_{\mu\nu}$ in an arbitrarily chosen order by $\alpha_1, \dots, \alpha_{n^2}$ and the $b_{\mu\nu}$ in the corresponding order by $\beta_1, \dots, \beta_{n^2}$. We can then write

$$\begin{aligned} \varphi(\lambda) &= P(\alpha_1, \dots, \alpha_{n^2}), & \psi(\lambda) &= P(\beta_1, \dots, \beta_{n^2}), \\ \varphi(\lambda) - \psi(\lambda) &= \sum_{\kappa=1}^{n^2} \Delta_\kappa \end{aligned} \quad (\text{K.7})$$

with

$$\Delta_\kappa = P(\alpha_1, \dots, \alpha_\kappa, \beta_{\kappa+1}, \dots, \beta_{n^2}) - P(\alpha_1, \dots, \alpha_{\kappa-1}, \beta_\kappa, \dots, \beta_{n^2}).$$

Since $P(\alpha_1, \dots, \alpha_{n^2})$ is linear with respect to any of the α_v , we have

$$\Delta_\kappa = \pm (\alpha_\kappa - \beta_\kappa) T_\kappa \quad (\kappa = 1, \dots, n^2), \quad (\text{K.8})$$

where T_κ is obtained from a minor of order $n-1$ of P by replacing there some of α_i by the corresponding β_i . Further, in every row of T_κ we have to subtract λ from at the most one of the elements α_v, β_v . Therefore, the Euclidean norm of every line in T_κ is

$$\leq \sqrt{(n-2)M^2 + (n+1)^2M^2} < (n+2)M.$$

It follows by Hadamard's estimate of the modulus of a determinant that

$$|T_\kappa| \leq (n+2)^{n-1}M^{n-1},$$

and therefore by (K.7), (K.8), and (K.3)

$$|\varphi(\lambda) - \psi(\lambda)| \leq (n+2)^{n-1}M^{n-1} \sum_{\kappa=1}^{n^2} |\alpha_\kappa - \beta_\kappa| \leq (n+2)^n M^n \delta,$$

that is, (K.6).

3. We can now prove our theorem. By Theorem 19.1, for any root λ' of $\psi(\lambda)$ we have $|\lambda'| \leq nM$ and therefore by (K.6)

$$|\varphi(\lambda')| = |\varphi(\lambda') - \psi(\lambda')| \leq (n+2)^n M^n \delta,$$

$$\prod_{v=1}^n |\lambda' - \lambda_v| \leq [(n+2)M\delta^{1/n}]^n; \quad (\text{K.9})$$

therefore for at least one of the λ_v we have (K.4).

If we now denote the right-hand expression in (K.4) by ε , the argument used in Sections 5–8 of Appendix A can be applied without any change. We see that by ordering the λ_v and λ'_v conveniently, we have

$$|\lambda'_v - \lambda_v| \leq 2n(n+2)M\delta^{1/n} \leq 2(n+1)^2M\delta^{1/n} \quad (v = 1, \dots, n);$$

(K.5) is proved.

L

The Determinantal Formulas for Divided Differences

1. The general divided difference of f is linear in the corresponding values of f :

$$[x_1, \dots, x_m] f = \sum_{v=1}^m U_v f(x_v), \quad (\text{L.1})$$

where the U_v are rational expressions in x_1, \dots, x_m , the values of which are obtained immediately from (1A.6):

$$U_m = \frac{1}{(x_m - x_1)(x_m - x_2) \cdots (x_m - x_{m-1})}. \quad (\text{L.2})$$

2. To write (L.1) as a quotient of determinants, introduce two *column* vectors $Z(x)$ and $N(x)$ by

$$Z(x) = (1, x, \dots, x^{m-1}, f(x))', \quad N(x) = (1, x, \dots, x^{m-1}, x^m)'. \quad (\text{L.3})$$

Then we easily see that

$$[x_1, \dots, x_m] f = \frac{|Z(x_1), \dots, Z(x_m)|}{|N(x_1), \dots, N(x_m)|}. \quad (\text{L.4})$$

Indeed, since both sides in (L.4) are symmetric, it is sufficient to compare the coefficient of $f(x_m)$.

This coefficient on the right is the quotient of two Vandermonde's determinants,

$$\frac{\prod_{m-1 \geq \mu > v \geq 1} (x_\mu - x_v)}{\prod_{m \geq \mu > v \geq 1} (x_\mu - x_v)} = \frac{1}{\prod_{v=1}^{m-1} (x_m - x_v)},$$

and this is just the above value of U_m .

3. In order to obtain the corresponding representation of the divided difference in the *confluent case*, consider k systems of variables

$$x_1, \dots, x_{m_1}; y_1, \dots, y_{m_2}; \dots; z_1, \dots, z_{m_k} \quad (m_1 + \dots + m_k = n),$$

supposed to be all distinct, and the corresponding divided difference

$$\begin{aligned} & [x_1, \dots, x_{m_1}; y_1, \dots, y_{m_2}; \dots; z_1, \dots, z_{m_k}] f \\ &= \frac{|Z(x_1) Z(x_2) \cdots Z(x_{m_1}) Z(y_1) \cdots Z(z_{m_k})|}{|N(x_1) N(x_2) \cdots N(x_{m_1}) N(y_1) \cdots N(z_{m_k})|}. \end{aligned} \quad (\text{L.5})$$

Subtracting in both the numerator and the denominator the first column from the second and dividing by $x_2 - x_1$ we see that the second columns can be replaced respectively by

$$[x_2, x_1] Z(x_1), \quad [x_2, x_1] N(x_1).$$

Applying the same procedure to every x_v column ($1 < v \leq m_1$) we can replace each x_v column respectively by

$$[x_v, x_1] Z(x_1), \quad [x_v, x_1] N(x_1).$$

Operating in the same way and starting from the second column, we can replace the third column in the numerator and denominator by

$$[x_3, x_2, x_1] Z(x_1), \quad [x_3, x_2, x_1] N(x_1).$$

Applying the same procedure repeatedly we can finally replace the first m_1 columns in the numerator and denominator respectively by

$$Z(x_1), \quad [x_2, x_1] Z(x_1), \dots, [x_{m_1}, \dots, x_2, x_1] Z(x_1);$$

$$N(x_1), \quad [x_2, x_1] N(x_1), \dots, [x_{m_1}, \dots, x_2, x_1] N(x_1).$$

The same procedure can be applied to the columns depending on the y_v, \dots, z_v and we obtain finally

$$\begin{aligned} & [x_1, \dots, x_{m_1}; y_1, \dots, y_{m_2}; \dots; z_1, \dots, z_{m_k}] f \\ &= \frac{|Z(x_1)[x_1, x_2] Z(x_1) \cdots [x_1, \dots, x_{m_1}] Z(x_1) Z(y_1) \cdots [z_1, \dots, z_{m_k}] Z(z_1)|}{|N(x_1)[x_1, x_2] N(x_1) \cdots [x_1, \dots, x_{m_1}] N(x_1) \cdots [z_1, \dots, z_{m_k}] N(z_1)|}. \end{aligned} \quad (\text{L.6})$$

4. Assume now k distinct variables t_1, \dots, t_k and in (L.6) let all x_v tend to t_1 , all y_v tend to t_2, \dots , all z_v tend to t_k ; then we obtain

$$\begin{aligned} & [t_1^{m_1}, \dots, t_k^{m_k}] f = Z^*/N^*, \\ & Z^* = |Z(t_1) Z'(t_1) \cdots Z^{(m_1-1)}(t_1) Z(t_2) \cdots Z^{(m_k-1)}(t_k)|, \\ & N^* = |N(t_1) N'(t_1) \cdots N^{(m_1-1)}(t_1) \cdots N^{(m_k-1)}(t_k)|, \end{aligned} \quad (\text{L.7})$$

provided N^* is $\neq 0$.

5. We are now going to show that indeed N^* is $\neq 0$ if the t_k are all distinct, as we will prove that

$$N^* = \prod_{\mu > v} (t_\mu - t_v)^{m_\mu m_v}. \quad (\text{L.8})$$

In order to prove (L.8), put

$$\Delta(x) = \begin{vmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_{m_1} \\ \vdots & & \vdots \\ x_1^{m_1-1} & \cdots & x_{m_1}^{m_1-1} \end{vmatrix} = \prod_{\mu > v} (x_\mu - x_v)$$

and define similarly $\Delta(y), \dots, \Delta(z)$.

Then, if we denote the denominator in (L.5) by N and that in (L.6) by N_0 , we have obviously

$$N_0 = \frac{N}{\Delta(x)\Delta(y)\cdots\Delta(z)}. \quad (\text{L.9})$$

Put, further,

$$R(x, y) = \prod (x_\mu - y_v) \quad (\mu = 1, \dots, m_1; \quad v = 1, \dots, m_2)$$

and define similarly $R(x, z)$, $R(y, z)$, etc. Then, writing in (L.9) the Vandermonde determinant N as the difference product of the arguments, we obtain

$$N_0 = R(x, y) \cdots R(x, z) \cdots R(y, z) \cdots$$

and this tends indeed to the right-hand product in (L.8).

6. The above gives a second proof of the existence of the confluent divided difference (1B.4) and also a formula for this difference, formally different from (1B.4).

M

Remainder Terms in Interpolation Formulas

1. Formulas (1B.12) (and (1B.14)) for the remainder term were deduced in the real case under the assumption that $f^{(n)}(x)$ (or $f^{(n+1)}(x)$) is *continuous* in $\langle x_1, \dots, x_m \rangle$.

However, these formulas can also be proved assuming in the case of (1B.12) only the *existence* of the derivative $f^{(n)}$ throughout the corresponding interval and in the case of (1B.14) that the derivative $f^{(n+1)}$ exists and is uniformly bounded in its interval.

As to the case of *confluent divided differences*, even the existence of the *confluent* divided difference (1B.4) has been proved in Chapter 1B, Section 1 only under the assumption of the *continuity* of the $f^{(m_k)}(x)$ in the corresponding neighborhoods. This raises the question whether (1B.4) still exists if only the *existence* of the $f^{(m_k)}(x)$ is assumed. We are going to show in an example that this is not the case.

Although these problems have no great practical importance, they present a certain theoretical interest.

Before discussing them, we prove some lemmata.

2. **Lemma 1.** *Let $F(x)$ be defined in J_x . Assume that $F(x_v) = 0$ ($v = 1, \dots, n+1$), $x_v \in J_x$, and that $F^{(n)}(x)$ exists in J_x . Then there exists a $\xi \in J_x$ such that $F^{(n)}(\xi) = 0$. ξ lies even in (J_x) unless all x_v are concentrated in one of the end points.*

Remarks. The zeros of $F(x)$ may be *multiple*. Any multiple zero must be counted a corresponding number of times. This lemma is a generalization of Rolle's theorem. As an illustration of the exceptional case mentioned in the last sentence of the lemma, consider

$$F(x) = x^{n+1}, \quad J_x: \quad 0 \leq x \leq 1;$$

then $\xi = 0$ and does not lie in (J_x) .

3. **Proof.** Let $g(x) := F'(x)$. We begin by showing that $g(x)$ has at least n zeros in J_x . Suppose first that the x_v ($v = 1, \dots, n+1$) are all distinct. If $F(x)$ has two distinct zeros, then by Rolle's theorem, $g(x)$ has a zero *between* these

two. More generally, if $F(x)$ has k distinct zeros, then $g(x)$ has $k - 1$ zeros separating those k zeros of $F(x)$. Clearly, if all x_v are simple, then $g(x)$ has n distinct zeros.

Assume now more generally that $F(x)$ has k different multiple zeros in J_x , where a *simple* zero is considered a zero of multiplicity *one*. If $F(x)$ has x_0 as a zero of multiplicity v , then $g(x)$ will have x_0 as a zero of multiplicity $v - 1$. Thus, each multiple zero loses one unit of its multiplicity in $g(x)$. If there are k multiple zeros, then $g(x)$ has at least $(n+1) - k + (k-1) = n$ zeros, where the term $(k-1)$ is the number of "new," Rollian zeros of $g(x)$. Hence, in any case, $g(x)$ has at least n zeros in J_x .

4. We consider now two cases:

Case I. We have $x_1 = x_2 = \dots = x_{n+1}$. Then x_1 is a zero of $F^{(n)}(x)$, i.e., $\xi = x_1$, and the lemma is true.

Case II. None of the x_v has the multiplicity $n+1$. Then, if $n = 1$, Rolle's theorem can be applied and the lemma is true. Assume the lemma is true for all smaller values of n . Then the lemma is true for $g(x)$ and, since $g(x)$ has at least one zero in (J_x) , there exists a $\xi \in (J_x)$ such that $g^{(n-1)}(\xi) = F^{(n)}(\xi) = 0$,

Q.E.D.

5. Lemma 2. Let $f(x), g(x)$ be defined and n times differentiable in J_x . Assume that there exist n common zeros $x_v \in J_x$ ($v = 1, \dots, n$) of $f(x)$ and $g(x)$, where, if a zero is counted with the multiplicity k , it must have at least the multiplicity k for both $f(x)$ and $g(x)$. Assume further that $g^{(n)}(x)$ does not vanish in J_x . Then for any $x_0 \neq x_v$ from J_x there exists a $\xi \in (J_x)$ such that

$$\frac{f(x_0)}{g(x_0)} = \frac{f^{(n)}(\xi)}{g^{(n)}(\xi)}, \quad \xi \in (J_x) \quad (x_0 \neq x_v; \quad v = 1, \dots, n). \quad (\text{M.1})$$

6. Proof. First, x_0 is not a zero of $g(x)$. For, otherwise, since $x_0 \neq x_v$, this would mean that $g(x)$ has $n+1$ zeros in J_x and by Lemma 1 there exists a $\xi \in J_x$ such that $g^{(n)}(\xi) = 0$, contrary to our hypothesis.

Let $\lambda = f(x_0)/g(x_0)$ and consider $F(x) = f(x) - \lambda g(x)$. $F(x)$ satisfies the hypothesis of Lemma 1, for x_v ($v = 1, \dots, n$) and x_0 are zeros of $F(x)$. Furthermore, since $g^{(n)}(x)$ and $f^{(n)}(x)$ exist in J_x , so does $F^{(n)}(x)$. Since $x_0 \neq x_v$ ($v = 1, \dots, n$), $F(x)$ has at least two *distinct* zeros and by Lemma 1 there exists a ξ in (J_x) such that

$$F^{(n)}(\xi) = f^{(n)}(\xi) - \lambda g^{(n)}(\xi) = 0$$

or

$$\frac{f^{(n)}(\xi)}{g^{(n)}(\xi)} = \lambda = \frac{f(x_0)}{g(x_0)}. \quad \text{Q.E.D.}$$

In the above lemma and in Lemma 1 it would be sufficient to assume the differentiability of $F(x), f(x), g(x)$ only in (J_x) and to require at the end points belonging to J_x only the continuity unless there are multiple zeros at the end points of J_x belonging to J_x . In this last case, it is sufficient to assume as many derivatives in the corresponding end points as are implied by the multiplicity of the zeros.

7. We are now going to show that $[x_2, x_1]f$ does not necessarily converge to $f'(t)$ for $x_2 \rightarrow t, x_1 \rightarrow t$, even if $f'(x)$ is assumed to exist in any point of a whole neighborhood of t . Consider, for instance, the function

$$f(x) = x^2 \sin(1/x) \quad (x \neq 0), \quad f(0) = 0,$$

which has for $x \neq 0$ the derivative $2x \sin(1/x) - \cos(1/x)$ and at $x = 0$ the derivative 0. Putting

$$\frac{1}{u_v} = \frac{\pi}{4} + 2v\pi \quad (v = 0, 1, \dots),$$

we have

$$f'(u_v) = \frac{\sqrt{2}}{\pi/4 + v\pi} - \sqrt{\frac{1}{2}},$$

and we see that the values of $f'(u_v)$ tend to $-\sqrt{\frac{1}{2}}$. Choose now $x_1^{(v)} \neq x_2^{(v)}$ and both so close to u_v that $[x_2^{(v)}, x_1^{(v)}]f(x) = -\sqrt{\frac{1}{2}} + O(1/v)$. Then we have indeed

$$x_1^{(v)} \rightarrow 0, \quad x_2^{(v)} \rightarrow 0, \quad [x_2^{(v)}, x_1^{(v)}]f(x) \rightarrow -\sqrt{\frac{1}{2}} \neq f'(0).$$

8. We return now to the general interpolation formulas (1B.9)–(1B.11) and assume that the interpolation abscissas x_1, \dots, x_n become partly confluent so that we have k distinct values t_1, \dots, t_k , each t_κ with the multiplicity m_κ , $\sum_{\kappa=1}^k m_\kappa = n$. Then formulas (1B.9), (1B.10) remain valid. If we assume that the $f^{(n)}(t)$ exists in the interval $J = \langle x, t_1, \dots, t_k \rangle$ and therefore $f^{(n-1)}(t)$ is continuous there, then, as long as x remains distinct from all t_κ , the expression (1B.15) for the remainder exists and remains valid. Indeed, in this case we have $\text{Max } m_\kappa - 1 \leq n - 1$.

9. We write now $L_{n-1}(x) = L(x)$ and consider a function $g(t)$ for which $g^{(n)}(t)$ exists and does not vanish in J , and which satisfies the relations

$$g^{(\mu)}(t_\kappa) = 0 \quad (\mu = 0, 1, \dots, m_\kappa - 1; \quad \kappa = 1, \dots, k). \quad (\text{M.2})$$

Then the quotient $[f(x) - L(x)]/g(x)$ satisfies the conditions of Lemma 2 and there exists therefore a ξ from J for which this quotient is equal to $f^{(n)}(\xi)/g^{(n)}(\xi)$, as $L^{(n)} \equiv 0$. We obtain therefore

$$f(x) = L_{n-1}(f, x) + \frac{f^{(n)}(\xi)}{g^{(n)}(\xi)} g(x). \quad (\text{M.3})$$

If we take in particular

$$g(x) = \prod_{v=1}^n (x - x_v) = \prod_{k=1}^k (x - t_k)^{m_k}, \quad g^{(n)} = n!,$$

(M.3) becomes

$$f(x) = L_{n-1}(x) + \frac{f^{(n)}(\xi)}{n!} \prod_{v=1}^n (x - x_v), \quad \xi \in \langle x, t_1, \dots, t_k \rangle. \quad (\text{M.4})$$

Comparing this with (1B.15), we see that we have

$$[x, t_1^{m_1}, \dots, t_k^{m_k}] f = \frac{1}{n!} f^{(n)}(\xi), \quad \xi \in \langle x, t_1, \dots, t_k \rangle \quad (\text{M.5})$$

for distinct x, t_1, \dots, t_k .

N

Generalization of Schröder's Series to the Case of Multiple Roots

1. If we are in the neighborhood of a *multiple root* ζ of the equation $f(x) = 0$ so that for an integer $p > 1$ we have

$$f(x) = (x - \zeta)^p F(x), \quad F(\zeta) \neq 0, \quad p > 1, \quad (\text{N.1})$$

we can obtain an analogous development to that of Schröder by applying Schröder's development to

$$u(x) = \sqrt[p]{f(x)} = (x - \zeta) F^{1/p}(x), \quad (\text{N.2})$$

where any of the values of $\sqrt[p]{F(\zeta)}$ can be chosen and $F^{1/p}(x)$ in the neighborhood of ζ is then determined by continuity.

However, while the theory of this method in the case of analytic functions works out easily along the classical lines, this theory requires in the real case some additional discussions, since we have to translate the assumptions concerning the higher derivatives of $f(x)$ into those implying the derivatives of $u(x)$. The essential difficulty consists in obtaining information about the derivatives of $F(x)$ in (N.1). To this purpose we derive first some lemmata.

2. In what follows we denote by $J(\zeta)$ a *one-sided neighborhood* of the point ζ , which does not contain ζ , and by $\bar{J} = \bar{J}(\zeta)$ this neighborhood closed in ζ .

Lemma 1. Assume that we have for the function $f(x)$ of the real variable x defined and continuous with its $n-1$ first derivatives in $\bar{J}(\zeta)$,

$$f(\zeta) = f'(\zeta) = \cdots = f^{(n)}(\zeta) = 0 \quad (\text{N.3})$$

so that

$$f(x) = (x - \zeta)^n \varphi(x), \quad (\text{N.4})$$

where $\varphi(x)$ is continuous in $\bar{J}(\zeta)$ with $\varphi(\zeta) = 0$.[†] Then we have for $k = 0, 1, \dots$

[†] That $\varphi(x)$ has 0 as limit in ζ follows from (N.3) if we apply the Bernoulli-L'Hôpital rule $(n-1)$ times. Then we get

$$\lim \varphi(x) = \lim \frac{f(x)}{(x - \zeta)^n} = \frac{1}{n!} \lim \frac{f^{(n-1)}(x)}{x - \zeta} = \frac{1}{n!} f^{(n)}(\zeta) = 0.$$

$n-1$:

$$(x-\zeta)^k \varphi^{(k)}(x) \rightarrow 0 \quad (x \rightarrow \zeta, \quad x \in J(\zeta), \quad k = 0, 1, \dots, n-1). \quad (\text{N.5})$$

Proof. The assertion is obvious for $k = 0$. Assume that we have already proved that

$$(x-\zeta)^\kappa \varphi^{(\kappa)}(x) \rightarrow 0 \quad (\kappa = 0, 1, \dots, k-1).$$

Differentiating (N.4) k times we obtain in $J(\zeta)$

$$f^{(k)}(x) = \sum_{\kappa=0}^k \binom{k}{\kappa} \varphi^{(\kappa)}(x) (k-\kappa)! \binom{n}{k-\kappa} (x-\zeta)^{n-k+\kappa}.$$

Let x now go to ζ out of J . It follows from (N.3) that we have, since $f^{(k)}(x)$ has in ζ a zero with the multiplicity $n-k+1$,

$$f^{(k)}(x) = (x-\zeta)^{n-k} \varphi_k(x), \quad \varphi_k(x) \rightarrow 0 \quad (x \rightarrow \zeta).$$

Thence, if we divide by $(x-\zeta)^{n-k}$,

$$\begin{aligned} \varphi_k(x) &= \sum_{\kappa=1}^{k-1} \binom{k}{\kappa} \varphi^{(\kappa)}(x) (x-\zeta)^\kappa (k-\kappa)! \binom{n}{k-\kappa} \\ &\quad + (x-\zeta)^k \varphi^{(k)}(x) + k! \binom{n}{k} \varphi(x). \end{aligned}$$

But here, for $x \rightarrow \zeta$, all terms in the first right-hand sum tend to 0, and since we have $\lim k! \binom{n}{k} \varphi(x) = \lim \varphi_k(x) = 0$, the second right-hand term tends to 0. This proves (N.5).

3. Lemma 2. Put, under the assumptions of Lemma 1, for a natural $p \leq n$,

$$F(x) = \frac{f(x)}{(x-\zeta)^p}. \quad (\text{N.6})$$

Then we have, for x from $J(\zeta)$ tending to ζ and for $k = 0, 1, \dots, n-1$,

$$F^{(k)}(x) = o((x-\zeta)^{n-k-p}) \quad (x \rightarrow \zeta; \quad k = 0, 1, \dots, n-1). \quad (\text{N.7})$$

Proof. We have from (N.4) and (N.6) $F(x) = (x-\zeta)^{n-p} \varphi(x)$. Differentiating this k times we get

$$F^{(k)}(x) = \sum_{\kappa=0}^k \binom{k}{\kappa} \varphi^{(\kappa)}(x) (k-\kappa)! \binom{n-p}{k-\kappa} (x-\zeta)^{n-p-k+\kappa},$$

$$\frac{F^{(k)}(x)}{(x-\zeta)^{n-p-k}} = \sum_{\kappa=0}^k \binom{k}{\kappa} (k-\kappa)! \binom{n-p}{k-\kappa} [(x-\zeta)^\kappa \varphi^{(\kappa)}(x)],$$

and here all expressions in brackets tend to 0, by (N.5). Relation (N.7) is proved.

4. From (N.7) we have obviously for $k \leq n-p$

$$F^{(k)}(x) \rightarrow 0 \quad (x \rightarrow \zeta; \quad x \in J(\zeta); \quad k = 0, 1, \dots, n-p). \quad (\text{N.8})$$

In order to derive from this result some information about $F^{(k)}(\zeta)$, we need a further lemma.

Lemma 3. Assume that for $n \geq 1$ the function $f(x)$ is continuous with its first n derivatives in J , and further that if x in J tends to ζ , $f^{(n)}(x)$ has a finite limit α_n :

$$f^{(n)}(x) \rightarrow \alpha_n \quad (x \rightarrow \zeta, \quad x \in J). \quad (\text{N.9})$$

Then $f(x)$ has a finite limit α as x tends to ζ in J . If $f(x)$ is assigned in ζ the value α , the function $f(x)$ is continuous with its first n derivatives in \bar{J} .

5. Proof. Define in \bar{J} the function $\psi(x)$ as having the values $f^{(n)}(x)$ in J and α_n at ζ . The function $\psi(x)$ is continuous in \bar{J} . We have then in J

$$f^{(n-1)}(x) = f^{(n-1)}(b) + \int_b^x \psi(u) du,$$

where b is a point of J , arbitrarily chosen.

If in this formula we let x tend to ζ , we obtain

$$\lim_{x \rightarrow \zeta} f^{(n-1)}(x) = f^{(n-1)}(b) + \int_b^\zeta \psi(u) du =: \alpha_{n-1} \quad (x \in J).$$

We see that $f^{(n-1)}(x)$ has a limit α_{n-1} for $x \rightarrow \zeta$. Applying this result repeatedly, we see that generally

$$f^{(v)}(x) \rightarrow \alpha_v \quad (x \rightarrow \zeta, \quad x \in J, \quad v = 0, 1, \dots, n).$$

We define now $f(x)$ at ζ as α_0 . We have then for $x \in J$ by the mean value theorem

$$\frac{f(x) - f(\zeta)}{x - \zeta} = \frac{f'(\xi)}{1},$$

where ξ lies inside the interval between x and ζ , that is, in J . For $x \rightarrow \zeta$, it follows that $f'(\zeta) = \alpha_1$. Applying the same argument to $f'(x)$ we obtain $f''(\zeta) = \alpha_2$ and proceeding in the same way we have generally

$$f^{(v)}(\zeta) = \alpha_v \quad (v = 0, 1, \dots, n).$$

Lemma 3 is proved.

6. From Lemma 3 and the relation (N.8) we have now:

Corollary. Under the conditions of Lemma 2 we have, defining $F(\zeta)$ as 0,

$$F^{(k)}(\zeta) = 0 \quad (k = 0, 1, \dots, n-p). \quad (\text{N.10})$$

7. With the next lemma we prove a little more than we need in our special case, but the result is of some general interest.

Lemma 4. Assume two natural numbers, p, n , $p \leq n$, and consider two functions $g(x), G(x)$ defined on J and satisfying the relations

$$g(x) = (x-\zeta)^p G(x), \quad g(\zeta) = 0. \quad (\text{N.11})$$

Assume that $g(x)$ satisfies the condition

$A_{n,p}$. $g(x)$ is continuous with its derivatives up to the order $n-1$ in \bar{J} , $g^{(n)}(\zeta)$ exists, and we have

$$g(\zeta) = g'(\zeta) = \dots = g^{(p-1)}(\zeta) = 0. \quad (\text{N.12})$$

We have then

$$G(\zeta) := \lim_{x \rightarrow \zeta} G(x) = \frac{1}{p!} g^{(p)}(\zeta), \quad (\text{N.13})$$

and $G(x)$ satisfies the condition

$B_{n,p}$. $G^{(v)}(x)$ exists and is continuous in \bar{J} for $v = 0, 1, \dots, n-p$ and we have for x going to ζ out of J

$$G^{(\lambda)}(x) = o((x-\zeta)^{n-p-\lambda}) \quad (\lambda = n-p+1, \dots, n-1). \quad (\text{N.14})$$

Conversely, if $G(x)$ satisfies the condition $B_{n,p}$, then $g(x)$ satisfies the condition $A_{n,p}$.

8. Proof. Consider, assuming for $g(x)$ the property $A_{n,p}$, the Taylor polynomial of order n for $g(x)$ at ζ ,

$$T(x) = \sum_{v=0}^{n-p} \frac{g^{(p+v)}(\zeta)}{(p+v)!} (x-\zeta)^{p+v}, \quad (\text{N.15})$$

and put

$$g(x) - T(x) = f(x). \quad (\text{N.16})$$

Then $f(x)$ satisfies the conditions of Lemmata 1 and 2. Defining $F(x)$ by (N.6) we have

$$G(x) = \frac{T(x)}{(x-\zeta)^p} + F(x).$$

The existence and the continuity of the first $n-p$ derivatives of $G(x)$ in J as well as the formula (N.13) follow now at once from Lemma 2, Lemma 3, (N.10), and (N.15).

As to the derivatives $G^{(\lambda)}(x)$ with $n-p < \lambda < n$, the formulas (N.14) follow from Lemma 2 and in particular from (N.7), since we have for these λ values $G^{(\lambda)}(x) = F^{(\lambda)}(x)$.

9. Assume now that $G(x)$ has the property $B_{n,p}$. Then, differentiating (N.11) λ times for $\lambda < n$, we have

$$\begin{aligned} g^{(\lambda)}(x) &= \sum_{v=0}^{\lambda} \binom{\lambda}{v} (\lambda-v)! \binom{p}{\lambda-v} (x-\zeta)^{p-\lambda+v} G^{(v)}(x), \\ g^{(\lambda)}(x) &= \sum_{\kappa=0}^{n-p} \binom{\lambda}{\kappa} (\lambda-\kappa)! \binom{p}{\lambda-\kappa} (x-\zeta)^{p-\lambda+\kappa} G^{(\kappa)}(x) \\ &\quad + (x-\zeta)^{n-\lambda} \sum_{\kappa=n-p+1}^{\lambda} \binom{\lambda}{\kappa} (\lambda-\kappa)! \binom{p}{\lambda-\kappa} (x-\zeta)^{p-n+\kappa} G^{(\kappa)}(x). \end{aligned}$$

The second right-hand sum is 0 for $\lambda \leq n-p$ and, for $\lambda > n-p$, as $x \rightarrow \zeta$, $o((x-\zeta)^{n-\lambda}) = o(1)$, by (N.14). We can therefore write, as $x \rightarrow \zeta$,

$$g^{(\lambda)}(x) = \sum_{\kappa=0}^{n-p} \binom{\lambda}{\kappa} (\lambda-\kappa)! \binom{p}{\lambda-\kappa} (x-\zeta)^{p-\lambda+\kappa} G^{(\kappa)}(x) + o((x-\zeta)^{n-\lambda}).$$

In this sum, $\binom{\lambda}{\kappa} = 0$ for $\kappa > \lambda$ and $\binom{\lambda}{\lambda-\kappa} = 0$ for $\lambda-\kappa > p$, $\kappa < \lambda$. We obtain

$$\begin{aligned} g^{(\lambda)}(x) &= \sum_{\substack{\text{Min}(\lambda, n-p) \\ \text{Max}(0, \lambda-p)}}^{\lambda} \binom{\lambda}{\kappa} (\lambda-\kappa)! \binom{p}{\lambda-\kappa} (x-\zeta)^{p-\lambda+\kappa} G^{(\kappa)}(x) \\ &\quad + o((x-\zeta)^{n-\lambda}) \quad (x \in J, x \rightarrow \zeta, 0 < \lambda < n). \end{aligned} \quad (\text{N.17})$$

For $\lambda < p$ it follows now that $g^{(\lambda)}(x) \rightarrow 0$. For $\lambda \geq p$ we can write, taking the term with $v = \lambda - p$ out,

$$\begin{aligned} g^{(\lambda)}(x) &= \sum_{v=\lambda-p+1}^{\text{Min}(\lambda, n-p)} C_{\lambda v} (x-\zeta)^{v-\lambda+p} G^{(v)}(x) \\ &\quad + o((x-\zeta)^{n-\lambda}) + p! \binom{\lambda}{\lambda-p} G^{(\lambda-p)}(x), \end{aligned} \quad (\text{N.18})$$

where $C_{\lambda v}$ are convenient numerical constants and $x \in J$.

In this formula we let x go to ζ . Again, in the first right-hand sum the derivatives $G^{(v)}(x)$ have finite limits, while the exponents of $(x-\zeta)$ are positive. It follows now that for $p = \lambda$, $g^{(p)}(x) \rightarrow p! G(\zeta)$, while for $p < \lambda$, $g^{(\lambda)}(x) \rightarrow \binom{\lambda}{p} p! G^{(\lambda-p)}(\zeta)$. With that, using Lemma 3, all assertions of Lemma 4 will be proved, if we prove that concerning $g^{(n)}(\zeta)$.

To prove the existence of $g^{(n)}(\zeta)$, take (N.17) for $\lambda = n - 1$. If $p < n$, we have

$$\begin{aligned} g^{(n-1)}(x) &= p! \binom{n-1}{p} G^{(n-p-1)}(x) \\ &\quad + p! \binom{n-1}{p-1} (x - \zeta) G^{(n-p)}(x) + o(x - \zeta), \end{aligned}$$

while for $p = n$ we get

$$g^{(n-1)}(x) = n! (x - \zeta) G(x) + o(x - \zeta).$$

Using Lemma 3 we obtain $g^{(n-1)}(\zeta) = p! \binom{n-1}{p} G^{(n-p-1)}(\zeta)$. For $p = n$ this is also true as $\binom{n-1}{n} = 0$.

We subtract on both sides

$$g^{(n-1)}(\zeta) = p! \binom{n-1}{p} G^{(n-p-1)}(\zeta).$$

Then we obtain, dividing by $x - \zeta$, for $p < n$, by (N.14) for $\lambda = n - p - 1$

$$\frac{g^{(n-1)}(x) - g^{(n-1)}(\zeta)}{x - \zeta} = p! \binom{n}{p} G^{(n-p)}(x) + o(1),$$

while for $p = n$,

$$\frac{g^{(n-1)}(x) - g^{(n-1)}(\zeta)}{x - \zeta} = n! G(\zeta) + o(1). \quad (\text{N.19})$$

But here, for $x \rightarrow \zeta$, the right-hand expressions have in both cases, by virtue of the assumption $B_{n,p}$ about the existence and continuity of $G^{(n-p)}(x)$ in \bar{J} , a limit. This proves the existence of $g^{(n)}(\zeta)$ and Lemma 4 is proved.

10. Lemma 5. Let $G(x)$ be a function satisfying the condition $B_{n,p}$ of Lemma 4. Denoting by U an interval containing all values assumed by $G(x)$ in $\bar{J}(\zeta)$, assume that $z(\omega)$ is continuous for $\omega \in U$ together with $z', \dots, z^{(n-1)}$. Put

$$G^*(x) = z(G(x)). \quad (\text{N.20})$$

Then $G^*(x)$ also satisfies condition $B_{n,p}$ of Lemma 4.

11. Proof. It is clear from our hypotheses that, for a natural $m < n$, $G^{*(m)}(x)$ exists in J . The differentiation of (N.20) gives

$$G^{*(m)}(x) = \sum c_{\alpha_1, \dots, \alpha_m, \beta} G'^{\alpha_1}(x) \cdots G^{(m)\alpha_m}(x) z^{(\beta)}(G), \quad (\text{N.21})$$

where $\sum_{\mu=1}^m \alpha_{\mu} = \beta \leq m$ and

$$\sum_{\mu=1}^m \mu \alpha_{\mu} = m. \quad (\text{N.22})$$

These relations are proved easily by induction while the exact values of the numerical constants $c_{\alpha_1, \dots, \alpha_m, \beta}$ due to Faà di Bruno do not matter for our purpose.

For $m \leq n-p$, $G^{*(m)}(x)$ obviously has a finite limit if x tends, from $J(\zeta)$, to ζ .

12. Assume now that $m > n-p$,

$$m = n - p + \bar{m}, \quad \bar{m} > 0. \quad (\text{N.23})$$

The general term of (N.21) is, by (N.14),

$$o((x-\zeta)^{\omega}), \quad \omega = \sum_{\mu=n-p+1}^{n+p+\bar{m}} (n-p-\mu) \alpha_{\mu} = - \sum_{\mu=1}^{\bar{m}} \mu \alpha_{n-p+\mu}. \quad (\text{N.24})$$

On the other hand, we can write (N.22) in the form

$$\sum_{\kappa=1}^{n-p} \kappa \alpha_{\kappa} + (n-p) \sum_{\mu=1}^{\bar{m}} \alpha_{n-p+\mu} + \sum_{\mu=1}^{\bar{m}} \mu \alpha_{n-p+\mu} = n - p + \bar{m}$$

and therefore we have, using (N.24),

$$(n-p) \sum_{\mu=1}^{\bar{m}} \alpha_{n-p+\mu} - \omega \leq n - p + \bar{m}. \quad (\text{N.25})$$

If the first left-hand sum in this formula vanishes we have, by (N.24), $\omega = 0$. If this sum is > 0 , the first left-hand term in (N.25) is $\geq n-p$ and we have therefore

$$-\omega \leq \bar{m}, \quad \omega \geq -\bar{m} = n - p - m,$$

and this is also true if $\omega = 0$.

We see that every term in (N.21) is $o((x-\zeta)^{n-p-m})$.

Lemma 5 is proved.

13. Theorem. Assume that $f(x)$ is continuous with its first $n-1$ derivatives in a neighborhood $U(\zeta)$ of ζ where U can be also a one-sided neighborhood of ζ but is assumed to contain ζ . Assume further that $f^{(n)}(\zeta)$ exists and that we have for a natural p , $1 < p < n$:

$$f(\zeta) = f'(\zeta) = \dots = f^{(p-1)}(\zeta) = 0, \quad f^{(p)}(\zeta) \neq 0, \quad (\text{N.26})$$

so that $f(x)$ has in ζ a root of the exact multiplicity p . We have then

$$f(x) = u(x)^p, \quad (\text{N.27})$$

where $u(x)$ is continuous with its first $n-p$ derivatives in a neighborhood U_1 of ζ , where U_1 can be chosen as a part of U containing ζ inside if U does so. Further, $u^{(n-p+1)}(\zeta)$ exists.

14. Proof. We have obviously

$$f(x) = (x - \zeta)^p F(x), \quad F(x) = F(\zeta)(1 + \omega(x)), \quad \omega(\zeta) = 0, \quad (\text{N.28})$$

where $F(x)$ and therefore also $\omega(x)$ have the property $B_{n,p}$ of Lemma 4, since $f(x)$ has the property $A_{n,p}$ of this lemma.[†] We have further obviously

$$u(x) = \alpha(x - \zeta)(1 + \omega^*(x)), \quad \alpha^p = F(\zeta), \quad (1 + \omega^*(x))^p = 1 + \omega(x). \quad (\text{N.29})$$

15. Observe now that the function of ω ,

$$z(\omega) = (1 + \omega)^{1/p} = \sum_{v=0}^{\infty} \binom{1/p}{v} \omega^v,$$

is analytic as long as $|\omega| < 1$. Denote by U_1 a neighborhood of ζ , a part of U , in which $|\omega(x)| < 1$ and which contains ζ in its interior if U does so; then it follows from Lemma 5 that $z(\omega(x)) = 1 + \omega^*(x)$ has the property $B_{n,p}$ of Lemma 4 in U_1 .

In particular we see that $\omega^{*(\kappa)}(x)$ exists and is continuous in U_1 for $\kappa = 0, 1, \dots, n-p$. The same holds then also for the function $G(x) = (1 + \omega^*(x))$. But we have

$$u(x) = (x - \zeta) G(x).$$

The assertion of the theorem can now be obtained from Lemma 4 if we change the notation there correspondingly. However, we prefer giving a direct proof.

That the first $n-p$ derivatives of $u(x)$ exist and are continuous in U_1 follows by direct differentiation. In particular we have

$$u^{(n-p)}(x) = (x - \zeta) G^{(n-p)}(x) + (n-p) G^{(n-p-1)}(x).$$

It follows then that $u^{(n-p)}(\zeta) = (n-p) G^{(n-p-1)}(\zeta)$, and therefore

$$\frac{u^{(n-p)}(x) - u^{(n-p)}(\zeta)}{x - \zeta} = G^{(n-p)}(x) + (n-p) \frac{G^{(n-p-1)}(x) - G^{(n-p-1)}(\zeta)}{x - \zeta}.$$

Hence we obtain as $x \rightarrow \zeta$

$$u^{(n-p+1)}(\zeta) = (n-p+1) G^{(n-p)}(\zeta).$$

Our theorem is now proved.

The reader will easily recognize that the proof which we gave of our theorem can be simplified insofar as only a part of Lemma 4 and only a very elementary part of Lemma 5 are needed. However, the lemmata as we proved them allow a better insight into the background of the whole situation.

[†] If ζ lies inside U , Lemma 4 has to be applied to both one-sided neighborhoods of ζ into which U is decomposed by ζ .

16. Under the conditions of our theorem the development (2.19) can now be applied to $u(x)$ if we replace in (2.19) the letter y by the letter u . However, for practical use the corresponding coefficients must be expressed through the derivatives of $y = f(x)$. We obtain then for the first coefficients in (2.20)

$$\begin{aligned} k &= -\frac{u}{u'} = -p \frac{y}{y'}, \\ X_1 &= 1, \quad \frac{1}{2!} X_2 = -\frac{1}{2} \frac{u''}{u'} = \frac{(p-1)y'^2 - pyy''}{2pyy'}, \\ \frac{1}{3!} X_3 &= \frac{3u''^2 - u'u'''}{6u'^2} \\ &= \frac{(p-2)(p-1)y'^4 - 3p(p-1)yy'^2y'' - p^2y^2y'y''' + 3p^2y^2y''^2}{6p^2y^2y'^2}. \end{aligned}$$

Instead of (2.20) we obtain in our case

$$\begin{aligned} \zeta - x &= k + \frac{(p-1)y'^2 - pyy''}{2pyy'} k^2 \\ &\quad + \frac{(p-2)(p-1)y'^4 - 3p(p-1)yy'^2y'' - p^2y^2y'y''' + 3p^2y^2y''^2}{6p^2y^2y'^2} k^3 \\ &\quad + \frac{1}{4!} X_4 k^4 + O((x-\zeta)^5). \end{aligned} \tag{N.30}$$

17. For certain purposes we need the values of $X_v(u'/u', u''/u', \dots, u^{(v)}/u')$ for $x = \zeta$. These values are obtained in the simplest way by developing u directly in powers of $x - \zeta$.

Indeed, it is clear that the values of the X_v can be obtained by direct differentiation of $u = y^{1/p}$, independently of the analytic character of $f(x)$. But then we obtain the same expressions as in the case where $f(x)$ is analytic in ζ .

18. Now write, putting $x - \zeta = \xi$,

$$f(x) = \xi^p \alpha (1 + \beta \xi + \gamma \xi^2 + \delta \xi^3 + O(\xi^4)),$$

where

$$\begin{aligned} \alpha &= \frac{f^{(p)}(\zeta)}{p!}, & \beta &= \frac{f^{(p+1)}(\zeta)}{f^{(p)}(\zeta)} \frac{1}{p+1}, \\ \gamma &= \frac{f^{(p+2)}(\zeta)}{(p+1)(p+2)f^{(p)}(\zeta)}, & \delta &= \frac{f^{(p+3)}(\zeta)}{(p+1)(p+2)(p+3)f^{(p)}(\zeta)}. \end{aligned} \tag{N.31}$$

Then we have

$$\begin{aligned} (1 + \beta\xi + \gamma\xi^2 + \delta\xi^3 + O(\xi^4))^{1/p} &= 1 + \xi \frac{\beta}{p} + \xi^2 \left(\frac{\gamma}{p} + \frac{\beta^2(1-p)}{2p^2} \right) \\ &\quad + \xi^3 \left(\frac{\delta}{p} + \frac{\beta\gamma}{p^2}(1-p) \right. \\ &\quad \left. + \frac{\beta^3}{6p^3}(1-p)(1-2p) \right) + O(\xi^4). \end{aligned}$$

Since $u = \alpha^{1/p}\xi(1 + \beta\xi + \gamma\xi^2 + \delta\xi^3 + O(\xi^4))^{1/p}$ (with the choice of $\alpha^{1/p}$ corresponding to u), we obtain

$$u' = \alpha^{1/p}, \quad \frac{u''}{u'} = \frac{2\beta}{p} = \frac{2}{p(p+1)} \frac{f^{(p+1)}(\zeta)}{f^{(p)}(\zeta)} \quad (\text{N.32})$$

$$\begin{aligned} \frac{u'''}{u'} &= 3 \frac{2p\gamma - (p-1)\beta^2}{p^2} \\ &= 3 \frac{2p(p+1)f^{(p)}(\zeta)f^{(p+2)}(\zeta) - (p-1)(p+2)(f^{(p+1)}(\zeta))^2}{(p+1)^2(p+2)p^2f^{(p)}(\zeta)^2}, \quad (\text{N.33}) \end{aligned}$$

$$\begin{aligned} \frac{u^{(4)}}{u'} &= 4 \frac{(p-1)(2p-1)\beta^3 - 6p(p-1)\beta\gamma + 6p^2\delta}{p^3} \\ &= 4 \left(\frac{(p-1)(2p-1)(p+2)(p+3)f^{(p+1)}(\zeta)^3}{p^3(p+3)(p+1)^3(p+2)f^{(p)}(\zeta)^3} \right. \\ &\quad - \frac{6p(p^2-1)(p+3)f^{(p)}(\zeta)f^{(p+1)}(\zeta)f^{(p+2)}(\zeta)}{p^3(p+3)(p+1)^3(p+2)f^{(p)}(\zeta)^3} \\ &\quad \left. + \frac{6p^2(p+1)^2f^{(p)}(\zeta)^2f^{(p+3)}(\zeta)}{p^3(p+3)(p+1)^3(p+2)f^{(p)}(\zeta)^3} \right). \quad (\text{N.34}) \end{aligned}$$

For the corresponding values of X_1, \dots, X_4 we have now

$$\begin{aligned} X_1 &= 1, \quad X_2 = -\frac{2}{p(p+1)} \frac{f^{(p+1)}(\zeta)}{f^{(p)}(\zeta)}, \\ \frac{1}{3}p^2(p+1)^2(p+2)f^{(p)}(\zeta)^2X_3 &= (p+2)(p+3)f^{(p+1)}(\zeta)^2 \\ &\quad - 2p(p+1)f^{(p)}(\zeta)f^{(p+2)}(\zeta), \\ \frac{1}{8}p^3(p+1)^3(p+2)(p+3)f^{(p)}(\zeta)^3X_4 &= 3p(p+1)(p+3)(p+4)f^{(p)}(\zeta)f^{(p+1)}(\zeta)f^{(p+2)}(\zeta) \\ &\quad - (p+2)^2(p+3)(p+4)f^{(p+1)}(\zeta)^3 - 3p^2(p+1)^2f^{(p)}(\zeta)^2f^{(p+3)}(\zeta). \end{aligned}$$

O

Laguerre Iterations

1. Consider n real quantities a_1, \dots, a_n and form with them

$$a_1 + \dots + a_n = a, \quad a_1^2 + \dots + a_n^2 = b. \quad (\text{O.1})$$

Denoting by N an arbitrary number $\geq n$ we have by the Cauchy-Schwarz inequality

$$\begin{aligned} (a - a_1)^2 &= (a_2 + \dots + a_n)^2 \\ &\leq (n-1)(a_2^2 + \dots + a_n^2) \leq (N-1)(b - a_1^2). \end{aligned}$$

The inequality between the first and the last term of this relation can be written simplified as

$$Na_1^2 - 2aa_1 + a^2 - (N-1)b \leq 0.$$

We see that the equation

$$\varphi_N(u) \equiv Nu^2 - 2au + a^2 - (N-1)b = 0 \quad (\text{O.2})$$

has two real roots

$$u_{1,2} = \frac{1}{N} \left(a \pm \sqrt{(N-1)(Nb - a^2)} \right) \quad (\text{O.3})$$

and that if $u_1 \leq u_2$, we have

$$u_1 \leq a_1 \leq u_2. \quad (\text{O.4})$$

2. We will now discuss the behavior of u_1, u_2 with growing N . We have obviously in (O.2), for $u = u_1$ or $u = u_2$,

$$\varphi_N(u) = N(u^2 - b) - (2au - a^2 - b) = 0. \quad (\text{O.5})$$

We are going to prove that the expression $u^2 - b$ in this formula is always ≤ 0 .

Since b is ≥ 0 , we can assume in this proof that $a \neq 0$, $|a| = A$, $a = \varepsilon A$, as otherwise clearly $u^2 - b \leq 0$.

Multiplying both sides of (O.5) by N we have

$$N^2(u^2 - b) = 2aNu - Na^2 - Nb.$$

Put

$$\alpha := Nb - A^2, \quad \beta := (N-1)A^2; \quad (\text{O.6})$$

α is certainly ≥ 0 , since the roots (O.3) are real.

Then we have

$$N^2(u^2 - b) = 2aNu - (\alpha + \beta) - 2A^2. \quad (\text{O.7})$$

It is sufficient to take as u in (O.5) the root with the greatest modulus and this root is, as follows from (O.3) and (O.6),

$$u = \frac{\varepsilon}{N}(A + \sqrt{(N-1)\alpha}), \quad \varepsilon = \operatorname{sgn} a.$$

Then we have further

$$2aNu = 2A^2 + 2\sqrt{\alpha\beta}$$

and, introducing this into (O.7), we get

$$N^2(u^2 - b) = -(\sqrt{\alpha} - \sqrt{\beta})^2 \leq 0,$$

which proves our assertion about $u^2 - b$ in (O.5).

3. Consider now an $\bar{N} > N$ and a root \bar{u} of the corresponding equation $\varphi_{\bar{N}}(u) = 0$. Then we have

$$\bar{N}(\bar{u}^2 - b) = 2a\bar{u} - a^2 - b$$

and it follows therefore for $\varphi_N(\bar{u})$ that

$$\varphi_N(\bar{u}) = N(\bar{u}^2 - b) - (2a\bar{u} - a^2 - b) = (N - \bar{N})(\bar{u}^2 - b) \geq 0,$$

$$\varphi_N(u) = (\bar{N} - N)(u^2 - b) \leq 0.$$

But then \bar{u} certainly does not lie strictly between the both bounds u_1, u_2 in (O.4) and u not outside the interval $\langle \bar{u}_1, \bar{u}_2 \rangle$.

We see that *with increasing N the bound u_2 in (O.4) increases and u_1 decreases*.

4. We consider now again as in Sections 1 and 3 of Chapter 15 the polynomial $f(x)$ in (15.1) with the real roots (15.2) and take as the a_v in (O.1) the single terms of (15.6), $1/(x - \zeta_v)$, in an arbitrary order. Then we have from (15.6) and (15.7)

$$a = \frac{f'(x)}{f(x)}, \quad b = \frac{f'(x)^2 - f(x)f''(x)}{f(x)^2}.$$

By (O.1), in this case always $b > 0$. Now (O.4) gives the inequality, valid for any ζ_v ,

$$u_1 \leq \frac{1}{x - \zeta_v} \leq u_2 \quad (\text{O.8})$$

where u_1 is the smaller and u_2 the greater of the roots given by

$$\begin{aligned} u_{1,2} &= \frac{1}{Nf(x)} [f'(x) \pm \sqrt{(N-1)((N-1)f'(x)^2 - Nf(x)f''(x))}], \\ u_1(x) &\leq u_2(x). \end{aligned} \quad (\text{O.9})$$

Since $b > 0$, none of the u_1, u_2 can vanish.

5. We put in what follows

$$\begin{aligned} v_1(x) &= \frac{1}{Nf(x)} [f''(x) - \sqrt{(N-1)((N-1)f'(x)^2 - Nf(x)f''(x))}], \\ v_2(x) &= \frac{1}{Nf(x)} [f''(x) + \sqrt{(N-1)((N-1)f'(x)^2 - Nf(x)f''(x))}]. \end{aligned} \quad (\text{O.10})$$

Assume now that x lies between consecutive distinct roots ζ_v, ζ_{v+1} , that is,

$$\zeta_v < x < \zeta_{v+1}, \quad (\text{O.11})$$

and that $f'(x) \neq 0$. Then it follows from (O.8), since the $x - \zeta_\kappa$ are positive as well as negative,

$$u_1(x) < 0 < u_2(x). \quad (\text{O.12})$$

Further, applying the first inequality (O.8) to ζ_{v+1} and the second to ζ_v , we have

$$\zeta_{v+1} > x - \frac{1}{u_1(x)}, \quad \zeta_v < x - \frac{1}{u_2(x)}. \quad (\text{O.13})$$

6. It follows now from (O.13), since $u_1 < 0, u_2 > 0$, that both numbers $x - 1/u_1, x - 1/u_2$ still lie in the interval (ζ_v, ζ_{v+1}) and therefore the same argument holds for both of them taken instead of x .

Form now the two sequences y_κ', z_κ' by the iteration rules:

$$y_0' = x, \quad y_{\kappa+1}' = y_\kappa' - \frac{1}{u_2(y_\kappa')} \quad (\kappa = 0, 1, \dots), \quad (\text{O.14})$$

$$z_0' = x, \quad z_{\kappa+1}' = z_\kappa' - \frac{1}{u_1(z_\kappa')} \quad (\kappa = 0, 1, \dots). \quad (\text{O.15})$$

Then it follows from our discussion that the y_κ' monotonically decrease and remain $> \zeta_v$, while the z_κ' monotonically increase and remain $< \zeta_{v+1}$. Both sequences are therefore convergent and, since by (O.9) in the limit points $f(x)$ must vanish, we see that

$$y_\kappa' \downarrow \zeta_v, \quad z_\kappa' \uparrow \zeta_{v+1}.$$

7. The formation of the sequences (O.14), (O.15) requires, however, at any κ , the determination of the sign at the square root in (O.9) corresponding to the $u_2(y_\kappa')$ and $u_1(z_\kappa')$.

On the other hand, since $f(x)$ keeps constant sign in the interval (ζ_v, ζ_{v+1}) , it is clear that a fixed sign at the square root corresponds to all $u_2(y_\kappa')$ and the opposite sign to all $u_1(z_\kappa')$. We can therefore formulate our result as

Theorem 1. Assume that for a polynomial $f(x)$ of degree n and without complex roots, x lies between two consecutive roots ζ, η and $f'(x) \neq 0$. Assume a fixed $N \geq n$ and form the two sequences y_κ, z_κ by the iteration rules

$$y_0 = z_0 = x, \quad y_{\kappa+1} = y_\kappa - \frac{1}{v_1(y_\kappa)}, \quad z_{\kappa+1} = z_\kappa - \frac{1}{v_2(z_\kappa)} \quad (\kappa = 0, 1, \dots), \quad (O.16)$$

where v_1, v_2 are defined by (O.10).

Then the sequences y_κ, z_κ remain between ζ and η and one of them converges monotonically to ζ while the other goes monotonically to η .

It is seen immediately that if $f(x) > 0$ between ζ and η then the y_κ tend to $\text{Max}(\zeta, \eta)$ and the z_κ tend to $\text{Min}(\zeta, \eta)$, while these limits must be interchanged if $f(x) < 0$ between ζ and η .

8. Assume now that we have $x > \zeta_n$. Then $u_2(x)$ is > 0 and we have, using (O.8) for $v = n$,

$$x - \frac{1}{u_2(x)} > \zeta_n.$$

It follows now immediately that the sequence (O.14), starting with our x , tends decreasingly to ζ_n . In a completely analogous manner, we see, if $x < \zeta_1$, that the sequence (O.15), starting with x , tends increasingly to ζ_1 .

9. Assume again $x > \zeta_n$. What can be said about the sequence (O.15)?

If $u_1(x) > 0$, then it follows from (O.8), applied to $v = 1$, that we have $x - 1/u_1(x) < \zeta_1$ and we see that the sequence (O.15) starting with this value tends to ζ_1 , that is to say, the sequence (O.15) starting with x tends to ζ_1 monotonically from a certain κ on.

If $u_1(x) < 0$ then we have in (O.15) $z_1' > z_0 = x > \zeta_n$. We see that as long as the $u_1(z_\kappa')$ remain < 0 the z_κ' are monotonically increasing.

If now all $u_1(z_\kappa')$ remain <0 , the z_κ' form a monotonically increasing sequence which, however, cannot have a finite limit because this limit would be a zero $>\zeta_n$ of $f(x)$. Therefore, the z_κ' tend then to infinity. On the other hand, since by (O.8) $u_2(z_\kappa')$ remains >0 , the product $u_1(x)u_2(x)$ is then <0 for all sufficiently great x . For this product, however, we have

$$u_1(x)u_2(x) = \frac{(N-1)f(x)f''(x) - (N-2)f'(x)^2}{Nf(x)^2} \quad (\text{O.17})$$

and this expression is equivalent, with $x \rightarrow \infty$, to

$$n \frac{n+1-N}{Nx^2},$$

as long as $N \neq n+1$.

10. We see that if $n \leq N < n+1$, $u_1(x)u_2(x)$ becomes >0 for sufficiently large x , so that in this case, too, the sequence (O.15) starting with x converges to ζ_1 , monotonically increasing from a certain κ on.

If, however, $N > n+1$, we see that if we start with a sufficiently large x the sequence (O.15) tends monotonically to ∞ .

If, finally, $N = n+1$, the asymptotic behavior of (O.17) depends on the sign of the next coefficient of $f(x)$, that is, on the sign of the sum $\zeta_1 + \dots + \zeta_n$.

We have not yet considered the case in which a $u_1(z_\kappa')$ vanishes. In this case the sequence (O.15) breaks down.

We collect the essential part of our results in:

Theorem 2. *If under the hypothesis of Theorem 1, x is assumed outside of an interval containing all roots of $f(x)$, then both sequences (O.16) converge, one to the smallest and one to the largest root of $f(x)$ as long as $n \leq N < n+1$. The convergence is monotonic from a certain κ on.*

If $N > n+1$, the sequence (O.14) starting with any $x > \zeta_n$ tends decreasingly to ζ_n and the sequence (O.15) starting with any $x < \zeta_1$ tends increasingly to ζ_1 .

11. We assume now that one of the sequence (O.16) converges to a root of $f(x)$ and will investigate the asymptotic behavior of this sequence. We denote the corresponding expression (O.10) by $u(x)$ and the corresponding sequence y_v or z_v by x_v . We have then obviously, as follows from (O.14) and (O.15), $|u(x_v)| \rightarrow \infty$ and even, since the convergence is monotonic, $u(x_v) \rightarrow \pm \infty$.

On the other hand, observe that in the neighborhood of ζ we certainly have $\operatorname{sgn}(x - \zeta) = \operatorname{sgn}(f(x)/f'(x))$. From the monotony of convergence of the x_v to ζ it follows also by (O.16) that, from a v onward, the $u(x_v)$ have the same sign as $x_v - \zeta$, so that we have finally from a certain v onward

$$\operatorname{sgn}(x_v - \zeta) = \operatorname{sgn} \frac{f(x_v)}{f'(x_v)} = \operatorname{sgn} u(x_v).$$

12. It follows further from (O.10) that

$$u(x) = \frac{f'(x)}{Nf(x)} \left[1 \pm (N-1) \sqrt{1 - \frac{N}{N-1} \frac{f(x)f''(x)}{f'(x)^2}} \right]. \quad (\text{O.18})$$

We assume from now on that $N > 2$ and further that ζ is a simple zero of $f(x)$. Then for $x = x_v \rightarrow \zeta$ the square root in (O.18) tends to 1. Since for $x \rightarrow \zeta$, $f'(x)/f(x) \sim 1/(x - \zeta)$, we see that $1/u(x)$ is, in the case of the plus sign, $\sim x - \zeta$, and, in the case of the minus sign, $\sim -[N/(N-2)](x - \zeta)$. But then it follows that

$$\frac{x - 1/u(x) - \zeta}{x - \zeta}$$

tends, in the case of the plus sign, to 0, and, in the case of the minus sign, to $1 + N/(N-2) > 1$. We see that ζ is, in the case of the minus sign, a point of repulsion for our iteration while, in the case of the plus sign, we have even a superlinear convergence.

We must therefore have for our $u(x)$ the plus sign. Putting

$$\sigma(x) = \frac{f(x)f''(x)}{f'(x)^2}, \quad (\text{O.19})$$

we have therefore

$$u(x) = \frac{1}{N} \frac{f'(x)}{f(x)} \left[1 + (N-1) \sqrt{1 - \frac{N}{N-1} \sigma(x)} \right]. \quad (\text{O.20})$$

13. Since $\sigma(x_v) \rightarrow 0$, we can develop, from a v onward, writing x for x_v , until the following formula (O.25),

$$\begin{aligned} \sqrt{1 - \frac{N}{N-1} \sigma(x)} &= 1 - \frac{1}{2} \frac{N}{N-1} \sigma(x) - \frac{1}{8} \left(\frac{N}{N-1} \right)^2 \sigma(x)^2 + O(\sigma(x)^3), \\ u(x) &= \frac{f'(x)}{f(x)} \left[1 - \frac{1}{2} \sigma(x) - \frac{1}{8} \frac{N}{N-1} \sigma(x)^2 + O(\sigma(x)^3) \right], \\ \frac{1}{u(x)} &= \frac{f(x)}{f'(x)} \left[1 + \frac{1}{2} \sigma(x) + \frac{1}{8} \frac{N}{N-1} \sigma(x)^2 + \frac{1}{4} \sigma(x)^2 + O(\sigma(x)^3) \right], \\ \frac{1}{u(x)} &= \frac{f(x)}{f'(x)} \left[1 + \frac{\sigma(x)}{2} + \frac{3}{8} \left(1 + \frac{1}{3(N-1)} \right) \sigma(x)^2 + O(\sigma(x)^3) \right]. \end{aligned} \quad (\text{O.21})$$

14. We now use the Schröder series (2.20), replacing k by $-f(x)/f'(x)$, and y and the derivatives of y , respectively, by $f(x), f'(x), f''(x), f'''(x)$. Then

we obtain

$$x - \zeta = \frac{f(x)}{f'(x)} + \frac{1}{2} \frac{f''(x)f(x)^2}{f'(x)^3} + \frac{3f''(x)^2 - f'(x)f'''(x)}{6f'(x)^5} f(x)^3 + O(f(x)^4).$$

Again using the notation (O.19) and observing that $f(x) = O(x - \zeta)$ we see that this becomes

$$x - \zeta = \frac{f(x)}{f'(x)} \left[1 + \frac{1}{2}\sigma(x) + \frac{3f''(x)^2 - f'(x)f'''(x)}{6f'(x)^2} \left(\frac{f(x)}{f'(x)} \right)^2 \right] + O((x - \zeta)^4). \quad (\text{O.22})$$

Subtracting from this (O.21) and replacing $\sigma(x)^2$ by the square of the expression (O.19), we have

$$\begin{aligned} x - \frac{1}{u(x)} - \zeta &= \frac{f(x)^3}{f'(x)^3} \left[\frac{3f''(x)^2 - f'(x)f'''(x)}{6f'(x)^2} - \left(\frac{3}{8} + \frac{1}{8(N-1)} \right) \frac{f''(x)^2}{f'(x)^2} \right] \\ &\quad + O((x - \zeta)^4) \\ &= \left(\frac{f(x)}{f'(x)} \right)^3 \frac{1}{24f'(x)^2} \left[3f''(x)^2 - 4f'(x)f'''(x) - \frac{3}{N-1} f''(x)^2 \right] \\ &\quad + O((x - \zeta)^4). \end{aligned}$$

Now putting

$$f_0' := f'(\zeta) \neq 0, \quad f_0'' := f''(\zeta), \quad f_0''' := f'''(\zeta), \quad (\text{O.23})$$

we obtain

$$\frac{x - (1/u(x)) - \zeta}{(x - \zeta)^3} \rightarrow \frac{1}{24f_0'^2} \left[3f_0''^2 - 4f_0'f_0''' - \frac{3}{N-1} f_0''^2 \right] \quad (x \rightarrow \zeta). \quad (\text{O.24})$$

Replacing x in formula (O.24) by x_v and observing that the left-hand numerator becomes $x_{v+1} - \zeta$, we have

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^3} \rightarrow \frac{1}{24f_0'^2} \left[3f_0''^2 - 4f_0'f_0''' - \frac{3}{N-1} f_0''^2 \right]. \quad (\text{O.25})$$

15. In formulas (O.24), (O.25) we have, of course, $N \geq n$. We are now going to prove that the right-hand expression in these formulas is never negative. We can assume $n \geq 3$.

Writing $x - \zeta = y$ and denoting $f(x)/(x - \zeta)$ by $g(y)$, we have

$$yg(y) = f_0'y + \frac{1}{2}f_0''y^2 + \frac{1}{6}f_0'''y^3 + \dots,$$

$$g(y) = f_0' + \frac{1}{2}f_0''y + \frac{1}{6}f_0'''y^2 + \dots.$$

The polynomial $g(y)$ has $n - 1$ real roots. If we write it with binomial coefficients as

$$g(y) = a + \left(\frac{n-1}{1}\right)by + \left(\frac{n-1}{2}\right)cy^2 + \dots,$$

we have, by Newton's inequality (B.30) of Appendix B,

$$b^2 \geq ac. \quad (\text{O.26})$$

On the other hand, we have

$$a = f_0', \quad b = \frac{f_0''}{2(n-1)}, \quad c = \frac{f_0'''}{3(n-1)(n-2)}.$$

Introducing these values into (O.26), we get

$$3f_0''^2(n-1)(n-2) \geq 4f_0'f_0'''(n-1)^2$$

and dividing on both sides by $(n-1)^2$,

$$3f_0''^2\left(1 - \frac{1}{n-1}\right) \geq 4f_0'f_0''',$$

and this inequality remains *a fortiori* true if we replace in it n by any $N > n$:

$$3f''(\zeta)^2 - 4f'(\zeta)f'''(\zeta) \geq \frac{3}{N-1}f''(\zeta)^2 \quad (N \geq n). \quad (\text{O.27})$$

16. If $f_0'' = 0$, the limit in (O.24), (O.25) does not depend on N . But if $f_0'' \neq 0$, we see that the right-hand limit in both relations is positive and monotonically increasing with the increasing $N > n$. Thus this limit is the smallest for $N = n$ and the largest for $N = \infty$.

Thus, Laguerre's formula, corresponding to $N = n$, is asymptotically the best, while formula (15.10) corresponding to $N = \infty$ is asymptotically the worst.

However, because the convergence is cubic anyway, this does not matter very much in numerical computation, since the formula (15.10) is simpler to apply and is not restricted to polynomials of a fixed degree.

As a matter of fact, it has already been proved in Section 3 that the single steps become smaller with growing N , if we start from the same x . However, this result only states something about a purely "tactical" situation, while our discussion of Section 15 confirms this from the "strategic" point of view.

17. We consider finally the case where ζ is a multiple root of multiplicity $p > 1$. In this case, we have for $x \rightarrow \zeta$

$$f(x) \sim \alpha(x - \zeta)^p, \quad \alpha = \frac{f^{(p)}(\zeta)}{p!} \neq 0,$$

$$f'(x) \sim p\alpha(x - \zeta)^{p-1}, \quad f''(x) \sim p(p-1)\alpha(x - \zeta)^{p-2}.$$

It follows with $x \rightarrow \zeta$ that

$$\frac{f(x)}{f'(x)} \sim \frac{1}{p}(x - \zeta), \quad \frac{f(x)f''(x)}{f'(x)^2} \rightarrow 1 - \frac{1}{p}. \quad (\text{O.28})$$

We have now from (O.18)

$$N(x - \zeta)u(x) \rightarrow [1 \pm \sqrt{(N-1)(N-1-N(1-1/p))}],$$

$$\frac{1}{u(x)} / (x - \zeta) \rightarrow \frac{N}{p \pm w}, \quad w = \sqrt{(N-1)(N-p)p}. \quad (\text{O.29})$$

If we have $p = N$, then we must obviously have $N = n$,

$$f(x) = \alpha(x - \zeta)^n, \quad f'(x) = \alpha n(x - \zeta)^{n-1},$$

$$f''(x) = \alpha(n-1)n(x - \zeta)^{n-2},$$

w vanishes, and we have $1/u(x) = x - \zeta$, so that already $x - 1/u(x) = \zeta$ and the iteration stops.

If we have $p = N-1$ it follows that $w = N-1$. We have then in (O.29) the plus sign, since otherwise we would have $u(x) = 0$. We obtain then in this case

$$\frac{x - \zeta - 1/u(x)}{x - \zeta} \rightarrow 1 - \frac{N}{2N-2} = \frac{N-2}{2N-2}$$

and the convergence is linear.

We assume from now on that $p < N-1$, so that $w > 0$.

18. We have from (O.29)

$$\frac{x - 1/u(x) - \zeta}{x - \zeta} \rightarrow 1 - \frac{N}{p \pm w} \quad (x \rightarrow \zeta). \quad (\text{O.30})$$

The modulus of this cannot be > 1 , since we would then have divergence. We have therefore in any case

$$\frac{N}{p \pm w} \geq 0, \quad p \pm w > 0.$$

If we had the minus sign at w it would follow that $w < p$, or, squaring,

$$(N-1)(N-p) < p, \quad N < p,$$

while $p < N-1$.

We see that we have the plus sign at w .

19. The limit in (O.30) becomes now

$$\frac{\sqrt{p(N-1)(N-p)} - (N-p)}{\sqrt{p(N-1)(N-p)} + p}. \quad (\text{O.31})$$

This is certainly $\neq 0$, since otherwise we would have $p(N-1) = N-p$, $pN = N$. The expression (O.31) is obviously < 1 . If it were ≤ -1 , we would have $w \leq N/2 - p$ and, squaring,

$$p(N-1)(N-p) \leq p^2 - pN + N^2/4.$$

Thence we obtain after some reductions $p(N-p) \leq N/4$, which is obviously impossible, since one of the left-hand factors is in any case $\geq N/2$.

We see that in the case of a multiple root the convergence of the generalized Laguerre iteration is *strictly linear*.

P

Approximation of Equations by Algebraic Equations of a Given Degree. Asymptotic Errors for Multiple Zeros

1. In this appendix we take up the discussion of the iteration method given in Chapter 18. We use the notations introduced in Chapters 17 and 18, assuming now that the multiplicity p of ζ is ≥ 2 . We assume further that all $|z_v - \zeta|$ ($v = 1, 2, \dots$) are positive. We shall need several lemmas.

2. Lemma 1. Assume n and p as integers with $n > p > 1$ and T as a positive number. Assume a δ with

$$0 < \delta < \frac{1}{2^{n+3}}, \quad \delta < 2^{n+1}T. \quad (\text{P.1})$$

Assume $n+1$ positive numbers $\eta_1, \dots, \eta_n, \eta$ such that we have

$$1 > \frac{\delta}{2^{n+1}T} > \eta_1 \geq \dots \geq \eta_{n-1}, \quad \eta_n < \delta^{p+1}\eta_{n-1}, \quad \eta < \frac{\delta}{2^{n+1}T}, \quad (\text{P.2})$$

$$\eta_n^{p-1} > \frac{T}{4\delta^{p+1}} \prod_{\kappa=1}^{n-1} \eta_\kappa, \quad (\text{P.3})$$

$$\eta^{p-1} \leq \frac{4T}{\delta} \prod_{\kappa=1}^{n-1} (\eta + \eta_\kappa). \quad (\text{P.4})$$

Then

$$\eta < \delta\eta_n. \quad (\text{P.5})$$

3. Proof. Denote by t , $1 \leq t \leq n$, an integer such that

$$\eta_1 \geq \dots \geq \eta_{t-1} > \eta \geq \eta_t \geq \dots \geq \eta_{n-1}, \quad (\text{P.6})$$

where we have $t = 1$ if $\eta \geq \eta_1$ and $t = n$ if $\eta < \eta_{n-1}$. Put, for $\pi = 1, 2, \dots, n-1$,

$$P_\pi = T \prod_{\kappa=1}^{\pi-1} \eta_\kappa \quad (\text{P.7})$$

where the product is 1 for $\pi = 1$.

4. We have from (P.4), using (P.6) and replacing $\eta + \eta_\kappa$ by either $2\eta_\kappa$ or 2η , according as $\kappa < t$ or $\kappa \geq t$,

$$\eta^{p-1} \leq \frac{2^{n+1}T}{\delta} \eta^{n-t} \prod_{\kappa=1}^{t-1} \eta_\kappa.$$

This becomes, if we divide on both sides by η^{n-t} , use (P.7), and put

$$s = p - 1 - (n-t): \quad (\text{P.8})$$

$$\eta^s \leq \frac{2^{n+1}P_t}{\delta}. \quad (\text{P.9})$$

It is now easy to see that we have $s > 0$, $t > 1$. Indeed, for $t = 1$ we would have $P_1 = T$, $s = p-n$, and it would follow from (P.9) that

$$\frac{\delta}{2^{n+1}T} \leq \eta^{n-p} \leq \eta,$$

while the left-hand expression is $> \eta$ by (P.2).

We can therefore assume $t > 1$. But then the right-hand expression in (P.9) is $\leq (2^{n+1}T/\delta)\eta_1$, and this is by (P.2) < 1 . From (P.9) it follows now that $\eta^s < 1$ and this is only possible if $s \geq 1$, since we have $\eta < 1$, by (P.2) and (P.1).

5. Dividing on both sides of (P.3) by η_n^{n-t} , we get, using (P.7),

$$\eta_n^s \geq \frac{P_t}{4\delta^{p+1}} \prod_{\kappa=t}^{n-1} \frac{\eta_\kappa}{\eta_n}$$

or, using (P.2), as each quotient η_κ/η_n is $> \delta^{-(p+1)}$,

$$\eta_n^s > \frac{P_t}{4} \delta^{-(p+1)(n-t+1)}. \quad (\text{P.10})$$

Dividing (P.9) by (P.10) and using (P.8), we have

$$\left(\frac{\eta}{\eta_n}\right)^s < \delta^s 2^{n+3} \delta^{(p+1)(n-t+1)-s-1}.$$

By (P.8) and (P.1), since $(p+1)(n-t+1)-(p+1)+n-t = (p+2)(n-t)$, the right-hand expression is equal to

$$\delta^s (2^{n+3} \delta) \delta^{(p+1)(n-t+1)-(p+1)+n-t} \leq \delta^s (2^{n+3} \delta) < \delta^s.$$

Inequality (P.5) now follows immediately from $s > 1$.

6. Before formulating the next lemma, we will verify that from (P.1) follows

$$(1-\delta)^n > \frac{1}{2}. \quad (\text{P.11})$$

Indeed, by Bernoulli's inequality

$$(1-\delta)^n \geq 1 - n\delta > 1 - \frac{n}{2^{n+3}},$$

so that we have only to prove that $n < 2^{n+2}$ and this follows immediately by induction for any natural n .

7. From now on we put, generally,

$$|z_v - \zeta| =: \xi_v, \quad (\text{P.12})$$

Lemma 2. Assume under the hypotheses of Theorem 18.1 that $p > 1$. Put $|\gamma| =: T$ and assume that δ satisfies (P.1), and N is such that we have

$$\xi_v < \frac{\delta}{2^{n+1}T}, \quad \xi_v \operatorname{Max}(C_1, C_2) \leq \frac{1}{2} \quad (v \geq N). \quad (\text{P.13})$$

Then, if we have for a $v \geq N$,

$$\xi_{v+n+1} < \delta^{p+1} \operatorname{Min}(\xi_{v+1}, \dots, \xi_{v+n}), \quad v \geq N, \quad (\text{P.14})$$

we also have

$$\xi_{v+n+2} \geq \delta \xi_{v+n+1}. \quad (\text{P.15})$$

8. Proof. Without loss of generality we can and will assume $\zeta = 0$. Assume that (P.15) is false, so that we have

$$\xi_{v+n+2} \geq \delta \xi_{v+n+1}. \quad (\text{P.16})$$

From (17.22), we have, replacing ε by $m \equiv \operatorname{Max}(\xi_{v+1}, \dots, \xi_{v+n})$ and since $|\gamma| = T$,

$$\begin{aligned} \xi_{v+n+1}^p &= U_v T \prod_{\kappa=1}^n |z_{v+\kappa} - z_{v+\kappa}|, \\ U_v &= \frac{1 + \theta_2 m C_2}{1 + \theta_1 m C_1}, \quad |\theta_2| \leq 1, \quad |\theta_1| \leq 1. \end{aligned} \quad (\text{P.17})$$

It follows then from the second inequality (P.13) that

$$\frac{1}{2} \leq U_v \leq 2 \quad (v \geq N). \quad (\text{P.18})$$

We have therefore from (P.17) and (P.14)

$$\xi_{v+n+1}^p > \frac{T}{2} \prod_{\kappa=1}^n \xi_{v+\kappa} (1-\delta)^n$$

and, using (P.11) and again (P.14),

$$\xi_{v+n+1}^{p-1} > \frac{T}{4} \frac{\xi_{v+1}}{\xi_{v+n+1}} \prod_{\kappa=2}^n \xi_{v+\kappa}, \quad \xi_{v+n+1}^{p-1} > \frac{T}{4\delta^{p+1}} \prod_{\kappa=2}^n \xi_{v+\kappa}. \quad (\text{P.19})$$

9. On the other hand, replacing v in (P.17) by $v+1$ and using (P.18), we have

$$\xi_{v+n+2}^p \leq 2T|z_{v+n+2} - z_{v+n+1}| \prod_{\kappa=2}^n (\xi_{v+n+2} + \xi_{v+\kappa}).$$

By (P.16) we have here, however,

$$|z_{v+n+2} - z_{v+n+1}| \leq \left(1 + \frac{1}{\delta}\right) \xi_{v+n+2} \leq \frac{2}{\delta} \xi_{v+n+2}$$

and get therefore

$$\xi_{v+n+2}^{p-1} \leq \frac{4T}{\delta} \prod_{\kappa=2}^n (\xi_{v+n+2} + \xi_{v+\kappa}). \quad (\text{P.20})$$

Consider now the $n-1$ numbers $\xi_{v+2}, \dots, \xi_{v+n}$. Arrange them in decreasing order and denote them in this order by $\eta_1, \eta_2, \dots, \eta_{n-1}$. Further, put

$$\xi_{v+n+1} = \eta_n, \quad \xi_{v+n+2} = \eta.$$

Then (P.19) and (P.20) become (P.3) and (P.4). The condition (P.2) of Lemma 1 follows then, too, from the first inequality (P.13) and from (P.14). Therefore, by Lemma 1, we have $\eta < \delta\eta_n$, in contradiction to (P.16). (P.15) and Lemma 2 are proved.

10. Lemma 3. *Under the conditions of Lemma 2 assume N so large that we have*

$$\xi_v < \frac{\delta^{(p+1)^{2n}}}{2^{n+1}|\gamma|} \quad (v \geq N) \quad (\text{P.21})$$

and that, beyond (P.14),

$$\xi_{v+n+1} \leq \delta^{(p+1)^{2n}} \min(\xi_{v+1}, \dots, \xi_{v+n}). \quad (\text{P.22})$$

Then we have

$$\xi_{v+n+\kappa+1} \leq \delta^{(p+1)^{2n-\kappa}} \xi_{v+n+\kappa} \quad (\kappa = 1, 2, \dots, 2n). \quad (\text{P.23})$$

Proof. To prove Lemma 3 it is sufficient to apply $2n$ times Lemma 2, replacing there δ by

$$\delta^{(p+1)^{2n-1}}, \quad \delta^{(p+1)^{2n-2}}, \dots, \delta^{p+1}, \delta.$$

11. In particular, it follows from (P.23) and (P.22) that we have, as soon as (P.21) and (P.22) are satisfied, putting $v+n = \mu$,

$$\xi_{\mu+\kappa+1} \leq \delta \xi_{\mu+\kappa} \quad (\kappa = 0, 1, \dots, 2n). \quad (\text{P.24})$$

It is easy to see that (P.22) is satisfied for infinitely many indices v . Indeed, this will follow from

$$\liminf_{v \rightarrow \infty} \frac{\xi_{v+n+1}}{\text{Min}(\xi_{v+1}, \dots, \xi_{v+n})} = 0. \quad (\text{P.25})$$

If (P.25) were not true, there would exist a positive $P < 1$, such that for all $v \geq 0$

$$\xi_{v+n+1} \geq P \text{Min}(\xi_{v+1}, \dots, \xi_{v+n}).$$

But this again signifies that to every $\mu > n+1$ there exists an index μ' such that

$$\xi_\mu \geq P \xi_{\mu'}, \quad \mu > \mu' \geq \mu - n.$$

Applying this repeatedly we obtain a sequence μ_τ , $\tau = 0, 1, \dots, t$, $\mu_0 = \mu$, such that

$$\begin{aligned} \xi_{\mu_\tau} &\geq P \xi_{\mu_{\tau+1}} & (\tau = 0, 1, \dots, t-1), \\ \mu_\tau &> \mu_{\tau+1} \geq \mu_\tau - n & (\tau = 0, 1, \dots, t-1), \quad \mu_t \leq n+1. \end{aligned}$$

Here we have obviously $t < \mu_0 = \mu$ and

$$\xi_\mu \geq P^t \text{Min}(\xi_1, \dots, \xi_n), \quad \xi_\mu \geq P^\mu \text{Min}(\xi_1, \dots, \xi_{n+1}),$$

in contradiction to (18.3).

We see that (P.22) is satisfied for infinitely many v .

12. Lemma 4. Assume the hypotheses and notations of Theorem 18.1 and $p > 1$, $|\gamma| = T$ as well as (P.13). If for a $\mu \geq N$ we have (P.24), then we have also

$$\xi_{\mu+2n+2} \leq \delta \xi_{\mu+2n+1}. \quad (\text{P.26})$$

Proof. From (P.17) we have by (P.18)

$$\xi_{\mu+n+\pi+2}^p \geq \frac{T}{2} \prod_{\kappa=1}^n |z_{\mu+n+\pi+2} - z_{\mu+\pi+1+\kappa}| \quad (\pi \geq 0), \quad (\text{P.27})$$

$$\xi_{\mu+2n+2}^p \leq 2T \prod_{\kappa=1}^n (\xi_{\mu+2n+2} + \xi_{\mu+n+1+\kappa}). \quad (\text{P.28})$$

13. By (P.24) for $1 \leq \pi \leq n-1$, $1 \leq \kappa \leq n$, we have

$$|z_{\mu+n+\pi+2} - z_{\mu+\pi+1+\kappa}| \geq (1-\delta) \xi_{\mu+\pi+1+\kappa}.$$

Therefore, by (P.11) we have from (P.27)

$$\xi_{\mu+n+\pi+2}^p \geq \frac{T}{4} \prod_{\kappa=1}^n \xi_{\mu+\pi+1+\kappa} \quad (1 \leq \pi \leq n-1).$$

Here we decompose the right-hand product into two subproducts:

$$\prod_{\kappa=1}^n = \prod_{\kappa=1}^{n-\pi} \prod_{\kappa=n-\pi+1}^n.$$

The first right-hand product here is

$$\xi_{\mu+\pi+2} \xi_{\mu+\pi+3} \cdots \xi_{\mu+n+1}.$$

All factors here occur in (P.24) and are $\geq \xi_{\mu+n+1}$, so that

$$\prod_{\kappa=1}^{n-\pi} \geq \xi_{\mu+n+1}^{n-\pi}.$$

As to the second right-hand product, we have

$$\prod_{\kappa=n-\pi+1}^n \xi_{\mu+n+1+\kappa} = \prod_{\sigma=1}^{\pi} \xi_{\mu+n+1+\sigma},$$

writing $\kappa = n - \pi + \sigma$.

Putting generally for $\tau = 0, 1, \dots, n$

$$P_\tau = \frac{T}{4} \prod_{\sigma=1}^{\tau} \xi_{\mu+n+1+\sigma}, \quad (\text{P.29})$$

we obtain finally

$$\xi_{\mu+n+\pi+2}^p > P_\pi \xi_{\mu+n+1}^{n-\pi} \quad (0 < \pi < n). \quad (\text{P.30})$$

14. Now put $\xi := \xi_{\mu+2n+2}$. There exists a uniquely determined integer λ , $0 \leq \lambda \leq n$, such that we have

$$\begin{aligned} \xi &< \xi_{\mu+n+\lambda+1}, & 0 < \lambda &\leq n, \\ \xi &\geq \xi_{\mu+n+\lambda+2}, & 0 &\leq \lambda < n. \end{aligned} \quad (\text{P.31})$$

These inequalities express of course that for $\lambda = 0$ we have $\xi \geq \xi_{\mu+n+2}$, for $\lambda = n$ we have $\xi < \xi_{\mu+2n+1}$, while ξ for $0 < \lambda < n$ lies in the half-open interval between $\xi_{\mu+n+\lambda+2}$ and $\xi_{\mu+n+\lambda+1}$; observe that all ξ_v occurring in relations (P.24) are monotonically decreasing.

15. Considering now (P.28), observe that we have for the general factor of the right-hand product for the λ defined by (P.31)

$$\xi + \xi_{\mu+n+1+\kappa} \leq \begin{cases} 2\xi_{\mu+n+1+\kappa} & (\kappa \leq \lambda) \\ 2\xi & (\kappa > \lambda). \end{cases}$$

The right-hand product in (P.28) is, therefore, using the notation (P.29),

$$\prod_{\kappa=1}^n \leq 2^n \xi^{n-\lambda} \prod_{\kappa=1}^{\lambda} \xi_{\mu+n+1+\kappa} \leq \frac{4}{T} 2^n P_\lambda \xi^{n-\lambda}$$

and we have therefore

$$\xi^p \leq 2^{n+3} P_\lambda \xi^{n-\lambda}. \quad (\text{P.32})$$

For $\lambda = 0$, we have in (P.29) $P_0 = T/4$, so that (P.32) becomes

$$\xi^{n-p} 2^{n+1} T \geq 1, \quad 2^{n+1} T \xi \geq 1,$$

while by (P.13)

$$\xi < \frac{\delta}{2^{n+1} T}, \quad 2^{n+1} T \xi < \delta < 1.$$

We see that $\lambda = 0$ is impossible.

16. Assuming $\lambda < n$, we can replace ξ on the right in (P.32) by $\xi_{\mu+n+\lambda+1}$ and on the left by $\xi_{\mu+n+\lambda+2}$.

We obtain

$$\xi_{\mu+n+\lambda+2}^p \leq 2^{n+3} P_\lambda \xi_{\mu+n+\lambda+1}^{n-\lambda}$$

and, comparing this with (P.30) for $\pi = \lambda$,

$$2^{n+3} \xi_{\mu+n+\lambda+1}^{n-\lambda} > \xi_{\mu+n+1}^{n-\lambda},$$

$$\frac{1}{2^{n+3}} < \left(\frac{\xi_{\mu+n+\lambda+1}}{\xi_{\mu+n+1}} \right)^{n-\lambda} \leq \left(\frac{\xi_{\mu+n+2}}{\xi_{\mu+n+1}} \right)^{n-\lambda} \leq \delta^{n-\lambda} \leq \delta,$$

in contradiction to (P.1). We see that $\lambda = n$, that is,

$$\xi < \xi_{\mu+2n+1}. \quad (\text{P.33})$$

Using (P.33) we have now from (P.28)

$$\xi^p \leq 2^{n+1} T \prod_{\kappa=1}^n \xi_{\mu+n+1+\kappa}.$$

On the other hand, putting in (P.30) $\pi = n-1$, we obtain

$$\xi_{\mu+2n+1}^p > \xi_{\mu+n+1} P_{n-1} = \frac{T}{4} \xi_{\mu+n+1} \prod_{\kappa=1}^{n-1} \xi_{\mu+n+1+\kappa},$$

and dividing these two inequalities term by term, by virtue of (P.24),

$$\frac{\xi^p}{\xi_{\mu+2n+1}^p} \leq 2^{n+3} \frac{\xi_{\mu+2n+1}}{\xi_{\mu+n+1}} \leq 2^{n+3} \delta^n \leq (2^{n+3} \delta) \delta^p,$$

and (P.26) follows from $\delta < 1/2^{n+3}$. Lemma 4 is proved.

Applying Lemma 4 repeatedly we see now that under the hypotheses of Lemma 4 we have from a certain v_0 onward

$$\xi_{v+1} \leq \delta \xi_v \quad (v \geq v_0),$$

and since $\delta > 0$ can be assumed arbitrarily small, finally, under the hypotheses of Theorem 18.1,

$$\frac{\xi_{v+1}}{\xi_v} \rightarrow 0 \quad (v \rightarrow \infty). \quad (\text{P.34})$$

17. Put now in (P.17)

$$\Gamma = T^{1/(n-p)} = \left(\frac{p!}{n!} \frac{|f^{(n)}(\zeta)|}{|f^{(p)}(\zeta)|} \right)^{1/(n-p)} \quad (\text{P.35})$$

Then we have by (P.34)

$$\xi_{v+n+1}^p = \Gamma^{n-p} (1 + O(\xi_{v+1})) \prod_{\kappa=1}^n \xi_{v+\kappa} \prod_{\kappa=1}^n \left(1 + O\left(\frac{\xi_{v+n+1}}{\xi_{v+\kappa}}\right) \right).$$

We can write this as

$$\xi_{v+n+1}^p = \Gamma^{n-p} \prod_{\kappa=1}^n \xi_{v+\kappa} (1 + \varepsilon_v) \quad (\text{P.36})$$

where

$$\varepsilon_v = O\left(\xi_{v+1} + \sum_{\kappa=1}^n \frac{\xi_{v+n+1}}{\xi_{v+\kappa}}\right) \rightarrow 0. \quad (\text{P.37})$$

Now put for $v = 1, 2, \dots$

$$t_v = -\log(\Gamma \xi_v) \rightarrow \infty, \quad w_v = t_{v+1} - t_v \rightarrow \infty, \quad v_v = w_v - 1. \quad (\text{P.38})$$

Then we obtain from (P.36)

$$pt_{v+n+1} - \sum_{\kappa=1}^n t_{v+\kappa} = O(\varepsilon_v). \quad (\text{P.39})$$

Replacing v in this formula by $v-1$ and subtracting, we get

$$pw_{v+n} - \sum_{\kappa=1}^n w_{v+\kappa-1} = o(1).$$

Introducing here the v_v by (P.38) and $\kappa-1$ instead of κ as the summation variable, we have finally

$$pv_{v+n} - \sum_{\kappa=0}^{n-1} v_{v+\kappa} = o(1) + n - p.$$

18. Since $v_v \rightarrow \infty$ it follows now that from a certain v onward, $v \geq v_0$, all v_v are positive and we have

$$pv_{v+n} - \sum_{\kappa=0}^{n-1} v_{v+\kappa} > 0 \quad (v \geq v_0).$$

Putting then $v_{v_0+v} = u_v$, the conditions of Theorem 12.3 are satisfied for the sequence u_v and we have therefore, for an $\alpha > 0$ and a σ which is, by (13.17), > 1 ,

$$u_v \geq \alpha \sigma^v \quad (v = 1, 2, \dots),$$

that is,

$$v_v \geq \frac{\alpha}{\sigma^{v_0}} \sigma^v \quad (v \geq v_0).$$

But then it follows from (P.38) that for a certain positive β we have

$$w_v \geq \beta \sigma^v \quad (v \geq v_0),$$

$$\frac{\xi_v}{\xi_{v+1}} \geq \exp(\beta \sigma^v) \quad (v \geq v_0),$$

and we see that we have

$$\frac{\xi_{v+1}}{\xi_v} = O(s^v)$$

for any arbitrarily small positive s . Since the same holds, by (18.13), for ξ_v , we have in (P.37) $\epsilon_v = O(s^v)$ so that (P.39) becomes

$$t_{v+n+1} - \frac{1}{p} \sum_{\kappa=1}^n t_{v+\kappa} = O(s^v) \quad (\text{P.40})$$

where the positive s can be taken as small as we wish.

19. But here the characteristic polynomial of the difference equation (P.40) is $f_{n,p}(x)$ of Chapter 13, Section 8, and all conditions of Theorem 12.1 are satisfied with $m = 1$, so that we can use (12.16) with $u_1 = \mu_{n,p}$, $u_2 = q_{n,p} < 1$. We obtain

$$t_v = \alpha \mu_{n,p}^v + O(q_{n,p}^v),$$

where α must be > 0 since $t_v \rightarrow \infty$. Using (P.38) we get finally

$$|z_v - \zeta| = \frac{1}{\Gamma} \exp(-\alpha \mu_{n,p}^v) (1 + O(q_{n,p}^v)), \quad (\text{P.41})$$

$$\frac{|z_{v+1} - \zeta|}{|z_v - \zeta|^{\mu_{n,p}}} = \Gamma^{\mu_{n,p}-1} + O(q_{n,p}^v), \quad (\text{P.42})$$

where Γ is given by (P.35) and we have for $\mu_{n,p}$ and $q_{n,p}$ the inequalities (13.20) and (13.21).

Q

Feedback Techniques for Error Estimates

1. We consider in this appendix a sequence x_v convergent to ζ in such a way that the errors obey the rule

$$|\zeta - x_{v+1}| \leq |\zeta - x_v| \varphi(|\zeta - x_v|) \quad (v = 0, 1, \dots) \quad (\text{Q.1})$$

where $\varphi(u)$ is, for positive u , positive and monotonically increasing. Such functions are often $q = \text{const}$, $0 < q < 1$; au , $a > 0$, or, more generally, au^α , $a \wedge \alpha > 0$.

If we try to apply such an error estimate in a computation conducted with d decimals, it can happen that the n th correction $x_{n+1} - x_n$ becomes already of the order 10^{-d} , while the error estimate obtained directly from (Q.1) is still considerably larger than the bound 10^{-d} .

2. It is possible, however, to use (Q.1) in such a way that we finally obtain from (Q.1) by a kind of feedback an estimate of the error in terms of $|x_{n+1} - x_n|$. We explain this method first on the example of $\varphi(u) = q$, $0 < q < 1$.

In this case we have

$$\begin{aligned} \zeta - x_v &= (x_{v+1} - x_v) + (\zeta - x_{v+1}), \\ |\zeta - x_v| &\leq |x_{v+1} - x_v| + |\zeta - x_{v+1}| \leq |x_{v+1} - x_v| + q|\zeta - x_v|, \\ (1-q)|\zeta - x_v| &\leq |x_{v+1} - x_v|, \quad |\zeta - x_v| \leq \frac{|x_{v+1} - x_v|}{1-q}. \end{aligned} \quad (\text{Q.2})$$

3. A more sophisticated example which can be treated rather completely is that of quadratic convergence. Here $\varphi(u)$ is au , with a positive a , and we can write (Q.1) in the form

$$|\zeta - x_{v+1}| \leq a|\zeta - x_v|^2. \quad (\text{Q.3})$$

Put

$$d_v := |\zeta - x_v|, \quad D_v := |x_{v+1} - x_v|. \quad (\text{Q.4})$$

4. We are now going to prove the

Theorem. Assume a positive $\mu \leq \frac{1}{4}$ and

$$aD_v < \mu, \quad ad_v \leq \frac{1}{2}(1 + \sqrt{1 - 4\mu}) =: \alpha_\mu. \quad (\text{Q.5})$$

Then

$$\frac{2}{1 + \sqrt{1+4\mu}} \leq \frac{d_v}{D_v} \leq \frac{2}{1 + \sqrt{1-4\mu}} \quad (\text{Q.6})$$

5. Proof. Since (Q.3) can be written as

$$ad_{v+1} \leq (ad_v)^2, \quad (\text{Q.7})$$

we can in our proof, without loss of generality, assume $a = 1$.

From the identity

$$(\zeta - x_v) - (x_{v+1} - x_v) = \zeta - x_{v+1}$$

it follows by the triangle inequality that

$$|d_v - D_v| \leq d_{v+1}, \quad d_v - D_v = \theta^* d_v^2, \quad -1 \leq \theta^* \leq 1. \quad (\text{Q.8})$$

If $\theta^* = 0$, then $d_v = D_v$. Otherwise it follows from the quadratic equation

$$\theta^* d_v^2 - d_v + D_v = 0 \quad (\text{Q.9})$$

that

$$d_v = \frac{1 \pm \sqrt{1-4\theta^*D_v}}{2\theta^*} = \frac{2D_v}{1 \mp \sqrt{1-4\theta^*D_v}}. \quad (\text{Q.10})$$

The assumption

$$\begin{aligned} d_v &= \frac{2D_v}{1 - \sqrt{1-4\theta^*D_v}} > \frac{2D_v}{1 - \sqrt{1-4D_v}} \\ &= \frac{1}{2}(1 + \sqrt{1-4D_v}) \geq \frac{1}{2}(1 + \sqrt{1-4\mu}) \end{aligned}$$

contradicts (Q.5). We have therefore

$$d_v = \frac{1 - \sqrt{1-4\theta^*D_v}}{2\theta^*} = \frac{2D_v}{1 + \sqrt{1-4\theta^*D_v}}$$

and (Q.6) follows immediately.

6. The left-hand side of formula (Q.6) is easier to apply if we write it in the form

$$\frac{d_v}{D_v} \geq 1 - \tau_\mu \mu \quad (\text{Q.11})$$

where τ_μ depends on μ .

Comparing (Q.11) with (Q.6), it follows that

$$\tau_\mu = \frac{1}{\mu} \frac{\sqrt{1+4\mu}-1}{\sqrt{1+4\mu}+1} = \frac{4}{(1+\sqrt{1+4\mu})^2}. \quad (\text{Q.12})$$

We see that τ_μ is monotonically increasing as μ decreases. For $\mu = \frac{1}{4}$ we obtain $\tau_{1/4} = 0.687$ (as the rounded up value) and therefore, as $\mu \downarrow aD_v$,

$$d_v > D_v(1 - 0.687aD_v) \quad (aD_v < \frac{1}{4}, \quad ad_v < \frac{1}{2}(1 + \sqrt{1 - 4aD_v})). \quad (\text{Q.13})$$

7. As to the right-hand side of formula (Q.6) we obtain, squaring it and using (Q.3),

$$\frac{d_{v+1}}{aD_v^2} \leq \frac{4}{(1+\sqrt{1-4\mu})^2}. \quad (\text{Q.14})$$

This formula is more easily used in the form

$$d_{v+1} \leq aD_v^2(1 + \tau_\mu^* \mu) \quad (\text{Q.15})$$

where τ_μ^* is given by

$$\tau_\mu^* = \left(\frac{4}{(1+\sqrt{1-4\mu})^2} - 1 \right) \frac{1}{\mu}. \quad (\text{Q.16})$$

Developing the expression (Q.16), we obtain

$$\tau_\mu^* = \frac{1}{\mu} \frac{1+2\mu-\sqrt{1-4\mu}}{1-2\mu+\sqrt{1-4\mu}} = 2 \frac{2+\mu}{1-2\mu-2\mu^2+\sqrt{1-4\mu}}$$

and we see that τ_μ^* is monotonically increasing with μ . In particular, for $\mu = \frac{1}{4}$ we obtain $\tau_{1/4}^* = 12$ and therefore

$$d_{v+1} < aD_v^2(1 + 12aD_v) \quad (aD_v < \frac{1}{4}, \quad ad_v < \frac{1}{2}(1 + \sqrt{1 - 4aD_v})). \quad (\text{Q.17})$$

8. Observe that from (Q.6) follow the right-hand bounds for d_v and ad_v :

$$d_v \leq 2D_v, \quad ad_v \leq 2aD_v < \frac{1}{2},$$

while for small values of aD_v , the value of d_v/D_v can be assumed to be very near to 1.

This is important since the values of D_v are obtained in the course of the computation, while those of d_v often require a theoretical discussion, and it is therefore desirable to have the required bound of ad_v as large as possible. On the other hand, if we want to obtain the convergence from (Q.3), we must have $ad_v < 1$.

In practice we will have to use a given bound for ad_v and to ask how small D_v must become in order that (Q.14) can be used.

Assuming $ad_v = \alpha < 1$, we then have to obtain the corresponding value of μ from

$$\frac{1}{2}(1 + \sqrt{1 - 4\mu}) > \alpha, \quad \mu < \alpha(1 - \alpha). \quad (\text{Q.18})$$

If for instance $\alpha \leq 0.9$, we obtain $\mu < 0.09$. If $\alpha \leq 0.99$, we obtain $\mu < 0.0099$.

We give finally the values of τ_μ , τ_μ^* , and α_μ corresponding to some standard values of μ :

$\mu:$	0.02	0.002	0.0002
$\tau_\mu:$	0.97	0.98	0.9998
$\tau_\mu^*:$	2.122	2.107	2.0021
$\alpha_\mu:$	0.979	0.997	0.999

(Q.19)

where the values of τ_μ^* and τ_μ are rounded up and those of α_μ are rounded down.

9. We now return to the general case of relation (Q.1), where we will assume that

$$\varphi(u) \downarrow 0 \quad (u \downarrow 0). \quad (\text{Q.20})$$

If we use again the notations (Q.4), it follows from $x_{v+1} - x_v = (\zeta - x_v) - (\zeta - x_{v+1}): d_v + d_{v+1} \geq D_v \geq d_v - d_{v+1} \geq d_v(1 - \varphi(d_v))$, that

$$\frac{1}{1 + \varphi(d_v)} \leq \frac{d_v}{D_v} \leq \frac{1}{1 - \varphi(d_v)}. \quad (\text{Q.21})$$

We must again assume that a bound α of d_v exists such that

$$d_v \leq \alpha, \quad \varphi(\alpha) < 1.$$

Then it follows from (Q.21) that

$$\frac{1}{1 + \varphi(\alpha)} \leq \frac{d_v}{D_v} \leq \frac{1}{1 - \varphi(\alpha)} \quad (\text{Q.22})$$

and further, using in (Q.21) the right-hand inequality (Q.22), that

$$\frac{1}{1 + \varphi\left(\frac{D_v}{1 - \varphi(\alpha)}\right)} \leq \frac{d_v}{D_v} \leq \frac{1}{1 - \varphi\left(\frac{D_v}{1 - \varphi(\alpha)}\right)}, \quad (\text{Q.23})$$

and this procedure can evidently be iterated indefinitely.

10. Consider for instance

$$\varphi(u) := \sqrt{u}, \quad d_v \leq \alpha := \frac{1}{4}. \quad (\text{Q.24})$$

Then, if $D_v \leq 10^{-4}$, we have, as $\varphi(\alpha) = \frac{1}{2}$,

$$\begin{aligned} d_v &\leq 2D_v, & \varphi(d_v) &\leq \frac{\sqrt{2}}{100} < 0.015, \\ d_v &< \frac{D_v}{0.985} < 0.0001016, & \varphi(d_v) &< 0.01008, \\ 0.99D_v &\leq d_v < \frac{D_v}{0.98992} < 1.0102D_v. \end{aligned} \quad (\text{Q.25})$$

11. The application of the feedback techniques is not restricted to the setup given by (Q.1). Consider for instance the sequence $x_v \rightarrow \zeta$ for which, using the notations (Q.4),

$$d_{v+1} \leq d_v \varphi_v \quad (\text{Q.26})$$

where φ_v is a sequence of constants such that

$$\lim_{n \rightarrow \infty} \prod_{v=1}^n \varphi_v = 0. \quad (\text{Q.27})$$

Here it follows again from $x_{v+1} - x_v = (\zeta - x_v) - (\zeta - x_{v+1})$ that

$$d(1 - \varphi_v) \leq D_v \leq d_v(1 + \varphi_v),$$

$$\frac{D}{1 + \varphi_v} \leq d_v \leq \frac{D_v}{1 - \varphi_v}. \quad (\text{Q.28})$$

If, for instance,

$$\varphi_v := 1 - \frac{\alpha}{v}, \quad \alpha \geq 0,$$

it follows from (Q.28) that

$$\frac{D_v}{2 - \alpha/v} \leq d_v \leq \frac{v}{\alpha} D_v,$$

which, if D_v go, for instance, geometrically to 0, still gives a satisfactory estimate.

12. Although the above discussion has been explicitly carried out in the domain of complex numbers, it is important for many applications that with slight modifications the above results also hold in a general metric space (see Chapter 32). We then have to replace definition (Q.4) by

$$D_v := |x_v, x_{v+1}|, \quad d_v := |x_v, \zeta|, \quad (\text{Q.29})$$

while inequalities (Q.2), (Q.8), (Q.21), and (Q.28) follow at once after applying the triangle inequality. All results concerning d_v and D_v then remain correct with the same constants.

In the case of an iteration sequence obtained from a contracting operator in a metric space the corresponding inequality was already deduced as the first inequality of formula (32.4).

A NUMERICAL EXAMPLE

13. We consider the iteration sequence obtained from the *regula falsi* in Section 17 of Chapter 3 in the case of the equation

$$f(x) \equiv x^3 - 2x - 5 = 0.$$

Here, however, we assume as the starting values the values denoted there as $x_2 =: y_0$, $x_3 =: y_1$. Then we have

$$y_4 = 2.09455114, \quad y_5 = 2.094551501.$$

In this case we obtain from (3.14), replacing there x_v by y_v , the value of $q := 0.0801$,

$$|\zeta - y_{v+1}| \leq q |\zeta - y_v|.$$

The *a priori* error of y_5 is then less than $q^5/10 < 3.3 \cdot 10^{-7}$. Since $y_4 - y_5 = 3.6 \cdot 10^{-7}$, we obtain for the *a posteriori* error of y_4 a value $< 4 \cdot 10^{-7}$, and multiplying it with q , for y_5 the estimate $3.3 \cdot 10^{-8}$, while the true error of y_5 is $\sim 2 \cdot 10^{-8}$.

R

Reduced Polynomial Equations

1. Reduced polynomial equations have been defined in Section 2 of Chapter 28 by (28.6).

If we have a general polynomial equation of exact degree n ,

$$F(y) := A_0 y^n + A_1 y^{n-1} + \cdots + A_n = 0, \quad A_0 \neq 0,$$

it can be reduced by the well-known transformation $y = x - A_1/(nA_0)$ to a form where the coefficient A_1 of x^{n-1} vanishes. We can therefore assume that already $A_1 = 0$. To bring then our equation into the reduced form, put, if $F(y) \not\equiv A_0 y^n$,

$$y = \rho x, \quad \rho := \min_{v>0} \left| \frac{A_0}{A_v} \right|^{1/v}.$$

If we then divide the polynomial by A_0 , we obtain finally a polynomial of the form (28.6)—and this is also true if $F(y) \equiv A_0 y^n$.

2. Consider for $n \geq 2$ the sequence of polynomials $\varphi_n(x)$ of exact degree n defined by

$$\varphi_2(x) = x^2 - 2, \quad \varphi_n(x) = x^n - x^{n-2} - \cdots - x - 2 \quad (n > 2). \quad (\text{R.1})$$

By Theorem 12.2 of Chapter 12, $\varphi_n(x)$ has exactly one positive root ρ_n^* , so that $\varphi_n(x) < 0$ if $0 < x < \rho_n^*$ and $\varphi_n(x) > 0$ if $x > \rho_n^*$.

We obtain immediately

$$\varphi_n(2) = 2^n - (2^{n-2} + 2^{n-3} + \cdots + 1) - 1 = 2^n - 2^{n-1} = 2^{n-1} > 0.$$

Therefore

$$\rho_n^* < 2 \quad (n \geq 2). \quad (\text{R.2})$$

On the other hand, the following identity is immediately verified:

$$\varphi_{n+1}(x) \equiv x\varphi_n(x) + x - 2 \quad (n \geq 2).$$

Replacing x here with ρ_n^* , it follows that $\varphi_{n+1}(\rho_n^*) = \rho_n^* - 2$. By (R.2) we see that $\varphi_{n+1}(\rho_n^*) < 0$,

$$\rho_n^* < \rho_{n+1}^* \quad (n \geq 2).$$

It follows therefore that

$$\sqrt{2} = \rho_2^* < \rho_3^* < \cdots < \rho_n^* \uparrow \rho^* \leq 2. \quad (\text{R.3})$$

In order to obtain the value of ρ^* , observe that

$$\begin{aligned} 0 &= \frac{\varphi_n(\rho_n^*)}{\rho_n^{*n}} = 1 - \frac{1}{\rho_n^{*2}} - \frac{1}{\rho_n^{*3}} - \cdots - \frac{1}{\rho_n^{*n-1}} - \frac{2}{\rho_n^{*n}} \\ &= 1 - \frac{1}{\rho_n^{*2}} \frac{1 - \rho_n^{*1-n}}{1 - 1/\rho_n^*} - \frac{1}{\rho_n^{*n}} \\ &= 1 - \frac{1 - \rho_n^{*1-n}}{\rho_n^{*2} - \rho_n^*} - \frac{1}{\rho_n^{*n}}, \end{aligned}$$

and this tends with $n \rightarrow \infty$, by virtue of (R.3), to

$$1 - \frac{1}{\rho^{*2} - \rho^*} = \frac{\rho^{*2} - \rho^* - 1}{\rho^*(\rho^* - 1)}.$$

We have therefore $\rho^{*2} - \rho^* - 1 = 0$ and, since $\rho^* > 0$, it follows that

$$\rho^* = \tau := \frac{\sqrt{5} + 1}{2} = 1.618 \dots. \quad (\text{R.4})$$

3. We are now going to prove that for the reduced polynomial $f(x)$ in (28.6) we have

$$|f(x)| > \left| \frac{x}{\rho_n^*} \right|^n \quad (|x| > \rho_n^*). \quad (\text{R.5})$$

Proof. It follows from (28.6) for $|x| > \rho_n^*$ that

$$\begin{aligned} \frac{f(x)}{x^n} &= 1 - \frac{a_2}{x^2} - \frac{a_3}{x^3} - \cdots - \frac{a_n}{x^n}, \\ \left| \frac{f(x)}{x^n} \right| &\geq 1 - \frac{1}{|x|^2} - \frac{1}{|x|^3} - \cdots - \frac{1}{|x|^{n-1}} - \frac{1}{|x|^n} \\ &> 1 - \frac{1}{\rho_n^{*2}} - \frac{1}{\rho_n^{*3}} - \cdots - \frac{1}{\rho_n^{*n-1}} - \frac{1}{\rho_n^{*n}} \\ &= \left(\frac{\varphi_n(x)}{x^n} + \frac{1}{x^n} \right)_{x=\rho_n^*} = \frac{1}{\rho_n^{*n}}. \end{aligned}$$

This is (R.5).

In particular, it follows that

$$|f(x)| \geq 1 \quad (|x| \geq \rho_n^*). \quad (\text{R.6})$$

This proves (28.7).

4. The reduced polynomial $f(x)$ in (28.6) is majorated by the polynomial

$$\chi_n(x) := x^n + x^{n-2} + \cdots + x + 1. \quad (\text{R.7})$$

We have therefore in particular

$$\frac{f^{(v)}(x)}{v!} \leq \frac{\chi_n^{(v)}(\rho_n^*)}{v!} \quad (|x| \leq \rho_n^*). \quad (\text{R.8})$$

The values

$$M_n^{(v)} := \frac{\chi_n^{(v)}(\rho_n^*)}{v!} \quad (\text{R.9})$$

have been tabulated for $3 \leq n \leq 20$ and $0 \leq v \leq n$.[†]

5. The constant M in (28.8) can obviously be taken as $2M_n^{(2)}$. We therefore give in the accompanying tabulation the values of $2M_n^{(2)}$ for $3 \leq n \leq 20$.

$n:$	3	4	5	6	7
$2M_n^{(2)}:$	9.1282	$3.1442 \cdot 10$	$9.1744 \cdot 10$	$2.3624 \cdot 10^2$	$5.5838 \cdot 10^2$
$n:$	8	9	10	11	12
$2M_n^{(2)}:$	$1.24222 \cdot 10^3$	$2.6430 \cdot 10^3$	$5.4350 \cdot 10^3$	$1.08828 \cdot 10^4$	$2.1334 \cdot 10^4$
$n:$	13	14	15	16	17
$2M_n^{(2)}:$	$4.1086 \cdot 10^4$	$7.8000 \cdot 10^4$	$1.46276 \cdot 10^5$	$2.7148 \cdot 10^5$	$4.9928 \cdot 10^5$
$n:$	18	19	20		
$2M_n^{(2)}:$	$9.1104 \cdot 10^5$	$1.65088 \cdot 10^6$	$2.9732 \cdot 10^6$		

6. A very simple and elegant estimate of $M_n^{(v)}$ can be obtained in terms of binomial coefficients, namely

$$M_n^{(v)} \leq 2 \binom{n}{v} \rho_n^{*n-v} \quad (n \geq 2, \quad 0 \leq v < n), \quad (\text{R.10})$$

[†] This has been done by M. P. Brodmann with the values rounded up at the 4th decimal place. See A. M. Ostrowski, "Some properties of polynomial equations," *SIAM J. Numer. Anal.* 8, 629 (1971).

where the factor 2 cannot be replaced for any couple v, n by a factor <1 . Formula (R.10) follows by virtue of (R.9) from the inequality

$$\binom{n}{v} x^{n-v} \leq \frac{1}{v!} \chi_n^{(v)}(x) \leq 2 \binom{n}{v} x^{n-v} \quad (n \geq 2, \quad 0 \leq v < n, \quad x \geq \rho_n^*). \quad (\text{R.11})$$

To prove (R.11), observe that

$$2x^n = \chi_n(x) + \varphi_n(x) + 1 \quad (\text{R.12})$$

and therefore

$$\frac{1}{v!} \chi_n^{(v)}(x) + \frac{1}{v!} \varphi_n^{(v)}(x) = 2 \binom{n}{v} x^{n-v} \quad (n \geq 2, \quad 0 < v < n). \quad (\text{R.13})$$

Now the left-hand side of inequality (R.11) is immediate for all positive x , since $\binom{n}{v} x^{n-v}$ is just the highest term of $(1/v!) \chi_n^{(v)}(x)$ and the other terms are nonnegative. We therefore need to prove only the right-hand side of inequality (R.11).

For $v = 0$ and $x \geq \rho_n^*$, $\varphi_n(x) + 1$ in (R.12) is ≥ 1 . This proves the right-hand side of inequality (R.11) for $v = 0$.

7. We therefore need to consider only $v > 0$, and it is, by virtue of (R.13), sufficient to prove that

$$\varphi_n^{(v)}(x) \geq 0 \quad (x \geq \rho_n^*). \quad (\text{R.14})$$

But

$$\frac{1}{v!} \varphi_n^{(v)}(x) = \binom{n}{v} x^{n-v} - \binom{n-2}{v} x^{n-2-v} - \binom{n-3}{v} x^{n-3-v} - \dots - \binom{v}{v}. \quad (\text{R.15})$$

For $v = n-1$ this is nx and therefore positive with x . We can therefore assume $0 < v \leq n-2$. Observe that generally $\binom{\mu}{v}$ is strictly monotonically increasing with μ for $\mu \geq v > 0$, since

$$\frac{\binom{\mu+1}{v}}{\binom{\mu}{v}} = \frac{\mu+1}{\mu-v+1}.$$

Therefore the coefficient $\binom{n}{v}$ of the first right-hand term in (R.15) is greater

than all following binomial coefficients. It follows for any positive x that

$$\frac{1}{v!} \varphi_n^{(v)}(x) > \binom{n}{v} (x^{n-v} - x^{n-v-2} - \dots - x - 2) = \binom{n}{v} \varphi_{n-v}(x) \quad (0 < v \leq n-2).$$

But, as $\rho_n^* > \rho_{n-v}^*$, $\varphi_{n-v}(x)$ is positive if $x \geq \rho_n^*$. This proves (R.14).

Relations (R.11) are now proved.

It follows further from (R.11), since $|f^{(v)}(x)|/v! \leq \chi_n^{(v)}(|x|)/v!$, that

$$\frac{1}{v!} |f^{(v)}(x)| \leq 2 \binom{n}{v} |x|^{n-v} \quad (n \geq 2, \quad 0 \leq v < n, \quad |x| \geq \rho_n^*). \quad (\text{R.16})$$

S

Discussion of the q -Acceleration

1. In our discussion we can assume, without loss of generality, that the zero ζ of the polynomial $f(z)$ is $=0$. We consider then $z \rightarrow 0$ and assume that $|z| < \frac{1}{10}$. Γ is an arbitrary but fixed integer > 2 .

We keep the notation of Chapter 30. If $p \geq 1$ is the exact multiplicity of ζ , we can write (28.2) and (28.4), putting $A_0 := f^{(p)}(0)/p!$,

$$f(z) = A_0 z^p (1 + \eta(z)), \quad (\text{S.1})$$

$$\frac{f(z)}{f'(z)} = \frac{z}{p} (1 + \varepsilon_0). \quad (\text{S.2})$$

Here $\eta(z)$ is a polynomial in z and ε_0 a rational function in z , both vanishing at the origin.

In the following discussion, we shall denote by the symbols $\varepsilon_0, \varepsilon_1, \varepsilon_2, \dots$ expressions depending on z and $z^{(\kappa)}$ and such that their modulus is $< C|z|$ where C is a convenient constant, for all $\kappa < k(z) + \Gamma + 2$ and all z with $|z| < \frac{1}{10}$.

2. We put generally $r := |z|$ and

$$m(R) := \min_{|z|=R} |f(z)|, \quad M(R) := \max_{|z|=R} |f(z)|.$$

Choose $R' = |z^{(s)}|$ in such a way that $s \leq k(z)$ and

$$R' := \max_{\kappa \leq k(z)} |z^{(\kappa)}|.$$

It follows by (30.2) that

$$m(R') \leq |f(z^{(s)})| < |f(z)| \leq M(r). \quad (\text{S.3})$$

On the other hand, since the values $M(R)$ and $m(R)$ are each time assumed on the circle $|z|=R$, it follows from (S.1) that

$$m(R') \sim |A_0| R'^p, \quad M(r) \sim |A_0| r^p \quad (r \rightarrow 0),$$

since by the lemma of Section 3, Chapter 30, $R' \rightarrow 0$ with r .

We therefore obtain from (S.3)

$$\overline{\lim}_{z \rightarrow 0} \max_{\kappa \leq k(z)} (|z^{(\kappa)}|/|z|) \leq 1 \quad (\text{S.4})$$

and thence, for sufficiently small $|z|$,

$$|z^{(\kappa)}|/|z| \leq 2, \quad |z^{(\kappa)} - z|/|z| \leq 2.5 \quad (\kappa \leq k(z)). \quad (\text{S.5})$$

Put now, using (30.1) and (S.2),

$$L_\kappa := t^* q^{\kappa-1}/p, \quad z^{(\kappa)} = z - L_\kappa z(1+\varepsilon_0); \quad (\text{S.6})$$

by virtue of (S.5), we can assume, taking $|z|$ sufficiently small,

$$L_\kappa < 3 \quad (\kappa \leq k(z)), \quad L_\kappa < 3q^{\Gamma+1} \quad (\kappa \leq k(z)+\Gamma+1). \quad (\text{S.7})$$

Further, we have

$$|z^{(\kappa+1)} - z^{(\kappa)}| = |(q-1)L_\kappa z(1+\varepsilon_0)| = O(L_\kappa|z|), \quad (\text{S.8})$$

where L_κ is bounded by virtue of (S.7), as long as $\kappa \leq k(z)+\Gamma$.

3. We can assume $|\eta(z^{(\kappa)})| \leq \frac{1}{10}$ ($\kappa \leq k(z)+\Gamma+1$). Put

$$M := \max |\eta'(u)| \quad (|u| \leq \frac{4}{10}(q^{\Gamma+1} + 1)). \quad (\text{S.9})$$

It follows then from (S.8) and (S.9) for $\kappa \leq k(z)+\Gamma$ that

$$\begin{aligned} |\eta(z^{(\kappa+1)}) - \eta(z^{(\kappa)})| &\leq ML_\kappa(q-1)|z||1+\varepsilon_0| = L_\kappa\varepsilon_1, \\ \frac{1+\eta(z^{(\kappa+1)})}{1+\eta(z^{(\kappa)})} &= 1 + \frac{\eta(z^{(\kappa+1)}) - \eta(z^{(\kappa)})}{1+\eta(z^{(\kappa)})} = 1 + \theta^* \frac{10}{9}L_\kappa\varepsilon_1 = 1 + L_\kappa\varepsilon_2, \\ \left[\frac{1+\eta(z^{(\kappa+1)})}{1+\eta(z^{(\kappa)})} \right]^{2/p} &= (1 + L_\kappa\varepsilon_2)^{2/p} = 1 + L_\kappa\varepsilon_3 \quad (\kappa \leq k(z)+\Gamma). \end{aligned} \quad (\text{S.10})$$

Using (S.1), we obtain now

$$\begin{aligned} \left| \frac{f(z^{(\kappa+1)})}{f(z^{(\kappa)})} \right| &= \left| \frac{z^{(\kappa+1)}}{z^{(\kappa)}} \right|^p \left| \frac{1+\eta(z^{(\kappa+1)})}{1+\eta(z^{(\kappa)})} \right| = \left| \frac{z^{(\kappa+1)}}{z^{(\kappa)}} \right|^p |1 + L_\kappa\varepsilon_3|^{p/2}, \\ \left| \frac{f(z^{(\kappa+1)})}{f(z^{(\kappa)})} \right|^{2/p} &= \left| \frac{z^{(\kappa+1)}}{z^{(\kappa)}} \right|^2 (1 + L_\kappa\varepsilon_4) \quad (\kappa \leq k(z)+\Gamma). \end{aligned} \quad (\text{S.11})$$

4. We consider now the first right-hand factor in (S.11). From (S.6) it follows, since L_κ is real, that

$$\left| \frac{z^{(\kappa)}}{z} \right|^2 = 1 - 2L_\kappa \operatorname{Re}(1+\varepsilon_0) + L_\kappa^2 |1+\varepsilon_0|^2 = 1 - 2L_\kappa(1+\varepsilon_5) + L_\kappa^2(1+\varepsilon_6);$$

it follows, since $L_{\kappa+1} = qL_\kappa$, that

$$\begin{aligned} \left| \frac{z^{(\kappa+1)}}{z^{(\kappa)}} \right|^2 &= \frac{1 - 2qL_\kappa(1+\varepsilon_5) + q^2L_\kappa^2(1+\varepsilon_6)}{1 - 2L_\kappa(1+\varepsilon_5) + L_\kappa^2(1+\varepsilon_6)}, \\ \left| \frac{z^{(\kappa+1)}}{z^{(\kappa)}} \right|^2 - 1 &= \frac{(q^2-1)L_\kappa^2(1+\varepsilon_6) - 2(q-1)L_\kappa(1+\varepsilon_5)}{|z^{(\kappa)}/z|^2} \\ &= (q^2-1)L_\kappa(1+\varepsilon_6) \left[L_\kappa - \frac{2}{q+1} \frac{1+\varepsilon_5}{1+\varepsilon_6} \right] / \left| \frac{z^{(\kappa)}}{z} \right|^2, \\ \left| \frac{z^{(\kappa+1)}}{z^{(\kappa)}} \right|^2 &= 1 + (q^2-1)L_\kappa(1+\varepsilon_6) \left[L_\kappa - \frac{2}{q+1} + \varepsilon_7 \right] / \left| \frac{z^{(\kappa)}}{z} \right|^2. \quad (\text{S.12}) \end{aligned}$$

We obtain now from (S.11), (S.7), and (S.12), putting $(1+\varepsilon_6)(1+L_\kappa\varepsilon_4) = 1+\varepsilon_8$,

$$\begin{aligned} \left| \frac{f(z^{(\kappa+1)})}{f(z^{(\kappa)})} \right|^{2/p} &= 1 + L_\kappa\varepsilon_4 + (1+\varepsilon_8)(q^2-1)L_\kappa \left[L_\kappa - \frac{2}{q+1} + \varepsilon_7 \right] / \left| \frac{z^{(\kappa)}}{z} \right|^2 \\ &= 1 + (1+\varepsilon_8)(q^2-1)L_\kappa \left[L_\kappa - \frac{2}{q+1} + \varepsilon_7 + \varepsilon_9 \left| \frac{z^{(\kappa)}}{z} \right|^2 \right] / \left| \frac{z^{(\kappa)}}{z} \right|^2. \end{aligned}$$

Since $|z^{(\kappa)}/z|$ is bounded, we can write $\varepsilon_7 + \varepsilon_9 |z^{(\kappa)}/z|^2 = \varepsilon_{10}$ and obtain finally the formula

$$\left| \frac{f(z^{(\kappa+1)})}{f(z^{(\kappa)})} \right|^{2/p} = 1 + \frac{(1+\varepsilon_8)(q^2-1)}{|z^{(\kappa)}/z|^2} L_\kappa \left[L_\kappa - \frac{2}{q+1} + \varepsilon_{10} \right] \quad (\kappa \leq k(z)+\Gamma). \quad (\text{S.13})$$

5. From (S.13) we easily see now that

$$\frac{2}{q+1} \leq \overline{\lim}_{z \rightarrow 0} L_{k(z)} \leq \frac{2q}{q+1}. \quad (\text{S.14})$$

Indeed, if we had $\overline{\lim}_{z \rightarrow 0} L_{k(z)} > 2q/(q+1)$, then for a convenient sequence $u_v \rightarrow 0$ we would have $L_{k(u_v)} > 2q/(q+1) + \alpha q$ for a convenient $\alpha > 0$. But then, if in (S.13) we put $\kappa = k(u_v) - 1$, it would follow that

$$L_{k(u_v)-1} - \frac{2}{q+1} + \varepsilon_{10} = \frac{1}{q} L_{k(u_v)} - \frac{2}{q+1} + \varepsilon_{10} > \alpha + \varepsilon_{10} > 0$$

and, for sufficiently small $|u_v|$,

$$|f(u_v^{(k)})| > |f(u_v^{(k-1)})|,$$

in contradiction to the definition of $k(z)$.

Similarly, if we had $\lim_{z \rightarrow 0} L_{k(z)} < 2/(q+1)$, it would follow for a convenient sequence $u_v \rightarrow 0$ that $L_{k(u_v)} < 2/(q+1) - \alpha$ with a positive α . But then from (S.13) with $\kappa = k(u_v)$ it would follow for sufficiently small $|u_v|$: $|f(u_v^{(\kappa+1)})| < |f(u_v^{(\kappa)})|$, and this is again contrary to the definition of $k(z)$.

6. From (S.14) it follows further for ε_0 in (S.6), since $1 + \varepsilon_0 \rightarrow 1$, that

$$\begin{aligned} \frac{2}{q+1} &\leq \overline{\lim}_{z \rightarrow 0} (L_k(1 + \varepsilon_0)) \leq \frac{2q}{q+1}, \\ -\frac{q-1}{q+1} &\leq \overline{\lim}_{z \rightarrow 0} [L_k(1 + \varepsilon_0) - 1] \leq \frac{q-1}{q+1}, \\ \overline{\lim}_{z \rightarrow 0} |L_k(1 + \varepsilon_0) - 1| &\leq \frac{q-1}{q+1}. \end{aligned} \quad (\text{S.15})$$

But by (S.6) $L_k(1 + \varepsilon_0) - 1 = -z^{(k)}/z$. Writing $G_q(z)$ for $z^{(k)}$, according to (30.5), we obtain, if we now drop the assumption $\zeta = 0$,

$$\overline{\lim}_{z \rightarrow \zeta} \left| \frac{G_q(z) - \zeta}{z - \zeta} \right| \leq \frac{q-1}{q+1}. \quad (\text{S.16})$$

From (S.1) and (S.16) it follows further that

$$\overline{\lim}_{z \rightarrow \zeta} \left| \frac{f(G_q(z))}{f(z)} \right| \leq \left(\frac{q-1}{q+1} \right)^p, \quad (\text{S.17})$$

while, on the other hand, it follows from (S.13) and (S.14) that

$$\left| \frac{f(z^{(\kappa+1)})}{f(z^{(\kappa)})} \right| > 1 \quad (k(z) < \kappa \leq k(z) + \Gamma). \quad (\text{S.18})$$

7. Consider now the method sketched in Section 14 of Chapter 30, which consists in using simultaneously with a given $q > 1$ also $Q = q^\gamma$, $\gamma > 1$, as in (30.14). We can assume again $\zeta = 0$.

Denote $k(z)$ corresponding to q by k and that corresponding to Q by K . Assume γ as an integer; then we will prove that

$$K = \left[\frac{k}{\gamma} \right] \vee \left[\left(\frac{k}{\gamma} \right) + 1 \right], \quad (\text{S.19})$$

for sufficiently small $|z|$.

Proof. Put, dividing k by γ ,

$$k = K_0 \gamma + \rho, \quad K_0 := \left[\frac{k}{\gamma} \right], \quad 0 \leq \rho < \gamma. \quad (\text{S.20})$$

Then the sequence of quotients

$$\left| \frac{f(z^{(\kappa\gamma)})}{f(z^{(\kappa\gamma-\gamma)})} \right| \quad (\kappa = 1, 2, \dots, K_0)$$

consists of numbers < 1 , since in this sequence $\kappa\gamma \leq K_0\gamma \leq k$ and the sequence (30.2) is strictly decreasing. Therefore $K \geq K_0$.

On the other hand, putting $\kappa := K_0 + 2$ it follows that

$$\left| \frac{f(z^{(K_0\gamma+2\gamma)})}{f(z^{(K_0\gamma+\gamma)})} \right| = \prod_{v=0}^{\gamma-1} \frac{|f(z^{(K_0\gamma+\gamma+v+1)})|}{|f(z^{(K_0\gamma+\gamma+v)})|} \geq 1,$$

by virtue of (S.18) with $\Gamma := 2\gamma - 1$, for sufficiently small $|z|$. Therefore $K \leq K_0 + 1$ and (S.19) is proved.

8. We shall finally discuss the order of magnitude of $k(z) := k$.

It follows from (S.15) that, as $z \rightarrow 0$,

$$L_k = 1 + \theta^* \frac{q-1}{q+1} + \delta$$

where δ , as well as in the following $\delta_0, \delta_1, \delta_2, \dots$, tends to 0 as $z \rightarrow 0$.

Therefore, by (S.6),

$$q^k t^* = pq + \theta^* pq \frac{q-1}{q+1} + \delta_0. \quad (\text{S.21})$$

On the other hand, it follows from (S.1) and (S.2) that for $T(z)$ as defined in (28.10)

$$T(z) \sim \frac{p^2 |A_0| |z|^{p-2}}{M}. \quad (\text{S.22})$$

9. If $p = 1$, we see that $T(z) \rightarrow \infty$ and t^* , as defined by (28.10), $= 1$, if $|z|$ is sufficiently small. It then follows from (S.21) that

$$k = \frac{1}{\lg q} \lg \left(q + \theta^* q \frac{q-1}{q+1} \right) + \delta_1.$$

If $p = 2$, we have

$$T(z) \rightarrow \frac{2 |f''(0)|}{M}, \quad t^* \rightarrow \text{Min} \left(1, \frac{2 |f''(0)|}{M} \right) =: t_0;$$

then

$$k = \frac{1}{\lg q} \lg 2q \frac{\left(1 + \theta^* \frac{q-1}{q+1} \right)}{t_0} + \delta_2.$$

In both cases it is probably not worthwhile using the q -acceleration at all, at least as long as $t^* = 1$.

10. Consider now the case where $p > 2$. In this case it follows from (S.22) that $T(z) \rightarrow 0$, and therefore

$$t^* = \frac{p^2 |A_0| |z|^{p-2} (1 + \delta_3)}{M}.$$

Putting $c := qM/pA_0$ it follows that

$$q^k = c \frac{1 + \theta^* \frac{q-1}{q+1}}{|z|^{p-2}} (1 + \delta_4), \quad k = \frac{p-2}{\lg q} \lg \frac{1}{|z|} + \theta^* C$$

where C is a constant depending on $f(z)$.

T

Remainder in the Taylor Formula for Analytic Functions

1. We give in what follows a formula for the remainder of the Taylor development in the case of analytic functions, which is very useful in applications but is usually not given in the standard texts. As a matter of fact, in special cases it was used in Chapters 7 and 14. However, in both cases it was derived directly without reference to the general formula.

Theorem. *Assume $f(z)$ regular in the circle $|z-a| \leq r$, $r > 0$. Then we have, if $|z-a|=:\rho \leq r$,*

$$\begin{aligned} R_m &:= f(z) - \sum_{v=0}^m \frac{f^{(v)}(a)}{v!} (z-a)^v \\ &= \theta^* \frac{(z-a)^{m+1}}{(m+1)!} f^{(m+1)}(a+\eta\rho), \quad |\eta| = 1, \quad |\theta^*| \leq 1, \end{aligned} \quad (\text{T.1})$$

$$|R_m| \leq \frac{\rho^{m+1}}{(m+1)!} \max_{|u-a|=\rho} |f^{(m+1)}(u)|. \quad (\text{T.2})$$

2. Proof. The proof follows easily from the formula

$$R_m = \int_a^z \frac{(z-u)^m}{m!} f^{(m+1)}(u) du, \quad (\text{T.3})$$

where we integrate along the rectilinear segment from a to z and which is familiar in the real case.

Since $f^{(m+1)}(z)$ is not changed if we subtract from $f(z)$ the sum of the first $m+1$ terms of its Taylor development at $z=a$, we can assume, proving (T.3), without loss of generality that

$$f(a) = f'(a) = \cdots = f^{(m)}(a) = 0. \quad (\text{T.4})$$

Equation (T.3) reduces in this case to

$$f(z) = \int_a^z \frac{(z-u)^m}{m!} f^{(m+1)}(u) du. \quad (\text{T.5})$$

But for $m = 0$, (T.5) follows at once from $f(a) = 0$. We can therefore prove our assertion by induction. For $m > 0$, integrating the right-hand integral by parts, we obtain

$$\begin{aligned} \int_a^z \frac{(z-u)^m}{m!} f^{(m+1)}(u) du &= \frac{(z-u)^m}{m!} f^{(m)}(u) \Big|_{u=a}^{u=z} + \int_a^z \frac{(z-u)^{m-1}}{(m-1)!} f^{(m)}(u) du \\ &= \int_a^z \frac{(z-u)^{m-1}}{(m-1)!} f^{(m)}(u) du, \end{aligned}$$

using $m > 0$ and $f^{(m)}(a) = 0$. Here, the last integral is $f(z)$ if we assume our assertion true for $m-1$. Equation (T.3) is proved.

3. Put now $z-a = \rho\varepsilon$, $|\varepsilon|=1$, and introduce in the integral in (T.3) as a new variable of integration the real variable t given by

$$u-a = \varepsilon t, \quad 0 \leq t \leq \rho.$$

Then it follows from (T.3) that

$$R_m = \varepsilon^{m+1} \int_0^\rho \frac{(\rho-t)^m}{m!} f^{(m+1)}(a+\varepsilon t) dt,$$

$$|R_m| \leq \operatorname{Max}_{|u-a| \leq \rho} |f^{(m+1)}(u)| \int_0^\rho \frac{(\rho-t)^m}{m!} dt,$$

since $\rho-t \geq 0$ along the interval of integration. Carrying out the integration, (T.2) follows immediately.

Observe now that $\operatorname{Max}_{|u-a| \leq \rho} |f^{(m+1)}(u)|$ is assumed, by the maximum property of the modulus of an analytic function, in a certain point on the circle $|u-a| = \rho$, $u = a + \eta\rho$. Then (T.1) follows immediately.

U

Equality Conditions for the Newton–Raphson Iteration

1. In the following lemma we will be concerned with the equality sign in formulas (39.2)–(39.7) and with (38.30). We will denote the equalities corresponding to these relations by (39.2), ..., (39.7).

A LEMMA

Lemma 1. *Assume the hypotheses of Lemma 39.1. Then from any of the equalities (39.2)–(39.7) follow (38.30) and (38.32).*

From (39.5) follow the equalities (39.2), (39.3), (39.4), and

$$\|h_1\| = \|Q_1\| \|f_1\|, \quad (\text{U.1})$$

and vice versa: from (39.2) \wedge (39.3) \wedge (U.1) follows (39.5).

From (39.2) follows

$$\|P(\xi_0 + t_2 h_0) - P(\xi_0 + t_1 h_0)\| = (t_2 - t_1) \|h_0\|/\sigma_0 \quad (0 \leq t_1 \leq t_2 \leq 1). \quad (\text{U.2})$$

From (39.4) follows (39.3).

From (39.7) follow (39.2) \wedge (39.3) \wedge (39.4).

From (39.6) follows $\varphi = 0$, $\alpha_2 = 2$, and (39.5).

2. Proof. Referring to Section 2 of Chapter 39, we see that if we have (39.2), we must obviously have $\|\omega\| = 1/\alpha_0$ and it follows from (39.10) that

$$\|P_1 - P_0\| = \|h_0\|/\sigma_0, \quad \|h_0\| = \|Q_0\| \|f_0\|.$$

The second relation is (38.30). The first relation can be written as

$$\|P(\xi_0 + h_0) - P(\xi_0)\| = \|h_0\|/\sigma_0,$$

and since we have by (38.18)

$$\|P(\xi_0 + t_2 h_0) - P(\xi_0 + t_1 h_0)\| \leq (t_2 - t_1) \|h_0\|/\sigma_0 \quad (0 \leq t_1 \leq t_2 \leq 1),$$

(U.2) follows by the triangle inequality, while (38.32) follows from (U.2).

3. If we have $(\overline{39.3})$, then (38.30) and (38.32) follow by Lemma 38.4.

Referring to the deduction of $(\overline{39.4})$ from (39.3) in the proof of Lemma 39.1, we see that $(\overline{39.3})$ follows from $(\overline{39.4})$.

Referring to the derivation of $(\overline{39.5})$ in the proof of Lemma 39.1, we see that $(\overline{39.5})$ holds if and only if we have (U.1) as well as $(\overline{39.3})$ and $(\overline{39.2})$. On the other hand, it follows from (U.1), (39.3) , and (39.2) that

$$\|h_1\| = \|f_1\| \|Q_1\| = \frac{\|f_0\| \alpha_0 \|Q_0\|}{2\alpha_0 - 1} = \frac{\|h_0\|}{2\alpha_0 - 2}$$

and we see that $(\overline{39.4})$ follows from $(\overline{39.5})$.

4. If $(\overline{39.6})$ holds, it follows from $e^\varphi + e^{-\varphi} \geq 2$ that we have $\varphi = 0$, $\alpha_0 = 2$, and $(\overline{39.5})$.

Finally, $(\overline{39.7})$ is only possible if we have $(\overline{39.2})$ and $(\overline{39.3})$. But then our assertions follow from those about $(\overline{39.3})$. Lemma 1 is proved.

EQUALITY CONDITIONS FOR NORMED SPACES

5. We are now going to discuss the conditions for the equality sign in relations (39.2) , (39.3) , (39.5) , (39.7) , and (38.2) written for the general index v . We rewrite these equations, expressing the upper bounds if possible in terms of the α_v . These are the relations where v runs from 0 to ∞ :

$$\left. \begin{array}{l} A_v \quad \frac{\|f_{v+1}\|}{\|f_v\|} \leq \frac{1}{2\alpha_v}, \\ B_v \quad \frac{\|Q_{v+1}\|}{\|Q_v\|} \leq \frac{\alpha_v}{\alpha_v - 1}, \\ C_v \quad \frac{\|h_{v+1}\|}{\|h_v\|} \leq \frac{1}{2\alpha_v - 2}, \end{array} \right\} \quad (U.3)$$

$$\left. \begin{array}{l} D_v \quad \|h_v\| \leq \|Q_v\| \|f_v\|, \\ E_v \quad \sigma_{v+1} \leq \sigma_v. \end{array} \right\} \quad (U.4)$$

Formulas (U.3) with the equality sign will be denoted resp. by \bar{A}_v , \bar{B}_v , \bar{C}_v , \bar{D}_v , \bar{E}_v . Observe that if $\sigma_1 < \sigma_0$, then α_1 can be replaced with a greater number and all α_v ($v > 0$) become greater. But this signifies that we have in the original relations (U.3) the strict inequality. Again, if we have for a $\mu > 0$: $\sigma_\mu < \sigma_{\mu-1}$, α_μ could be increased, if we were to start from ξ_μ anew, and then all relations (U.3) with $v \geq \mu$ must be strict inequalities.

Therefore, if one of the relations

$$\bar{A}_v, \quad \bar{B}_v, \quad \bar{C}_v \quad (\text{U.5})$$

holds for a $v = \mu \geq 0$, we must have

$$\sigma_0 = \sigma_1 = \cdots = \sigma_\mu. \quad (\text{U.6})$$

6. We are now going to prove

Lemma 2. *If for a $v = \mu > 0$ one of the relations (U.5) holds, then all relations (U.3) and (U.4) hold with the equality sign for $v < \mu$.*

Proof. Under the assumption of the lemma we have (U.6). On the other hand, we see by Lemma 1 that \bar{D}_0 , that is, (38.30), follows from one of the relations $\bar{A}_0, \bar{B}_0, \bar{C}_0$. Therefore, in our case \bar{D}_μ follows.

Further, by Lemma 1, relations (39.2) and (39.4), that is, \bar{B}_0 and \bar{A}_0 , follow from $\sigma_1 = \sigma_0$. Therefore all relations \bar{A}_v and \bar{B}_v ($v < \mu$) follow from (U.6), and from there on also \bar{D}_v ($v < \mu$).

As to the relations C_v , it follows from Lemma 1 that \bar{C}_0 , that is, (39.5), follows from \bar{D}_1, \bar{A}_0 , and \bar{B}_0 . Therefore \bar{C}_v ($v < \mu$) follows from D_{v+1}, A_v , and B_v . Our lemma is proved.

7. Assume now that we have the equality sign in (38.19) for a $v = n \geq 0$. The formula (38.19) was obtained from (39.20) and (39.21),

$$\|\xi_{n+1} - \zeta\| \leq \rho_n := \exp(-2^n \varphi) \|h_n\| \leq \|h_0\| \exp(-2^n \varphi) \sin \varphi / \sin 2^n \varphi.$$

From the equality sign in (38.19) it follows therefore that $\|\xi_{n+1} - \zeta\| = \rho_n$. But on the other hand, we have $\zeta - \xi_{n+1} = \sum_{v>n} h_v$. We obtain therefore

$$\left\| \sum_{v>n} h_v \right\| = \rho_n. \quad (\text{U.7})$$

Replacing the subscript 0 in (39.3) with v , we obtain $\|h_v\| \leq \rho_{v-1} - \rho_v$ and therefore $\sum_{v>n} \|h_v\| \leq \rho_n$. Relation (U.7) becomes

$$\left\| \sum_{v>n} h_v \right\| = \sum_{v>n} \|h_v\|.$$

Now it follows by the triangle inequality that

$$\left\| \sum_{v>n} h_v \right\| \leq \|h_{n+1}\| + \left\| \sum_{v>n+1} h_v \right\| \leq \sum_{v>n} \|h_v\|, \quad \left\| \sum_{v>n+1} h_v \right\| = \sum_{v>n+1} \|h_v\|.$$

We now see that if we have $\|\xi_{n+1} - \zeta\| = \rho_n$, we also have $\|\xi_{v+1} - \zeta\| = \rho_v$ for all $v \geq n$.

8. On the other hand, it follows that the above equality is only possible if we have $\rho_{v-1} - \rho_v = \|h_v\|$ for all $v > n$. Introducing the values of ρ_{v-1} and

ρ_v from (39.19), we obtain

$$\begin{aligned}\|h_v\| &= \exp(-2^{v-1}\varphi)\|h_{v-1}\| - \exp(-2^v\varphi)\|h_v\|, \\ \|h_v\|(1 + \exp(-2^v\varphi)) &= \|h_{v-1}\|\exp(-2^{v-1}\varphi), \\ 2(\alpha_v - 1)\|h_v\| &\equiv 2\|h_v\|\cos 2^{v-1}\varphi = \|h_{v-1}\|.\end{aligned}$$

But this is \bar{C}_{v-1} for all $v > n$. By what has been proved in Section 6, it follows now:

Theorem 1. *If under the conditions of one of Theorems 38.1–38.3 the equality sign holds in (38.19) for at least one value n of v , it holds also for any $v \geq n$. We have then*

$$\|\xi_v - \zeta\| = \rho_{v-1}, \quad \rho_{v-1} - \rho_v = \|h_v\| \quad (v > n) \quad (\text{U.8})$$

and the equality sign holds in all relations A_v, B_v, C_v, D_v, E_v of Section 5.

EQUALITY CONDITIONS FOR STRICTLY NORMED SPACES

9. In order to obtain further information in the case of the equality sign in (38.19), we must introduce restrictive conditions concerning the norms in the spaces X and Y .

A normed linear space X is called *strictly normed* if for any $\xi_1, \xi_2 \in X$ the relation

$$\|\xi_1 + \xi_2\| = \|\xi_1\| + \|\xi_2\| \quad (\text{U.9})$$

implies the existence of two nonnegative numbers t_1, t_2 with

$$\xi_1 = t_1(\xi_1 + \xi_2), \quad \xi_2 = t_2(\xi_1 + \xi_2), \quad t_1 + t_2 = 1. \quad (\text{U.10})$$

For instance, any Hilbert space is strictly normed.

Geometrically the above definition can be interpreted as the postulate that if the triangle with the vertices $0, \xi_1, \xi_1 + \xi_2$ lies in a strictly normed linear space X , the length of the side $\langle 0, \xi_1 + \xi_2 \rangle$ is less than the sum of the lengths of the two other sides $\langle 0, \xi_1 \rangle$ and $\langle \xi_1, \xi_1 + \xi_2 \rangle$ of this triangle, unless ξ_1 lies on the side $\langle 0, \xi_1 + \xi_2 \rangle$.

It is easy to see, as in elementary geometry, that if two points A, B of a strictly normed space X are connected by a polygon $AA_1A_2 \cdots A_nB$ in X , then the length of this polygon is greater than the norm of $B - A$, unless all A_v lie, in the order of v , on the interval $\langle A, B \rangle$.

10. Lemma 3. *Assume X a strictly normed linear space and $A(t)$ a function, with values in X , of the real variable t , $a \leq t \leq b$. Then, if $A(t)$ satisfies*

the functional equation

$$\|A(u) - A(v)\| = q|u - v| \quad (u, v \in \langle a, b \rangle) \quad (\text{U.11})$$

with the positive constant q , we have

$$A(t) = A(a) + (t-a)(A(b)-A(a))/(b-a). \quad (\text{U.12})$$

Proof. Indeed, from

$$\begin{aligned} \|A(b) - A(t)\| + \|A(t) - A(a)\| &= q(b-a) = q(b-t) + q(t-a) \\ &= \|A(b) - A(t)\| + \|A(t) - A(a)\| \end{aligned}$$

it follows, since X is strictly normed, that

$$A(t) - A(a) = t^*(A(b) - A(a)), \quad t^* \geq 0.$$

Taking the norms on both sides and using (U.11), we obtain

$$t^* = (t-a)/(b-a)$$

and the lemma is proved.

11. We assume now that we have in (38.19) the equality sign for $v \geq n$ and that the space X is *strictly normed*. Put

$$R_v := \xi_v - \zeta \quad (v \geq n).$$

It follows from Theorem 1 that we have then

$$\|R_v\| = \rho_v, \quad \rho_{v-1} - \rho_v = \|h_v\| \quad (v \geq n),$$

where, since $\|h_{v+1}\|/\|h_v\| = 1/(2\alpha_v - 2)$, all h_v , and therefore also all R_v , are $\neq 0$. From

$$R_v = R_{v+1} + h_v, \quad \|R_v\| = \|R_{v+1}\| + \|h_v\| \quad (v \geq n)$$

it follows, applying (U.10) to $v = n$ and $v = n+1$, since the corresponding t_1, t_2 are in this case positive,

$$R_{n+1} = ah_n, \quad h_{n+1} = bR_{n+1}, \quad a \wedge b > 0$$

and therefore $h_{n+1} = ch_n$, $c > 0$. But then $c = 1/(2\alpha_n - 2)$ and we obtain for all $v \geq n$

$$h_{v+1} = h_v/(2\alpha_v - 2) \quad (v \geq n). \quad (\text{U.13})$$

Multiplying, it follows for positive τ_v that

$$h_v = \tau_v h_n, \quad \tau_{v+1}/\tau_v = 1/(2\alpha_v - 2) \quad (v \geq n). \quad (\text{U.14})$$

Putting $T_v := \sum_{\kappa=n}^v \tau_\kappa$ ($v \geq n$), we have

$$\xi_{v+1} = \xi_n + T_v h_n \quad (v \geq n). \quad (\text{U.15})$$

Obviously

$$T_v \uparrow T^* = \|\zeta - \xi_n\| / \|h_n\|, \quad \zeta = \xi_n + T^* h_n. \quad (\text{U.16})$$

12. Put now

$$H(T) := f(\xi_n + Th_n) \quad (0 \leq T \leq T^*). \quad (\text{U.17})$$

Obviously, $H(T_v) = f_{v+1}$. If we now put, for $0 \leq t \leq 1$,

$$H_v(t) := f(\xi_v + th_v), \quad H'_v = P(\xi_v + th_v)h_v,$$

it follows from (U.15) and (U.16) that

$$H_v(t) = H(T), \quad T = T_{v-1} + t\tau_v \quad (v \geq n). \quad (\text{U.18})$$

By Lemma 38.4 the equality sign in (39.5), that is, \bar{C}_0 , entails relation (38.32). It therefore follows from \bar{C}_v , correspondingly, that

$$\|H'_v(t'') - H'_v(t')\| = |t'' - t'| \|h_v\|^2 / \sigma \quad (0 \leq t'' \leq t' \leq 1). \quad (\text{U.19})$$

Differentiating (U.18), we obtain $H'_v(t) = \tau_v H'(T)$. Therefore

$$\tau_v \|H'(T'') - H'(T')\| = |t'' - t'| \tau_v^2 \|h_v\|^2 / \sigma,$$

if $T'' = T_{v-1} + t''\tau_v$, $T' = T_{v-1} + t'\tau_v$, $T_{v-1} \leq T'' \wedge T' \leq T_v$. But then it follows that

$$\|H'(T'') - H'(T')\| = |T'' - T'| \|h_n\|^2 / \sigma \quad (T_{v-1} \leq T'' \wedge T' \leq T_v). \quad (\text{U.20})$$

13. We assume now that Y is *strictly normed*, too; then we can apply Lemma 3 of Section 10 to (U.20). It follows that we can write, for $T_{v-1} \leq T \leq T_v$,

$$H'(T) = 2A_v T + B_v, \quad A_v \wedge B_v \in Y.$$

For $T = T_{v-1}$ and $T = T_v$ we obtain, using (U.14),

$$2A_v T_{v-1} + B_v = H'(T_{v-1}) = P(\xi_v)h_n = P_v h_v / \tau_v = -f_v / \tau_v,$$

$$2A_v T_v + B_v = P_{v+1} h_n = P_{v+1} h_{v+1} / \tau_{v+1} = -f_{v+1} / \tau_{v+1} = (2 - 2\alpha_v) f_{v+1} / \tau_v.$$

Subtracting, we obtain

$$2\tau_v^2 A_v = f_v - (2\alpha_v - 2) f_{v+1}$$

and further

$$\tau_v^2 B_v = T_{v-1} (2\alpha_v - 2) f_{v+1} - T_v f_v.$$

14. For $H(T)$ we obtain now

$$H(T) = A_v T^2 + B_v T + C_v, \quad C_v \in Y \quad (T_{v-1} \leq T \leq T_v).$$

Since $H(T_{v-1}) = f_v$, $H(T_v) = f_{v+1}$, it follows by subtraction that

$$f_{v+1} - f_v = B_v(T_v - T_{v-1}) + A_v(T_v^2 - T_{v-1}^2) = \tau_v(B_v + A_v(T_v + T_{v-1})),$$

$$\tau_v(f_{v+1} - f_v) = T_{v-1}(2\alpha_v - 2)f_{v+1} - T_v f_v + (T_v + T_{v-1})\left(\frac{f_v}{2} - (\alpha_v - 1)f_{v+1}\right);$$

hence, bringing all terms with f_v to the left and all terms with f_{v+1} to the right, we have

$$\begin{aligned} f_v[-\tau_v + T_v - \frac{1}{2}(T_{v-1} + T_v)] &= f_{v+1}[(2\alpha_v - 2)T_{v-1} - \tau_v - (\alpha_v - 1)(T_{v-1} + T_v)], \\ -\frac{\tau_v}{2}f_v &= f_{v+1}((\alpha_v - 1)T_{v-1} - \tau_v - (\alpha_v - 1)T_v) \\ &= -\tau_v \alpha_v f_{v+1}, \\ f_{v+1} &= \frac{f_v}{2\alpha_v} \quad (v \geq n). \end{aligned} \tag{U.21}$$

Introducing this into the above expressions for $\tau_v^2 A_v$, $\tau_v^2 B_v$, we obtain

$$2\tau_v^2 A_v = f_v - \frac{2\alpha_v - 2}{2\alpha_v} f_v = \frac{f_v}{\alpha_v}, \tag{U.22}$$

$$2\alpha_v \tau_v^2 A_v = f_v, \tag{U.22}$$

$$\tau_v^2 B_v = f_v \left(\frac{2\alpha_v - 2}{2\alpha_v} T_{v-1} - T_v \right) = -\frac{\tau_v + T_{v-1}}{\alpha_v} f_v, \tag{U.23}$$

$$B_v = -(T_{v-1} + \alpha_v \tau_v) 2A_v. \tag{U.24}$$

15. We are now going to compute A_v , B_v , and C_v and prove that they are independent of v .

From (U.14) it follows, squaring and using (38.7),

$$\tau_{v+1}^2 = \frac{\tau_v^2}{4(\alpha_v - 1)^2} = \frac{\tau_v^2}{2\alpha_{v+1}}$$

and using (U.21), since $\tau_n = 1$,

$$\begin{aligned} \frac{f_{v+1}}{\alpha_{v+1} \tau_{v+1}^2} &= \frac{2f_v}{2\alpha_v \tau_v^2} = \frac{f_v}{\alpha_v \tau_v^2} = \frac{f_n}{\alpha_n}, \\ A_v &=: A = f_n/(2\alpha_n). \end{aligned} \tag{U.25}$$

On the other hand, $T_{v-1} + \alpha_v \tau_v$ is, by (38.7) and (U.14), independent of v . Indeed,

$$(T_v + \alpha_{v+1} \tau_{v+1}) - (T_{v-1} + \alpha_v \tau_v) = \tau_v - \alpha_v \tau_v + 2(\alpha_v - 1)^2 \frac{\tau_v}{2\alpha_v - 2} = 0,$$

and therefore, for $v = n + 1$, as $T_n = \tau_n = 1$,

$$\begin{aligned} T_{v-1} + \alpha_v \tau_v &= \tau_n + \alpha_{n+1} \tau_{n+1} = 1 + 2(\alpha_n - 1)^2 \frac{1}{2\alpha_n - 2} = \alpha_n, \\ B_v &= -2\alpha_n A = -f_n. \end{aligned} \quad (\text{U.26})$$

As to C_v , observe that we must have for $v > n$

$$A_{v-1} T_{v-1}^2 + B_{v-1} T_{v-1} + C_{v-1} = A_v T_{v-1}^2 + B_v T_{v-1} + C_v,$$

and since $A_{v-1} = A_v$, $B_{v-1} = B_v$, we must also have $C_{v-1} = C_v$ ($v > n$), and therefore $C_v = C_n = H(0) = f_n$. We see that

$$H(T) = f(\xi_n + Th_n) = f_n \left(\frac{T^2}{2\alpha_n} - T + 1 \right) \quad (0 \leq T \leq T^*). \quad (\text{U.27})$$

16. As the roots of the quadratic equation $T^2/(2\alpha_n) - T + 1 = 0$ are

$$1 + \exp(-2^n \varphi), \quad 1 + \exp(2^n \varphi),$$

it follows from the unicity theorem, Theorem 40.2, that the smaller one of them corresponds to ζ and that therefore

$$T^* = 1 + \exp(-2^n \varphi), \quad \zeta = \xi_n + (1 + \exp(-2^n \varphi)) h_n. \quad (\text{U.28})$$

On the other hand, if we have (U.27), the equality sign in (38.19) holds indeed for $v \geq n$. This follows immediately from the discussion of the polynomial F_{α_0} in (40.1) if we replace there α_0 with α_n and φ with $2^n \varphi$.

If we return now to the case of Theorem 40.1 where (38.19) holds for $v > n$, we see that in this case we obtain, as $\alpha_0 = 2$,

$$f(\xi_0 + Th_0) = f_0(T/2 - 1)^2. \quad (\text{U.29})$$

We have proved now:

Theorem 2. *If both X and Y are strictly normed, under the conditions of Theorem 1 we have relations (U.13), (U.21), and (U.27), and under the conditions of Theorem 40.1, formula (U.29).*

Bibliographical Notes

TEXTBOOKS AND PAPERS REPEATEDLY QUOTED

1. A. S. Householder, *Principles of Numerical Analysis*, McGraw-Hill, New York, 1953.
2. A. Ostrowski, *Vorlesungen über Differential- und Integralrechnung*, Vol. II, Birkhäuser, Basel, 1951.
3. A. Ostrowski, "Sur la convergence et l'estimation des erreurs dans quelques procédés de résolution des équations numériques," *Collection of Papers in Memory of D. A. Grave*, pp. 213–234, Moscow, 1940. An English translation appeared as *Tech. Rept. No. 7*, Aug. 30, 1960, of the Appl. Math. and Stat. Labs., Stanford Univ., Stanford, California.
4. A. Ostrowski, "Recherches sur la méthode de Gräffe et les zéros des polynômes et des séries de Laurent," *Acta Math.* **72**, 99–257, 1940.
5. E. Schröder, "Über unendlich viele Algorithmen zur Auflösung der Gleichungen," *Math. Ann.* **2**, 317–365, 1870.
6. J. F. Steffensen, *Interpolation*, Chelsea, Bronx, New York, 1950.
7. F. A. Willers, *Methoden der praktischen Analysis*, 2nd ed., de Gruyter, Berlin, 1954. Translation of the 1st edition: *Practical Analysis. Graphical and Numerical Methods*, Dover, New York, 1948.
8. A. Ostrowski, *Solution of Equations and Systems of Equations*, translated into Russian by L. S. Rumshiski and B. L. Rumshiski, I. I. L., Moscow, 1963.
9. J. F. Traub, *Iterative Methods for Solution of Equations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

Chapters 1A and 1B

Divided differences were introduced by Newton and later repeatedly investigated by Ampère, Cauchy, Stieltjes, and many other authors. The most detailed exposition of their properties can be found in Milne-Thomson, *The Calculus of Finite Differences*, pp. 1–19, London, 1933. In the last decades much work has been done in Rumania on divided differences. Compare the monograph of E. Popoviciu, *Teoreme de mediedin analiza matematică și legătura lor cu teoria interpolarii*, Dacia, Cluj, Rumania, 1972. The exposition in our text differs from the standard ones by the adopted notation which brings out more explicitly the operator character of the process implied, and by the more detailed treatment of the confluent case.

Chapter 2

See the remarks about inverse interpolation in Steffensen [6,* p. 80]. Darboux's theorem was first given in a fundamental paper by Darboux, "Mémoire sur les fonctions discontinues,"

* Numbers in square brackets refer to the bibliography on this page.

Ann. École Norm. Sup. Paris **4**, 1875; formulas (2.5) and (2.7) in Schröder [5, p. 330]. Theorem 2.2 is apparently due to Ostrowski, 1st ed. Theorems 2.3 and 2.3° have been added in the second edition. For references about the Schröder series see the note to Chapter 14.

Chapter 3

Regula falsi goes back to the early Italian algebraists. The material of Sections 4, 5, and 8–16 is due to Ostrowski, 1st ed.

Chapter 4

The distinction between points of attraction and points of repulsion has been introduced by the late J. F. Ritt. Theorem 4.2 in the case of points of attraction goes back to Schröder [5, p. 323]. The content of Sections 6–8 is due to Ostrowski, 1st ed. Theorems 4.3–4.5 are essentially contained in general theorems of functional analysis about existence of fixed points of transformations. Compare J. Weissinger, *Math. Nachr.* **8**, 193–212, 1952, and J. Schröder, *Arch. Math.* **7**, 471–484, 1956.

Chapter 5

Theorem 5.1 is an improved version of a theorem by R. von Mises and H. Geiringer, “Praktische Verfahren der Gleichungsauflösung, zusammenfassender Bericht,” *Z. Angew. Math. Mech.* **9**, 58–77, 1929.

Some essential points of the analysis of von Mises and Geiringer go back to M. Bauer and L. Féjer. Compare M. Bauer, “Zur Bestimmung der reellen Wurzeln einer algebraischen Gleichung durch Iteration,” *Jber. Deut. Math.-Verein.* **25**, 294–301, 1916. Theorem 5.2 was first published in Ostrowski [3] and later generalized to a larger class of comparison functions by J. Karamata, “Über das asymptotische Verhalten der Folgen, die durch Iteration definiert sind,” *Rec. Trav. Acad. Serbe Sci.* **35**, 60, 1953. The content of Section 10 is due to Ostrowski, 1st ed.

Chapter 6

The quadratic character of convergence has been discussed by J. B. J. Fourier. Compare *Oeuvres de Fourier*, Vol. II, pp. 249–250, Gauthier-Villars, Paris, 1890.

Chapter 7

For the distinction between *a priori* and *a posteriori* estimates, compare A. Ostrowski, “Sur la continuité relative des racines d'équations,” *Compt. Rend.* **209**, 777–779, 1939. The existence theorems of this chapter were published for the first time in this form by Ostrowski [3]. They were already given in a less precise form by A. L. Cauchy in his *Leçons sur le calcul différentiel*, Buré frères, Paris, 1823, reprinted in *Oeuvres complètes*, 2nd series, Vol. IV, Gauthier-Villars, Paris, 1899. Compare in particular in this reprint pp. 576, 578, and 600.

If we put in the real and the complex case, respectively,

$$m := \min_{J_0} |f'(z)|, \quad m := \min_{K_0} |f'(z)|,$$

Cauchy's conditions are correspondingly

$$m > 2|h_0|M, \quad m > 4|h_0|M.$$

The next results relevant in this connection were obtained by H. B. Fine, *Proc. Nat. Acad. Sci. USA*, **2**, 546–552, 1916. Fine's conditions are in our notations, respectively,

$$M|f(z_0)| < m^2, \quad M|f'(z_0)| < m^2/\sqrt{2}.$$

While Cauchy's conditions are clearly weaker than ours, this is not immediately seen in the case of Fine's conditions, since, though they contain everywhere m instead of $|f'(z_0)|$, Fine's constants are smaller. In order to compare these conditions put

$$q := |f'(z_0)|^2 / |Mf(z_0)|.$$

Then our condition is in both cases $q \geq 2$. From the conditions of Cauchy can be deduced $q > 3$, respectively, $q > 5$. The corresponding conditions that can be deduced from those of Fine are

$$q > \frac{1}{2}(3 + \sqrt{5}) = 2.618, \quad q > 1 + \sqrt{2}/2 + \sqrt{\frac{1}{2} + \sqrt{2}} = 3.090.$$

However, the importance of Fine's paper consists of its treatment of *systems* of equations. Compare the note to Chapter 42.

Chapter 8

Formula (8.1) is due to E. Schröder [5, p. 325]. The asymptotic relation (8.13) was given by Ostrowski, 2nd ed. Further unpublished material is added in this edition. Some special material with interesting numerical examples on the application of Schröder's formula in computational practice is to be found in L. B. Rall, "Convergence of the Newton process to multiple solutions," *Num. Math.* **9**, 23–37, 1966.

Chapter 9

The bounds (9.1) were given by J. B. J. Fourier in 1818. Compare *Oeuvres de Fourier*, Vol. II, p. 248, Gauthier-Villars, Paris, 1890. The discussion given and particularly the formulas (9.3), (9.22), and (9.23) are due to Ostrowski, 1st ed.

Chapter 10

The bounds (10.1) were proposed by G. P. Dandelin (1824), *N. Mém. Acad. Bruxelles* (1826). Theorem 10.1 is due to Ostrowski, 1st ed.

Chapter 11

The content of this chapter is due to Ostrowski, 1st ed.

Chapter 12

For the general theory of the difference equations with constant coefficients, compare, for instance, L. M. Milne-Thomson, *The Calculus of Finite Differences*, pp. 384–414, MacMillan, London, 1933; or N. E. Nörlund, *Differenzenrechnung*, pp. 295–300, Springer, Berlin, 1924. The treatment and several results are apparently due to Ostrowski, 1st ed.

Chapter 13

Most of the material in this chapter is unpublished. The theorem of Eneström and Kakeya is contained in a paper by Eneström in Swedish on the theory of pensions ("Härledning af en allmän formel för antalet pensionärer, som vid en godtycklig tidpunkt förefinnas

inom en sluten pensionskassa," *Ofversigt af vetenskaps Akad. Förhandl.* **50**, 405–415, 1893) and remained, of course, completely unknown. It was rediscovered by S. Kakeya, "On the limits of the roots of an algebraic equation with positive coefficients," *Tōhoku Math. J.* **2**, 140, 1912. The discussion of Sections 8–11 was given in the first edition only for $p = 1$ but the general case had to be treated in this edition since it is used in the Appendix P. This generalization has also been derived by essentially the same method by Traub [9].

Chapter 14

The infinite series (14.2) goes back to Newton and Euler and has also been considered (without discussion of convergence) by Theremin (*J. Reine Angew. Math.* **49**, 178–243, 1855) and E. Schröder [5, p. 329]. Theorems 14.1 and 14.2 are due to Ostrowski, 1st ed.

Chapters 15 and 16

The material in this chapter has not been previously published. For the background of the method compare the bibliographical note to Appendix O.

Chapters 17 and 18

Compare the paper by the author: "On the approximation of equations by algebraic equations," in the *J. Soc. Ind. Appl. Math.* **1**, series B, 104–130.

Chapter 19

Results about norms of vectors and matrices for $p = 1, \infty$ are given explicitly by V. N. Faddeeva in her book *Computational Methods of Linear Algebra*, Dover, New York, 1959, translated from the Russian 1950 edition by C. D. Benster. Previous formulations can be found in T. Rella, "Über den absoluten Betrag von Matrizen," *Proc. Intern. Congr. Math., Oslo, 1936*, Vol. II, pp. 29–31, and "Über positiv-homogene Funktionen ersten Grades einer Matrix," *Monatsh. Math. Physik* **48**, 84–95, 1939. For a more general treatment of norms of matrices compare A. Ostrowski, "Über Normen von Matrizen," *Math. Z.* **63**, 2–18, 1955, and A. S. Householder, "On norms of vectors and matrices," *Oak Ridge Nat. Lab. Rept. 1756*, 1954. Theorem 19.1 is due to Frobenius. Theorem 19.3 is due to Ostrowski, 1st ed.

Chapters 20 and 21

The results were published by the author without proofs in *Compt. Rend.* **244**, 288–289, 1957.

Chapter 22

Theorems 22.1 and 22.2 were published by the author without proofs in *Compt. Rend.* **244**, 288–289, 1957. The criterion (22.23) for the convergence of an iteration goes back to Scarborough, *Numerical Mathematical Analysis*, 1st ed., Johns Hopkins Press, Baltimore, Maryland, 1930. Another special case appears in G. Schulz, "Über die Lösung von Gleichungen durch Iteration," *Z. Angew. Math. Mech.* **22**, 234–235, 1942. In the second edition the condition of continuity of partial derivatives in the fixed point has been replaced by the condition of existence of the total differential in this point.

Chapter 23

The most part of the content of this chapter belongs to the classical theory of matrices. Estimate (23.16) is apparently due to U. Wegner, "Contributi alla teoria dei procedimenti iterativi per la risoluzione numerica dei sistemi di equazioni lineari algebriche," *Mem. Accad. Naz. Lincei* 4, 1–49, 1953.

Chapter 24

The content of Sections 1–9 is partly a further development of the material given in the author's paper, quoted in the note to Chapter 19. The Sections 10–11 develop a result of the author's paper, "Sur la variation de la matrice inverse d'une matrice donnée," *C.R. Acad. Sci. Paris* 231, 1019–1021, 1950. I owe the idea of using relation (23.21) to an observation by Professor F. L. Bauer.

Chapter 25

The idea of the method was indicated by Cauchy in a note in *Compt. Rend.* 25, 536–538, 1847. Its different modifications and developments were studied extensively for the case in which $f(\xi)$ is a quadratic polynomial (cf., for an account of this development, Householder [1, pp. 48–49]). Theorem 27.1 is unpublished in this form, but there are similar developments in the literature. However, usually the choice of r_μ in (27.12) was such as to make $f(\xi)$ a minimum either along the line $\xi = \xi_\mu + t\psi_\mu$ or in the neighborhood of this r_μ , while the discussion in Chapter 29, Sections 6–11, shows that this approach cannot be recommended in the general case. Compare Householder [1, pp. 48–49] and A. A. Goldstein, *Numerische Mathematik* 4, 146–150, 1962, where further bibliography can be found. The concept of convergence to the set Ω^* in this connection is apparently new.

Chapter 26

The results of this chapter are new.

Chapter 27

The results of this chapter are new, but the special case that $\psi_\mu = \phi_\mu$ was treated in Goldstein's paper, quoted in the bibliographical notes to Chapter 25.

Chapter 28

The approach discussed here was first published in the author's note "Une méthode générale de résolution automatique d'une équation polynomiale," pp. 179–182, in *Programmation en mathématique numérique*, Actes Colloq. Internat. C.N.R.S. No. 165, Besançon, 1966 (Editions Centre Nat. Recherche Sci., Paris, 1968). The Ω test was first published in author's note "A method for automatic solution of algebraic equations," pp. 209–224, in *Constructive Aspects of the Fundamental Theorem of Algebra, IBM Symposium, Zürich, 1967*, Wiley, New York, 1969.

Chapters 29 and 30

The material of these chapters was first published in the author's last paper mentioned in the note to Chapter 28.

Chapters 31–36

The material of these chapters belongs nowadays to the classical part of the functional analysis. Many minor details, though, are unpublished.

Chapter 37

Most of the material in this chapter is unpublished. For the bibliographical information, concerning the special case of a finite system of equations in the Euclidean space, see the bibliographical note to Chapter 24 of the second edition, p. 331.

Chapters 38 and 39

The first fundamental result in the direction of the subject of these chapters goes back to L. V. Kantorovich, who succeeded in transporting the proofs developed in the case of one equation and a system of $n > 1$ equations in Euclidean space, to the general case of Banach space, and even succeeded in obtaining the same constants. The observation that the existence and boundedness of the second gradient could be replaced by the Lipschitz condition for the first gradient, was then developed step-by-step in some American, Hungarian, and Russian publications. For the improved convergence bounds see the author's note, *C.R. Acad. Sci. Paris* **272** (A), 1251–1253, May 10, 1971. In the case of one equation in the Euclidean space these bounds were already given in the paper by the author, "On convergence conditions for the Newton–Raphson iteration," *BMN* **26**, Mimeo. prepublication copy, July 1969.

Chapter 40

Most of the material in this chapter is unpublished. In Theorem 40.2 the special case $S := S_0$ corresponds to author's first publications and the special case $S := K$ to a unicity theorem given by L. V. Kantorovich.

Chapter 41

The results of these sections, for a special choice of norms, were first published by the author in two notes in *Compt. Rend.* **231**, 1114–1116, 1950; **232**, 786–788, 1951. They were partly developed in more detail by the author in the article "Simultaneous systems of equations" in the proceedings of a symposium: *Simultaneous Linear Equations and the Determination of Eigenvalues*, Nat. Bureau of Standards, Appl. Math. Series **29**, 1953.

Chapter 42

The first convergence proof of the Newton–Raphson method for $n \geq 2$ was given by H. B. Fine in the article quoted in the note to Chapter 7. However, Fine used a bound of $\|J(\xi)^{-1}\|$ in a whole neighborhood of ξ_0 . The first proofs using only $\|J(\xi_0)^{-1}\|$ were given by the author (1936) for $n = 2$ and, for $n \geq 2$, by K. Bussmann (cf. *Z. Angew. Math. Mech.* **22**, 361–362, 1942).

Appendix A

The results were first given in Ostrowski [4, pp. 209–212, 218].

Appendix B

A more precise result is given in Ostrowski [4, pp. 212–217], but the proof given here is more elementary.

Appendix C

Published by the author without proof in 1957, *Compt. Rend.* **244**, 429–430. In the meantime I found in a paper by U. T. Bödewadt, “Die Kettenregel für höhere Ableitungen,” *Math. Z.* **48**, p. 740, 1942–1943, the formula (C.4) as the formula (21), although with a more complicated proof. The priority of the discovery of this formula belongs therefore to U. T. Bödewadt.

Appendix D

The exposition is only in formal respect different from that in Gauss’s quoted book. We give some further references for the latest developments in the direction of the generalized *regula falsi*:

1. W. M. Kincaid, “A two-point method for the numerical solution of systems of simultaneous equations,” *Quart. Appl. Math.* **18**, 313–324, 1960–1961.
2. J. W. Schmidt, “Eine Übertragung der Regula Falsi auf Gleichungen in Banach-Räumen,” *Z. Angew. Math. Mech.* **43**, 1–8, 97–110, 1963.
3. L. Bittner, “Mehrpunktverfahren zur Auflösung von Gleichungssystemen,” *Z. Angew. Math. Mech.* **43**, 111–126, 1963.

These papers contain further references.

Appendix E

The bibliography is given as footnotes in the text.

Appendix F

The content is due to Ostrowski, 1st ed.

Appendix G

The results of Sections 1–3 are contained in Ostrowski [3]. Sections 4–9 are due to Ostrowski, 1st ed. The original proof of (G.8) given in the first edition is here considerably simplified following a suggestion by L. S. and B. L. Rumshiski in Ref. [8].

Appendix H

The content is due to Ostrowski, 1st ed.

Appendix I

Improvement in several directions of the result which was published in A. M. Ostrowski, “A method of speeding up iterations with super-linear convergence,” *J. Math. Mech.* **7**, 117–120, 1958.

Appendix J

The discussion of this appendix arose from the desire to clarify the background of Whittaker's series from Section 5, which was published by E. T. Whittaker, "A formula for the solution of algebraic or transcendental equations," *Proc. Math. Soc. Edinburgh* **36**, 103–106, 1918, and E. T. Whittaker and G. Robinson, *The Calculus of Observation*, § 60, Blackie and Son, London and Glasgow, 1928. The idea of this approximation appears to go back to A. De Morgan, *J. Inst. Actuaries* **14**, 353, 1868. Otherwise the treatment is apparently due to Ostrowski, 1st ed.

Appendix K

The results were first given in A. Ostrowski, "Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrizenelementen," *Jber. Deut. Mat.-Verein.* **60**, 40–42, 1957. A translation of this note by G. C. Bump, edited by G. E. Forsythe, appeared as Tech. Report No. 2, Appl. Math. and Stat. Labs., Stanford Univ., Stanford, California.

Appendix L

The formulas of this appendix are classical but the complete formulas (L.7) and (L.8) appear to be missing in standard texts.

Appendix M

The general remainder formula goes back to Cauchy but the discussion of the confluent case had to be partly developed anew.

Appendix N

The content of this appendix is apparently new.

Appendix O

The original publication of Laguerre concerning the method in question is found in *Nouvelles Annales de Mathématiques* (2) **19**, 161–172, 193–202, 1880, reprinted in *Oeuvres de Laguerre* **1**, 87–103, Paris, 1898. The theory was first presented in detail by H. Weber in his *Lehrbuch der Algebra*, Vol. I, pp. 322–331, Braunschweig, 1895, and in R. Fricke's *Lehrbuch der Algebra*, Vol. I, pp. 252–258, Braunschweig, 1924.

The cubic character of convergence and the modifications necessary in the case of multiple roots were first discussed by E. Bodewig, *Proc. Acad. Amsterdam* **49**, 911–921, 1946, and J. G. van der Corput, *Proc. Acad. Amsterdam* **49**, 922–929, 1946. Our method of proof is essentially simpler than the standard one but also goes back to Laguerre, who alluded to it in a footnote in his original paper. (Compare also N. Obreschkoff's book, *Verteilung und Berechnung der Nullstellen reeller Polynome*, pp. 261–263, Berlin, 1963.)

In all standard presentations of this method N in formula (O.9) is the exact degree of $f(x)$. The use of these formulas with an N that is only restricted to being greater than or equal to the degree of $f(x)$ and the connected developments are unpublished. In this connection the method discussed in Chapters 15 and 16 is obtained for $N \rightarrow \infty$.

Appendix P

Compare the note to Chapter 18.

Appendix Q

This material was first published in the author's note in *Compt. Rendu* **275** (A), 275–278, July 24, 1972.

Appendix R

The material of this appendix was first published in the author's paper in *SIAM J. Numer. Anal.* **8**, 623–638, 1971.

Appendix S

The material of this appendix has been first given in the prepublication, "An improved general routine for solving algebraic equations," *BMN* **15**, August 1967.

Appendix T

The result of this appendix has probably often been published. But the author is unable to give any reference.

Appendix U

Most of the material of this appendix is unpublished. The concept of a strictly normed space is referred to by P. Fischer and Gy. Muszély in their paper, "On some new generalizations of the functional equation of Cauchy," *Can. Math. Bull.* **10**, 197–205, 1967.

This Page Intentionally Left Blank

Index

A

- Accelerating convergence*, methods for ... , 40, 296, 306, 320
Additive operator, 220
Aitken, A. C., 296
Aitken's transformation, 296
Algebraic equation in one unknown, solution by the gradient method, 177
Ampère, A. M., 399
Asymptotic behavior of solutions of linear differential equations, 90, ... errors, 39, 49, 56, 74, 82, 91, 100, 101, 112, 113, 134, 307–309, 359, 371
Attraction, points of ... , 38, 146

B

- Ball, 215
Banach, S., 210, 227
Bauer, F. L., 403
Bauer, M., 400
Bernoulli–L'Hôpital rule, 343
Bernoullian method for solution of equations, 332
Bittner, L., 405
Bodewig, E., 406
Bödewadt, U. T., 405
Borel's covering theorem, 229
Bounded operator, 219
Bussmann, K., 404

C

- Cauchy, A. L., 11, 91, 166, 399, 400, 401, 406
Cauchy–Bolzano condition, 209, 215
Cauchy sequence, 215
Cauchy's convergence theorem, 50
Cauchy's theorem on algebraic equations, 91

- Cauchy–Schwarz inequality, 137
Center of an iteration, 38
Characteristic equation, root, vector of a matrix, 140
Characteristic polynomial of a linear differential equation, 84, special ..., 96, table of zeros of ..., 102
Closed interval, 1, 4, 22
Coefficient field, 208
Coincident interpolation points, 53
Common zeros in numerator and denominator, 340
Compact, 210
Complete, 210
Complex variable, functions of ..., 23, 45, 60, 104, 127, 131, 351
Confluent divided differences, 8, 339
Conformal mapping, 104
Connected zeros, 277, 284
Continuity, relative, of roots, 281
Contraction bound, 215
Contracting operator, 215
Convergence, 209, cubic, 112, linear, 31, 43, weakly linear, 174, supralinear, 36, 296
Convex polygonal line, 288
van der Corput, J. G., 406

D

- Dandelin, G. P., 401, ... bounds, 73
Darboux, G., 19, 399
Darboux's theorem, 19, 399
Definiteness of a quadratic form, 156
Degree of convergence, 43
Derivatives of the inverse function, 20, 405
Derived set, 173
Diagonal form of a matrix, 141
Difference equation, linear, 32, 84
Directional derivatives, 236

- Divided differences*, 1, in the confluent case (with repeated arguments), 8
 Division of power series, 86
 Double precision, 37
- E**
- Efficiency index, 33, 55, 100
 Eigenvalue, eigenvector of a matrix, 140
 Eneström, G., 99, 401
 Entire functions, 124
 Error estimates *a priori, a posteriori*, 56
 Errors *a priori, a posteriori*, 372
 Euclidean length, 155
 Euler, L., 402
 Exact multiplicity, 61
 Examples of computation, 36, 83, 109, 306, 333
 Extrapolation versus interpolation, 306
- F**
- Faà di Bruno, F., 349
 Faddeeva, V. N., 402
F-differential, 239, *F*-gradient, 239
 Fatou differential, 239, Fatou gradient, 239
 Feedback, 68
 Fejér, L., 400
 Fibonacci sequence, 32
 Fine, H. B., 401, 404
 Fischer, P., 407
 Fixed point, *see* Center of an iteration
 Forsythe, G. E., 406
 Fourier, J. B., 400, 401, Fourier bounds, 69, Fourier conditions, 30, 69
 Fricke, R., 406
 Frobenius, G., 402
 Frobenius norm, 155
 Functional analysis, 55, 404
Fundamental equation, root, vector of a matrix, 140
- G**
- Gateau, *G*-differential, 237, ... gradient, 237
 Gauss, C. F., 294
 Geiringer-von Mises, H., 400
 Generating series, 84
 Genocchi, A., 11
 Goldstein, A. A., 403
 Günther, S., 331
- H**
- Hadamard, J., 332, ...'s estimate of a determinant, 335
 Hardy, G. H., 287
 Hermite's integral representation, 3
 Hermitian matrix, 156, ... form, 156
 Hessian matrix, 171
 Hilbert space, 208
 Hölder's inequality, 135
 Horner unit, 32
 Householder, A. S., 296, 298, 399, 402, 403
- I**
- Inner product of vectors, 136
 Integral representation of divided difference, 4, 5, 6, 10
 Interior of an interval, 15
 Interpolating function, 15
Interpolation abscissas, points, 15, multiple, 15
 Interpolation formula, 13, general ..., 14
 Interpolation function, 15
 Interval in a LS, 208
Interval, closed, 1, 3, 22, ... open, 1
 Inverse function, 20, 103, 290
 Inverse interpolation, 18, 28, 53
 Inverse operator, 224
 Isobaric polynomial, 21
 Iterating function, 38
Iteration, 38, ... of order k , 320
- J**
- Jacobian*, 176, ... matrix, 151
 Jordan's canonical form of a matrix, 141
 J_m -routine, 197, ... *J*-test, 196
- K**
- Kakaya, S., 99, 401, 402
 Kantorovich, I. V., 404
 Karamata, J., 400
 Kincaid, W. M., 405
 Kronecker, L., 329
- L**
- Lagrange interpolation, 15
 Laguerre, E. N., 360, 406, ... iteration, 353
L-differential, 238, ... -gradient, 238
 Linear, weakly linear convergence in the gradient method, 174, 178

- Linear functions as interpolating functions, 79
 Linear operator, 220
 Linear space, LS, 208
 Lipschitz condition, 231, ... inequality, 211,
 ... constant, 211
 Littlewood, J. E., 287
- M**
- Matrix*, 137, "size" of ..., 137
 Mapping, 215
 Mean value formula for divided difference, 5
 Metric space, MS, 214
 Metricization, 215
 Milne-Thomson, L.M., 399, 401
 Minowski inequality, 136, ... norm, 159
Minimum regular ... 174
 von Mises, R., 400
 Modulus, 207
 Monotonic iterating functions, 47
 de Morgan, A., 406
Multiple roots, 48, 61, 113, 339, 340, 343,
 ... interpolation abscissas, 103, 339
 Muszély, Gy., 407
- N**
- Neighborhood, one-sided, 38
 Newton, I., 53, 287, 399, 402
 Newton's interpolation formula, 12
Newton-Raphson method, 42, 53, 55, 56, 69,
 73, its generalization for several
 variables, 270, its analog for multiple
 roots, 61
 Nörlund, N. E., 401
 Norm, 135, 137, 155, 157, 159, 161, 208
 Normed linear space, NLS, 208
 Norms of vectors, 135
- O**
- Obreschkoff, N., 406
 Ω -test, 190
 One-sided neighborhood, 38
 Open interval, 1, 22
 Operator, 219
- P**
- Poincaré, H., 87
Points of attraction, 87, ... of repulsion, 150,
 152, ... of definite repulsion, 38
 Pólya, G., 287
- Polynomial interpolation, 15
 Popoviciu, E., 399
Positive definite, semidefinite quadratic
 forms, 156
Power series, use of, 85, 324
 Principle of permanence, 23
Products, infinite, 75
- Q**
- q* acceleration, 200
 Quadratic convergence, 62, 400
 Quadratic forms, 156
- R**
- Rall, L. B., 401
 Raphson, J., 53
Regula falsi, 27, 51, 55, 83, 90, for two
 equations, 294
Regular matrix, 140, ... minimum, 174
 Relative continuity, 281
 Rella, T., 402
 Remainder in Taylor's formula, 389
 Remainder terms of interpolation for-
 mulas, 13, 15, 339
Repulsion, points of ..., 150, 152, points
 of definite ..., 39
 Ritt, J. F., 400
 Robinson, G., 41, 400, 406
 Rolle's theorem, 97, 110, 339
 Roots of special algebraic equations, 91, 97
 Rouché's theorem, 105, 278
 Rounding-off rule, 33, 310
 Row vector, 135
 Rumshiski, L. S. and B. L., 405
 Runge, C., 12
 Runge's notation, 12
- S**
- Scarborough, J., 402
 Schmidt, J. W., 405
Schröder, E., 399, 400, 401, 402, ...'s series,
 26, 358, its analog for multiple roots,
 343
 Schröder, J., 400
 Schröder's formula, 61
 Schulz, G., 402
 Square root iteration, 110
 Stability of convergence of A^n , 145
Steepest descent, 166, method of ... 178
 Steffensen, J. F., 399

- Stieltjes, T. J., 399
 Strictly normed spaces, 394
 Strong convergence, 221
 Subharmonic functions, 172
 Supralinear convergence, 320
 Sylvester's determinant formula, 325
 Symbols, *see* List of Notations, p. xvii
 Symmetric matrices, 156
 Symmetry of divided differences, 2
- T
- Taylor approximation* to the roots, errors
 of ..., 106
 Taylor development of the root, 102
 Theremin, F., 402
 Total differentiability, 150
 Transform of a matrix, 140
 Transpose of a vector, 137
 Traub, J. F., 399, 402
 Triangle inequality, 136, 160
- Triangular scheme for divided differences,
 17
- V
- Vainberg, M. M., 238
 Vector, 135
- W
- Weak convergence, 221
 Weakly linear convergence, *see* Convergence
 Weber, H., 406
 Wegner, U., 403
 Weight in a polynomial, 21
 Weissinger, J., 400
 Whittaker, E. T., 41, 326, 400, 406
 Willers, T. A., 296, 399, 400
 Wronski, H., 325
- Z
- Zeros of interpolating polynomials*, 127 ...
 of special polynomials, 97