

SOLUTION OF EQUATIONS AND SYSTEMS OF EQUATIONS

Second Edition

A. M. OSTROWSKI

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BASEL
BASEL, SWITZERLAND

1966



ACADEMIC PRESS New York and London

COPYRIGHT © 1966, BY ACADEMIC PRESS INC.

ALL RIGHTS RESERVED.

**NO PART OF THIS BOOK MAY BE REPRODUCED IN ANY
FORM, BY PHOTOSTAT, MICROFILM, OR ANY OTHER MEANS,
WITHOUT WRITTEN PERMISSION FROM THE PUBLISHERS.**

ACADEMIC PRESS INC.

111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
Berkeley Square House, London W.1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 65-22762

PRINTED IN THE UNITED STATES OF AMERICA

Preface to the First Edition

The course from which this book has developed was given at the invitation of the National Bureau of Standards in the summer of 1952 at the American University, Washington, D.C. Notes of the lectures were taken by Mr. M. Ticson and reproduced in a small number of copies as the working paper CL-52-2 for the internal use of the NBS.

As this course contained a great deal of unpublished material, it appeared advisable to publish the whole as a book. For this purpose, the whole course was worked through anew and completely rewritten, much new material was incorporated in the lectures, and 11 appendices were added to deal with problems which could not be treated in the main text of the book. The treatment is still far from being complete, however.

While the lectures as they were given were intended more or less for the undergraduate level, in its present form the book presents methods and procedures of argumentation which conform more to the graduate level. I think that, in view of the increasing practical importance of numerical methods, it is essential for the mathematician to apply all his techniques to clarify the theoretical basis of these methods. Thus, I hope this book will help to a certain degree to bridge the gap that still exists between "pure" and "practical" mathematics.

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 in 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.

A. M. OSTROWSKI

June, 1960

Preface to the Second Edition

For this second edition, the entire text was thoroughly revised and much new material added so that the size of the book is almost doubled.

The new material deals with those methods which can be used with the automatic computer without any special preparation. The Laguerre iteration and its modifications are extensively analyzed, since this iteration can be used, at least for polynomials with only real zeros, starting with an arbitrary real value. In two chapters and one appendix we treat the approximation of a zero by zeros of interpolating polynomials, as the extensive experimentations by D. I. Muller make it appear probable that in many cases this method is not sensitive with respect to the choice of the starting value. In several chapters we deal with the method of steepest descent. Although, in the case of one variable, the complete working through of this method to its practical use with a computer was achieved too late to be included in the book, the method of steepest descent gives a nonsensitive although rather slow approach for large classes of systems of equations, a subject that was somewhat neglected in the first edition. In this respect the four chapters preceding the generalization of the Newton-Raphson method to the case of several variables may be welcome to "pure" as well as to "applied" numerical analysts. Finally, the discussion of the theory of divided differences, added in this edition, may help to make this theory more widely known.

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. Ruedi, 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.

A. M. OSTROWSKI

May, 1966

Contents

PREFACE TO THE FIRST EDITION	v
PREFACE TO THE SECOND EDITION	vii
1. DIVIDED DIFFERENCES	1
Divided Differences for Distinct Arguments	1
Symmetry	2
Integral Representation	3
Mean Value Formulas	5
Divided Differences with Repeated Arguments	6
A Formula for Confluent Divided Differences	7
Newton's Interpolation Formula	8
General Interpolation Problem	10
Polynomial Interpolation	12
The Remainder for a General Interpolating Function	12
Triangular Schemes for Computing Divided Differences	13
2. INVERSE INTERPOLATION. DERIVATIVES OF THE INVERSE FUNCTION. ONE INTERPOLATION POINT	15
The Concept of Inverse Interpolation	15
Darboux's Theorem on Values of $f'(x)$	16
Derivatives of the Inverse Function	17
One Interpolation Point	20
A Development of a Zero of $f(x)$	24
3. METHOD OF FALSE POSITION (REGULA FALSI)	26
Definition of the Regula Falsi	26
Use of Inverse Interpolation	27
Geometric Interpretation (Fourier's Conditions)	29

Iteration with Successive Adjacent Points	30
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	38
A Convergence Criterion for an Iteration	38
Points of Attraction and Repulsion	38
Improving the Convergence	40
5. FURTHER DISCUSSION OF ITERATIONS. MULTIPLE ZEROS	49
Iterations by Monotonic Iterating Functions.	49
Multiple Zeros	51
Connection of the Regula Falsi with the Theory of Iteration	54
6. NEWTON-RAPHSON METHOD	56
The Idea of the Newton-Raphson Method.	56
The Use of Inverse Interpolation	57
Comparison of Regula Falsi and Newton-Raphson Method	58
7. FUNDAMENTAL EXISTENCE THEOREMS FOR NEWTON-RAPHSON ITERATION	59
Error Estimates a Priori and a Posteriori	59
Fundamental Existence Theorems	59
8. AN ANALOG OF THE NEWTON-RAPHSON METHOD FOR MULTIPLE ROOTS	67
9. FOURIER BOUNDS FOR NEWTON-RAPHSON ITERATION	70
10. DANDELIN BOUNDS FOR NEWTON-RAPHSON ITERATION	75

11. THREE INTERPOLATION POINTS	82
Interpolation by Linear Fractions	82
Two Coincident Interpolation Points	83
Error Estimates	84
Use in Iteration Procedure	86
12. LINEAR DIFFERENCE EQUATIONS	88
Inhomogeneous and Homogeneous Difference Equations	88
General Solution of the Homogeneous Equation	89
Lemma on Division of Power Series	90
Asymptotic Behavior of Solutions of (12.1)	92
Asymptotic Behavior of Errors in the Regula Falsi Iteration	94
A Theorem on Roots of Certain Equations	96
13. n DISTINCT POINTS OF INTERPOLATION	99
Error Estimates	99
Iteration with n Distinct Points of Interpolation	100
Discussion of the Roots of Some Special Equations	102
14. $n + 1$ COINCIDENT POINTS OF INTERPOLATION AND TAYLOR DEVELOPMENT OF THE ROOT	109
Statement of the Problem	109
A Theorem on Inverse Functions and Conformal Mapping	109
Theorem on the Error of the Taylor Approximation to the Root	112
Discussion of the Conditions of the Theorem	113
15. THE SQUARE ROOT ITERATION	117
16. FURTHER DISCUSSION OF SQUARE ROOT ITERATION	127
17. A GENERAL THEOREM ON ZEROS OF INTERPOLATING POLYNOMIALS	136

18. APPROXIMATION OF EQUATIONS BY ALGEBRAIC EQUATIONS OF A GIVEN DEGREE. ASYMPTOTIC ERRORS FOR SIMPLE ROOTS	141
19. NORMS OF VECTORS AND MATRICES	145
20. TWO THEOREMS ON CONVERGENCE OF PRODUCTS OF MATRICES	153
21. A THEOREM ON DIVERGENCE OF PRODUCTS OF MATRICES	156
22. CHARACTERIZATION OF POINTS OF ATTRACTION AND REPULSION FOR ITERATIONS WITH SEVERAL VARIABLES	161
An Example	165
23. FURTHER DISCUSSION OF NORMS OF MATRICES. $\Delta q(A)$. .	167
Triangle Inequality	167
Bilinear and Quadratic Forms of Symmetric Matrices	169
Estimate of $\Delta_p(ABC)$	170
Variation of $\Delta_p(A^{-1})$	171
Length of Arc in the $ \xi _p$ Metric	173
24. AN EXISTENCE THEOREM FOR SYSTEMS OF EQUATIONS . .	176
Formulation of the Theorem	176
Proof of Theorem 24.1	177
A Uniqueness Theorem	179
Example	180
25. n-DIMENSIONAL GENERALIZATION OF THE NEWTON- RAPHSON METHOD. STATEMENT OF THE THEOREMS . .	183
Variation of the Jacobian Matrix	184
Statement of the n -Dimensional Analog of the Newton-Raphson Method	186

26. <i>n</i>-DIMENSIONAL GENERALIZATION OF THE NEWTON-RAPHSON METHOD. PROOFS OF THE THEOREMS	189
27. METHOD OF STEEPEST DESCENT. CONVERGENCE OF THE PROCEDURE	195
Idea of the Method	195
Convergence of the Procedure	198
Application to $ f(x + iy) ^2$	200
28. METHOD OF STEEPEST DESCENT. WEAKLY LINEAR CONVERGENCE OF THE ξ_μ	203
The Derived Set at the ξ_μ	203
Weakly Linear Convergence	204
Condition for the Regular Minimum of the Function (27.3)	207
Algebraic Equations with One Unknown	208
29. METHOD OF STEEPEST DESCENT. LINEAR CONVERGENCE OF THE ξ_μ	210
Example	213

APPENDICES

A. Continuity of the Roots of Algebraic Equations	220
B. Relative Continuity of the Roots of Algebraic Equations	225
C. An Explicit Formula for the nth Derivative of the Inverse Function	235
D. Analog of the Regula Falsi for Two Equations with Two Unknowns	239
E. Steffensen's Improved Iteration Rule	241
F. The Newton-Raphson Algorithm for Quadratic Polynomials	247
G. Some Modifications and Improvements of the Newton-Raphson Method	251
H. Rounding Off in Inverse Interpolation	256
I. Accelerating Iterations with Superlinear Convergence	267
J. Roots of $f(z) = 0$ in Terms of the Coefficients of the Development of $1/f(z)$	271

K. Continuity of the Fundamental Roots as Functions of the Elements of the Matrix	282
L. The Determinantal Formulas for Divided Differences	284
M. Remainder Terms in Interpolation Formulas	288
N. Generalization of Schröder's Series to the Case of Multiple Roots	293
O. Laguerre Iterations	305
P. Approximation of Equations by Algebraic Equations of a Given Degree. Asymptotic Errors for Multiple Zeros	316
Bibliographical Notes	327
INDEX	335

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

$$[x_1, x_2]_x f \equiv \frac{f(x_2) - f(x_1)}{x_2 - x_1} \quad (x_1 \neq x_2). \quad (1.1)$$

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

$$[x_1, x_1]_x f \equiv f'(x)_{x=x_1}, \quad (1.2)$$

provided $f'(x_1)$ exists.

The subscript x can be dropped either if $f(x)$ is defined as a function of *one* variable x only, or if the symbol $[x_1, x_2]f$ has the meaning

$$[x_1, x_2]f \equiv [x_1, x_2]_x f. \quad (1.3)$$

If $f(x)$ does not depend on x_1, x_2 as parameters, the expression (1.1) 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

$$[x_1, x][x_2, x]f = [x_2, x][x_1, x]f, \quad (1.4)$$

provided $f(x)$ does not depend on x_1 and x_2 . Indeed, the common value of both sides of (1.4) is seen at once to be

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

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

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

$$[x_1, x][x_2, x] \dots [x_m, x]_t f \equiv [x, x_1, \dots, x_m]_t f. \quad (1.5)$$

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

SYMMETRY

4. The expression (1.5) is, by (1.4), 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 (1.5) is *symmetric in all $m + 1$ arguments*. For this it is sufficient to prove that (1.5) does not change if x is interchanged with x_1 . But, if we put

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

the expression (1.5) becomes

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

and our assertion follows from what has been said about (1.3). In (1.5) $[x, x_1, \dots, x_m]$ 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 (1.5) 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 (1.5) follows (if we interchange x with x_k , replace m by $m - 1$, and correspondingly shift the indices) the general formula

$$[x_1, \dots, x_m]f = [x_1, \dots, x_k][x_k, \dots, x_m]f, \quad (1.6)$$

provided $f(x)$ does not contain any of the variables x_1, \dots, x_m as parameters, and, of course, the analogous decomposition in a product of several operators.

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

$$[x_1, x]x^m = \frac{x_1^m - x^m}{x_1 - x} = \sum x^p x_1^{q_1} \quad (p + q_1 = m - 1),$$

$$\begin{aligned} [x_2, x][x_1, x]x^m &= \sum_{p+q_1=m-1} x_1^{q_1} [x_2, x]x^p = \sum x_1^{q_1} x^p x_2^{q_2} \\ &\quad (p + q_1 + q_2 = m - 2), \end{aligned}$$

$$\begin{aligned} [x_k, x_{k-1}, \dots, x_1, x]x^m &= \sum x^p x_1^{q_1} \dots x_k^{q_k} \\ &\quad (p + q_1 + \dots + q_k = m - k), \end{aligned}$$

$$[x_{m-1}, \dots, x_1, x]x^m = x_{m-1} + x_{m-2} + \dots + x_1 + x,$$

$$\left. \begin{array}{l} [x_m, x_{m-1}, \dots, x_1, x]x^m = 1, \\ [x_{m+1}, x_m, \dots, x_1, x]x^m = 0. \end{array} \right\} \quad (1.7)$$

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

INTEGRAL REPRESENTATION

6. We will denote the smallest convex (closed) polygonal domain containing all x, x_v ($v = 1, \dots, m$) by $P(x, x_1, \dots, x_m)$ 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 $P(x, x_1, \dots, x_m)$. We have obviously

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

Introducing here the new variable of integration t_1 by $t = (1 - t_1)x_1 + t_1x_2$, $dt = -(x_1 - x_2)dt_1$, we obtain, replacing x_2 by x ,

$$[x_1, x]f = \int_0^1 f'[(1 - t_1)x_1 + t_1x] dt_1.$$

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

$$[x_2, x_1, x]f = \int_0^1 dt_1 \int_0^1 t_1 f''[(1 - t_1)x_1 + t_1(1 - \tau_2)x_2 + t_1\tau_2 x] d\tau_2;$$

here we introduce in the inner integral the new variable of integration $t_2 = t_1\tau_2$, $dt_2 = t_1 d\tau_2$, 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

$$\begin{aligned} [x_m, \dots, x_1, x]f &= \int_0^1 \int_0^{t_1} \dots \int_0^{t_{m-1}} f^{(m)}[(1 - t_1)x_1 + (t_1 - t_2)x_2 + \dots \\ &\quad + (t_{m-1} - t_m)x_m + t_m x] dt_m \dots dt_1 \end{aligned} \quad (1.8)$$

and this formula is immediately verified by induction.

For $f = x^m/m!$ the left-side expression in (1.8) becomes, by (1.7), $1/m!$ and $f^{(m)} \equiv 1$.

We see that we have

$$\int_0^1 \int_0^{t_1} \dots \int_0^{t_{m-1}} dt_m \dots dt_1 = \frac{1}{m!}. \quad (1.9)$$

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 (1.8) are *nonnegative*. One consequence of this is that the argument

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

lies in the real case in $\langle x, x_1, \dots, x_m \rangle$ and in the general case in $P(x, x_1, \dots, x_m)$.

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 the expression (1.8) replacing there $f^{(m)}$ by the constants Ω and ω , that is to say, replacing f by $\Omega x^m/m!$ and $\omega x^m/m!$. We see that (1.8) is contained between $\Omega/m!$ and $\omega/m!$. We have therefore, as $f^{(m)}$ is assumed to be continuous,

$$[x_m, \dots, x_1, x]f = \frac{f^{(m)}(\xi)}{m!}, \quad \xi \prec \langle x, x_1, \dots, x_m \rangle. * \quad (1.10)$$

If x, x_i are not all real, the modulus of $f^{(m)}$ assumes its maximum in a point ξ of $P(x, x_1, \dots, x_m)$ and the modulus of (1.8) is then

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

so that we have

$$[x_m, \dots, x_1, x]f = \theta \frac{f^{(m)}(\xi)}{m!}. \quad |\theta| \leq 1, \quad \xi \prec P(x, x_1, \dots, x_m). \quad (1.11)$$

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

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

Then we have, denoting by ζ an arbitrarily chosen point of $P(x, x_1, \dots, x_m)$, subtracting $f^{(m)}(\zeta)/m!$ from both sides of (1.8), and using (1.9),

* The sign \prec is the *symbol of inclusion*; the formula $a \prec K$ is to be read, a is contained in K . In the same way, the formula $K \succ a$ is to be read, K contains a .

$$[x_m, \dots, x_1, x]f - \frac{f^{(m)}(\zeta)}{m!} = \int_0^1 \int_0^{t_1} \dots \int_0^{t_{m-1}} F(t_1, \dots, t_m) dt_m \dots dt_1,$$

$$\begin{aligned} F(t_1, \dots, t_m) &= f^{(m)}[(1-t_1)x_1 + \dots \\ &\quad + (t_{m-1}-t_m)x_m + t_mx] - f^{(m)}(\zeta). \end{aligned}$$

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

$$\left| \int_0^1 \int_0^{t_1} \dots \int_0^{t_{m-1}} F(t_1, \dots, t_m) dt_m \dots dt_1 \right| \leq \frac{DM_{m+1}}{m!},$$

$$[x_m, \dots, x_1, x]f = \frac{f^{(m)}(\zeta)}{m!} + \theta D \frac{M_{m+1}}{m!}, \quad |\theta| \leq 1. \quad (1.12)$$

DIVIDED DIFFERENCES WITH REPEATED ARGUMENTS

10. From (1.10) it follows that if $x \rightarrow t$, $x_\nu \rightarrow t$ ($\nu = 1, \dots, m$), the expression (1.8) tends to $f^{(m)}(t)/m!$. This limit can therefore be considered as the reasonable definition of $[t, \dots, t]_x f$, where the argument t is repeated $m + 1$ times. We will denote this symbol more simply by $[t^{m+1}]_x f$ and we have

$$[t^{m+1}]_x f(x) = \frac{f^{(m)}(t)}{m!}. \quad (1.13)$$

In the complex case, $f(x)$ is assumed to be regular in $P(x, x_1, \dots, x_m)$, and therefore (1.12) can be used instead of (1.10). The above argument and the definition (1.13) remain valid, therefore, in the complex case, too.

More generally, consider k groups of (real or complex) variables

$$x_1, \dots, x_{m_1}; \quad y_1, \dots, y_{m_2}; \dots; \quad z_1, \dots, z_{m_k}$$

and the corresponding expression

$$[x_1, \dots, x_{m_1}, \quad y_1, \dots, y_{m_2}, \dots, \quad z_1, \dots, z_{m_k}]_x f, \quad (1.14)$$

where $f(x)$ does not contain the variables $x_\nu, y_\nu, \dots, z_\nu$.

Let now the x_v tend to t_1 , the y_v tend to t_2, \dots , the z_v tend to t_k . If the limit of (1.14) exists, it can be denoted by

$$[t_1, \dots, t_1, t_2, \dots, t_2, \dots, t_k, \dots, t_k]_x f,$$

where t_1 is repeated m_1 times, ..., and t_k is repeated m_k times. To indicate the number of repetitions we will use for the limit of (1.14) the notation

$$[t_1^{m_1}, t_2^{m_2}, \dots, t_k^{m_k}] f. \quad (1.15)$$

Observe that we can go to the limit in (1.8) (with $m_1 + m_2 + \dots + m_k = m + 1$), after having replaced the integration variables t_1, \dots, t_m , respectively, by τ_1, \dots, τ_m , and therefore also not simultaneously but separately to t_1, \dots, t_k and are sure to obtain the same limit. Indeed, this amounts to replacing the parameters x_v, y_v, \dots, z_v by their limits under the sign of integration and $f^{(m)}(x)$ is *uniformly* continuous. We speak in this case of the divided differences *with repeated arguments* or of *confluent* divided differences. The operator (1.15) is, of course, also a *linear operator*.

A FORMULA FOR CONFLUENT DIVIDED DIFFERENCES

11. We are now going to prove a formula for (1.15). Suppose that t_1, \dots, t_k are all distinct and that m_1, \dots, m_k are natural integers. Then we shall prove that we have

$$[t_1^{m_1}, \dots, t_k^{m_k}]_x f = \frac{\partial^{m_1-1}}{\partial t_1^{m_1-1}} \frac{\partial^{m_2-1}}{\partial t_2^{m_2-1}} \cdots \frac{\partial^{m_k-1}}{\partial t_k^{m_k-1}} [t_1, t_2, \dots, t_k]_x f, \quad (1.16)$$

assuming that the derivatives of $f(x)$ occurring in (1.16) are continuous in the points indicated and in their neighborhoods.

This formula is already proved for $k = 2$ and also for $m_1 = m_2 = \dots = m_k = 1$. Put $m_1 + m_2 + \dots + m_k = M$.

12. Then we can assume that (1.16) is already proved for all smaller values of M . Further, it can be assumed that not all m_k have the value 1 and that in particular $m_1 > 1$ since (1.14) is symmetric.

Finally, we assume that all $x_\mu, y_\mu, \dots, z_\mu$ are distinct among themselves and from t_1, \dots, t_k . We can decompose (1.14) by virtue of (1.6) as

$$[x_1, \dots, x_{m_1}]_{x_{m_1}} [x_{m_1}, y_1, \dots, y_{m_1}, \dots, z_1, \dots, z_{m_k}]_{xf}$$

and this is obviously equal to

$$[x_1, \dots, x_{m_1}]_t [t, y_1, \dots, y_{m_1}, \dots, z_1, \dots, z_{m_k}]_{xf}.$$

Now let x_1, \dots, x_{m_1} tend to t_1 ; then we obtain as the limit of (1.14)

$$\frac{\partial^{m_1-1}}{\partial t_1^{m_1-1}} [t_1, y_1, \dots, y_{m_1}, \dots, z_1, \dots, z_{m_k}]_{xf}.$$

But here (1.16) can be applied to

$$[t_1, y_1, \dots, y_{m_1}, \dots, z_1, \dots, z_{m_k}]_{xf}$$

and (1.16) is proved in the general case.

From (1.16) we conclude that the expression (1.15) is symmetric with respect to the $t_1^{m_1}, \dots, t_k^{m_k}$. From the definition of (1.15) as the limit of (1.14), it follows that the formulas (1.10), (1.11), and (1.12) remain valid if the x_ν are no longer assumed to be distinct.

NEWTON'S INTERPOLATION FORMULA

13. Consider a function $f(x)$ of one variable x and $n + 1$ *distinct* points x, x_1, \dots, x_n . Then we can write (1.1) 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

$$[x, x_1, \dots, x_{\nu-1}]f = [x_1, \dots, x_\nu]f + (x - x_\nu)[x, x_1, \dots, x_\nu]f \quad (1.17)$$

until we have for $\nu = n$

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

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

$$(x - x_1)(x - x_2) \dots (x - x_\mu) = \mathcal{X}^\mu \quad (\mu = 1, 2, \dots); \quad \mathcal{X}^0 = 1. \quad (1.18)$$

Obviously we have generally

$$\mathcal{X}^{\mu-1}(x - x_\mu) = \mathcal{X}^\mu \quad (\mu = 1, 2, \dots). \quad (1.19)$$

14. Multiplying now (1.17) by $\mathcal{X}^{\nu-1}$ we obtain by (1.19)

$$\mathcal{X}^{\nu-1}[x, x_1, \dots, x_{\nu-1}]f = \mathcal{X}^{\nu-1}[x_1, \dots, x_\nu]f + \mathcal{X}^\nu[x, x_1, \dots, x_\nu]f,$$

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

$$f(x) = \sum_{\nu=1}^n \mathcal{X}^{\nu-1}[x_1, \dots, x_\nu]f + \mathcal{X}^n[x, x_1, \dots, x_n]f$$

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

$$f(x) = L_{n-1}(f, x) + R_{n-1}(f, x), \quad (1.20)$$

where we put

$$\begin{aligned} L_{n-1}(f, x) &= f(x_1) + (x - x_1)[x_1, x_2]f + (x - x_1)(x - x_2)[x_1, x_2, x_3]f \\ &\quad + \dots + (x - x_1) \dots (x - x_{n-1})[x_1, \dots, x_n]f, \end{aligned} \quad (1.21)$$

and for the "remainder term"

$$R_{n-1}(f, x) = (x - x_1) \dots (x - x_n)[x, x_1, \dots, x_n]f. \quad (1.22)$$

15. This is the classical *interpolation formula of Newton*, in the case of distinct interpolation points x_ν . For real or not necessarily real x_1, \dots, x_n we can write R_{n-1} correspondingly, by (1.10), (1.11), or (1.12), in one of the forms

$$R_{n-1}(f, x) = (x - x_1) \dots (x - x_n) \frac{f^{(n)}(\xi)}{n!} \quad (1.23a)$$

$$R_{n-1}(f, x) = (x - x_1) \dots (x - x_n) \theta \frac{f^{(n)}(\xi)}{n!} \quad (1.23b)$$

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

$$R_{n-1}(f, x) = \frac{(x - x_1) \dots (x - x_n)}{n!} (f^{(n)}(\xi) + \theta DM_{n+1}), \quad (1.23c)$$

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

The formula (1.20) is an *identity*; we can therefore let groups of interpolation points become equal as in Section 10. Then R_{n-1} becomes, for instance,

$$R_{n-1}(f, x) = (x - t_1)^{m_1} \dots (x - t_k)^{m_k} [x, t_1^{m_1}, \dots, t_k^{m_k}] f. \quad (1.24)$$

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

GENERAL INTERPOLATION PROBLEM

16. In order to make clear the significance of the polynomial $L_{n-1}(f, x)$, observe that by (1.16) $f(x)$ enters into $L_{n-1}(f, x)$ and $R_{n-1}(f, x)$ only with the values

$$f^{(\mu)}(t_\kappa) \quad (\mu = 0, 1, \dots, m_\kappa - 1; \quad \kappa = 1, \dots, k). \quad (1.25)$$

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

$$\gamma^{(\mu)}(t_\kappa) = f^{(\mu)}(t_\kappa) \quad (\mu = 0, \dots, m_\kappa - 1; \quad \kappa = 1, \dots, k) \quad (1.26)$$

we have

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

Suppose now that the function $\gamma(x)$ satisfying (1.26) is a *polynomial of degree not exceeding $n - 1$* . Then, if we apply (1.23) 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 (1.26) is *uniquely determined*.

17. We are now going to prove that *there always exists a polynomial satisfying (1.26)*. We are going to prove even more.

Assuming *arbitrarily* n numbers

$$A_{\kappa}^{(\mu)} \quad (\kappa = 1, \dots, k; \mu = 0, 1, \dots, m_{\kappa} - 1), \quad (1.27)$$

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

$$\gamma^{(\mu)}(t_{\kappa}) = A_{\kappa}^{(\mu)} \quad (\kappa = 1, \dots, k; \mu = 0, 1, \dots, m_{\kappa} - 1). \quad (1.28)$$

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

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

the equations (1.28) represent a set of n linear equations in the unknowns u , 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

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

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

19. It follows now in particular that $L_{n-1}(f, x)$ is the unique polynomial $\gamma(x)$ of degree $< n$ satisfying the conditions (1.26):

$$L_{n-1}^{(\mu)}(f, t_{\kappa}) = f^{(\mu)}(t_{\kappa}) \quad (\kappa = 1, \dots, k; \mu = 0, 1, \dots, m_{\kappa} - 1). \quad (1.29)$$

20. A function $\gamma(t)$ satisfying the conditions (1.28) is generally called an *interpolating function to the system of values $A_{\kappa}^{(\mu)}$* , corresponding to the *interpolation points* (or *interpolation abscissas*) t_{κ} , such 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 (1.26), $\gamma(t)$ 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

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 (1.28) could be also considered for polynomials $\gamma(t)$ not necessarily of degree $< n$ but satisfying some other special or general conditions.

21. 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

$$L_{n-1}(f, x) = \sum_{v=1}^n f(t_v) \frac{F(x)}{(x - t_v) F'(t_v)}. \quad (1.30)$$

THE REMAINDER FOR A GENERAL INTERPOLATING FUNCTION

22. If $\gamma(x)$ is a general interpolating function satisfying the conditions (1.26) we can easily obtain an expression for the “remainder” $f(x) - \gamma(x)$ corresponding to (1.24), 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

$$f(x) - \gamma(x) = R_{n-1}(F, x) = \prod_{v=1}^n (x - x_v) [x, t_v]^{m_v} (f - \gamma) \quad (1.31)$$

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

$$f(x) - \gamma(x) = \prod_{\kappa=1}^k (x - t_\kappa)^{m_\kappa} \frac{f^{(n)}(\xi) - \gamma^{(n)}(\xi)}{n!}, \quad \xi \prec \langle x, t_1, \dots, t_k \rangle. \quad (1.32)$$

TRIANGULAR SCHEMES FOR COMPUTING DIVIDED DIFFERENCES

23. 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 :

$$x_1 \quad f(x_1)$$

$$[x_1, \quad x_2]f$$

$$x_2 \quad f(x_2)$$

$$[x_1, \quad x_2, \quad x_3]f$$

$$[x_2, \quad x_3]f$$

$$[x_1, \quad x_2, \quad x_3, \quad x_4]f$$

$$x_3 \quad f(x_3)$$

$$[x_2, \quad x_3, \quad x_4]f$$

$$[x_3, \quad x_4]f$$

$$x_4 \quad f(x_4)$$

.

$$[x_{m-2}, \quad x_{m-1}, \quad x_m]f$$

.

$$[x_{m-1}, \quad x_m]f$$

$$x_m \quad f(x_m)$$

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.

24. In the confluent case a similar scheme can be used. If we consider, for instance, $[x_1^3, \quad x_2^2, \quad x_3]f$, the corresponding scheme would be as shown on page 14.

x_1	$f(x_1)$						
		$f'(x_1)$					
x_1	$f(x_1)$		$f''(x_1)$				
				$\frac{\partial^2}{\partial x_1^2} [x_1, x_2]f$			
x_1	$f(x_1)$				$\frac{\partial^3}{\partial x_1^2 \partial x_2} [x_1, x_2]f$		
						$\frac{\partial^3}{\partial x_1^2 \partial x_2} [x_1, x_2, x_3]f$	
x_2	$f(x_2)$					$\frac{\partial^2}{\partial x_1 \partial x_2} [x_1, x_2]f$	
							$\frac{\partial^2}{\partial x_1 \partial x_2} [x_1, x_2, x_3]f$
x_2	$f(x_2)$						$\frac{\partial}{\partial x_2} [x_1, x_2, x_3]f$
			$f'(x_2)$				
x_2	$f(x_2)$						$\frac{\partial}{\partial x_2} [x_2, x_3]f$
				$[x_2, x_3]f$			
x_3	$f(x_3)$						

THE CONCEPT OF INVERSE INTERPOLATION

1. If a number a is used as an approximation for a number x , we write $a = x$; 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 $x = \phi(y)$ be the inverse function of $y = f(x)$. Then $\phi(y_v) = x_v$. The problem of finding a zero of $f(x)$ now becomes the problem of evaluating $\phi(0)$.

Let $T(y)$ be an interpolating polynomial $L_{n-1}(\phi, y)$ for $\phi(y)$, so that $T(y_v) = x_v$. We now evaluate $T(0) = \phi(0)$. An estimate of the error involved here may be obtained from (1.24) and (1.10):

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

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 $J_x: a \leqq x \leqq b$. 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 $x \prec J_x$.*

Remark. This property of $f'(x)$ is sometimes incorrectly given as a definition of a continuous function. The assertion of the Theorem 2.1 is, of course, trivial if $f'(x)$ is continuous in 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-sided and the left-sided 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 $f'(x) \neq 0$ ($x \prec J_x$) implies therefore that either $f'(x) > 0$ for all $x \prec J_x$ or $f'(x) < 0$ for all $x \prec J_x$.

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,

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

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

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

Here X_k is a *polynomial* in y' , y'' , \dots , $y^{(k)}$. 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 $\Phi(y)$. We write (2.5) with $k = n$ and get by differentiation, since $dy'/dy = y''/y'$,

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

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_ν instead of $y^{(\nu)}$. X_n is then a polynomial in y_1, y_2, \dots, y_n .

By induction we see from (2.7) that

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

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)

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

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

$$y_1 \sum_{\nu} y_{\nu+1} \frac{\partial}{\partial y_{\nu}} X_n \quad (2.10)$$

and becomes therefore also of the dimension n .

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

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

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

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

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-side 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.

We give in the next section 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 $a \geqslant b$.

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

$$\min f(x) = f(\xi) \quad (x \prec J_x, \quad \xi \prec J_x).$$

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 the differential calculus,

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

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

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

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 $(x_0 - 1/10^3, x_0 + 1/10^3)$. 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 $|f(z_0)| < \eta m$ and, everywhere in K_η , $|f'(z)| > m$. 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 $|a| < m$.*

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 $w = f(z)$ of $w = f(z)$ exists and is uniquely determined. We call $F(w)$ the local inverse of $f(z)$ in z_0 .

In virtue of $w = f(z)$ the path γ is transformed into a curve Γ in the w -plane and $w_0 = f(z_0)$ is contained in 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 $w = ta$ ($0 \leqq t \leqq 1$); $G(w)$ obviously can be continued along the segment $w = ta$ ($0 \leqq t < t_0$) 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 $w = ta$ ($0 \leqq t < t_1$), and we have

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

and

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

We have then

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

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

$$|G(ta)| < t \frac{|a|}{m} < 1 \quad (0 \leq t < t_1),$$

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

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

$$(\gamma) \quad z = z_t \quad (0 \leq t \leq t_1).$$

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 γ 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 $\kappa(z)$ 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 $> t_1$ contrary to the definition of t_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.

* 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 $x \uparrow \zeta$.

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

15. We can use the values of $\phi^{(v)}(y)$ 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

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

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:

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

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 $\phi^{(v)}(y)$ 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 $\Phi^{(n+1)}(\sigma)y'^{n+1}/(n+1)!$ 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*

$$h = \zeta - x = \sum_{v=1}^n \frac{1}{v!} X_v \left(\frac{y'}{y'}, \frac{y''}{y'}, \dots, \frac{y^{(v)}}{y'} \right) k^v + O(f(x)^{n+1}), \quad (2.19)$$

$$k = -\frac{f(x)}{f'(x)} \quad (x \rightarrow \zeta, \quad f(x) \rightarrow 0, \quad f'(\zeta) \neq 0).$$

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

$$h = \zeta - x = k - \frac{1}{2} \frac{y''}{y'} k^2 + \frac{3 y'^2 - y' y^{(3)}}{6 y'^2} k^3$$

$$+ \frac{10 y' y'' y^{(3)} - y'^2 y^{(4)} - 15 y'^3}{24 y'^3} k^4 + \dots \quad (2.20)$$

* 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.

3

Method of False Position (Regula 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 $x_1, x_2 \prec J_x$, $x_1 \neq x_2$, 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 :

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

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 (1.23a) for the remainder with $n = 2$, we have for each $x \prec J_x$

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

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

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

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

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

Since

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

we may rewrite (3.5) as follows:

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

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.

The 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 $x = \phi(y)$, 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 $\phi(0) = \zeta$. We now interpolate the function $\phi(y)$ 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 $\phi(0)$. 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 $\phi(y)$.

4. From (2.2) we get

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

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

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

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

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

and

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

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) in replacing there $f(x_3)$ by $f'(\xi')(x_3 - \xi)$, y_1 by $(x_1 - \xi)f'(\xi_1)$, y_2 by $(x_2 - \xi)f'(\xi_2)$, 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 $\xi' = \xi_3 = \xi$ 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 below. Taking x_0 and x_1 as our initial approximations, we

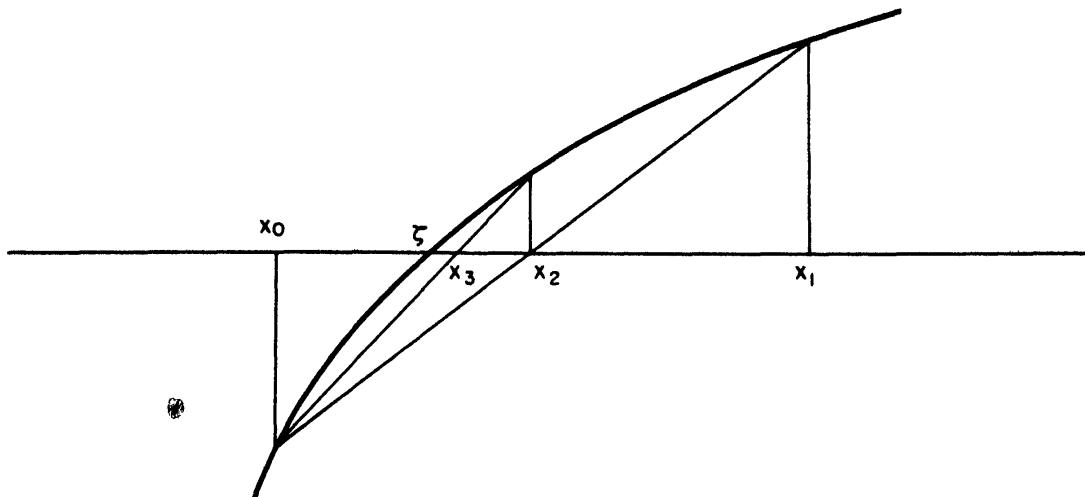


Fig. 1

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

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

Does x_ν ($\nu = 1, 2, \dots$) converge?

7. If the geometric situation is as we have drawn in Fig. 1, then x_ν 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

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

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

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 the 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_3 . Then

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

Put $K|\zeta - x_\nu| = d_\nu$; then we have, applying this repeatedly, $d_2 \leq d_0 d_1, \dots, d_{\nu+1} \leq d_\nu d_{\nu-1}$. 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 $d_\nu \leq d^{\alpha_\nu}$, where $\alpha_0 = 1, \alpha_1 = 1, \dots, \alpha_{\nu+1} = \alpha_\nu + \alpha_{\nu-1}$. We have here the *Fibonacci* sequence defined by

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

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

We try a solution of the form t^ν . Substituting in (3.16), we have $t^{\nu+1} = t^\nu + t^{\nu-1}$. Dividing by $t^{\nu-1}$, 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 $\beta_\nu = c_1 t_1^\nu + c_2 t_2^\nu$, where c_1 and c_2 are arbitrary constants. Now β_ν satisfies (3.16) for

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

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 $\zeta - x_v \rightarrow 0$ then we always have

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

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 volume 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

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

* 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 differential systems used in the method in question, 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.

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 the formula (3.13) is $1.618\dots$ *.

THE ROUNDING-OFF RULE

12. Since in using (3.2) the error of x_3 is $O(y_1y_2)$, 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 $O(y_1y_2)$ 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 (3.2a) that

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

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 we obtain for δ_1 and δ_2 the conditions

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

(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

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

* 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.

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 be 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 $1/f'(\zeta)$, only 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

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

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

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

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

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

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*
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

* 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

$$f'(\zeta) = 11.1614377, \quad f''(\zeta) = 12.56730888, \quad (3.25)$$

$$\frac{f''(\zeta)}{2f'(\zeta)} = 0.5629789.$$

EXAMPLE 2

(1)	(2)	(3)	(4)
v	x_v	$\zeta - x_v$	$\frac{x_{v+1} - \zeta}{(x_v - \zeta)(x_{v-1} - \zeta)}$
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	0.032727	0.565
5	2.0945 49431 75	0.05205	0.5624
6	2.0945 51481 228	0.09314	0.5629 779
7	2.0945 51481 54232 69542 1	0.153627	†

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

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.

4

Iteration

A CONVERGENCE CRITERION FOR AN ITERATION

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

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

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 $x_\nu \rightarrow \zeta_0$ (unless one of the x_ν becomes equal to ζ_0).*

Theorem 4.2. *Let $\psi(x)$ be an iterating function defined in J_x . Assume that $\zeta_0 = \psi(\zeta_0)$, $\zeta_0 \prec (J_x)$, 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*

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

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 $x_1 \prec V(\zeta_0)$ 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 $x_2 \prec V(\zeta_0)$. Applying repeatedly (4.4) to x_2, \dots, x_ν , we get $|x_3 - \zeta_0| \leq p|x_2 - \zeta_0|, \dots,$

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

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

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

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 $V(\zeta_0)$

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

* 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.

Hence for an $x_1 \prec V(\zeta_0)$, 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 , as long as it lies in $V(\zeta_0)$ and is $\neq \zeta_0$. Hence $x \rightarrow \zeta_0$, unless one of the x , 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 $1/10$. If p is greater than $1/2$, 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 insured 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 $\max(|1 - cM|, |1 - cm|)$ 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 $\psi^*(\zeta_0)$ 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 :

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

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

$$x_{n+1} = 0.5 - \log x_n$$

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 $\frac{1}{2}(x_n + x_{n+1})$ 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 $F_c(x) = cF(x)$,

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

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 $F'(\xi) > \frac{1}{2} \max F'(x)$. 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_ν , we get

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

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

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

Then we have also, replacing x here by x_ν ,

$$\overline{\lim}_{\nu \rightarrow \infty} \frac{|x_{\nu+1} - \zeta_0|}{|x_\nu - \zeta_0|^k} < \infty, \quad k > 1. \quad (4.12)$$

10. If we have a sequence x_ν going to ζ_0 and if we have a relation of the type (4.12) then we say that the convergence of the x_ν 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_ν .

On the other hand, if we have the relation $\overline{\lim}|x_{\nu+1} - \zeta_0|/|x_\nu - \zeta_0| < 1$, 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_ν 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 $J_x (\zeta_0 - d < x < \zeta_0 + d)$ 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_ν converging to ζ_0 and the convergence is at least linear.

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

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

and therefore, by (4.13),

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

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

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

and we see that indeed the x_ν 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 $\zeta_0 \pm d$. 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 nonnegative 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 $|x - \zeta_0| < d$. 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 $\leq \min(\rho, \delta)$.

For a general z from K_ε consider

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

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 $v = 1, 2, \dots, n+1$,

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

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

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

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

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

We see that

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

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

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

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

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

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

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

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

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

)

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

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

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

- (a) If $f(\zeta) = 0$, $\zeta \prec J_0$, 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

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

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 the case (α) we have a monotonic *increasing* sequence $x_0 < x_1 \leq x_2 \leq x_3 \leq \dots$ and in the 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, $\zeta_0 \prec J_0$. 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*

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

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 the differential calculus, we have

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

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

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

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 $x \rightarrow \zeta$,

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

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 $\alpha > 0$*

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

* 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.

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

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

7. Proof. We can assume without loss of generality, that A and the values of $f(x)$ in J_0 are positive. Since by (5.7) $f'(\zeta)$ exists, we have

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

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

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

Introduce the positive numbers

$$q_\nu = A |\zeta - x_\nu|^\alpha \quad (5.12)$$

and consider

$$d_\nu = \frac{1}{q_{\nu+1}} - \frac{1}{q_\nu} = \frac{1}{q_{\nu+1}} \left(1 - \frac{q_{\nu+1}}{q_\nu} \right). \quad (5.13)$$

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

$$p_\nu = \frac{f(x)}{x_\nu - \zeta}. \quad (5.14)$$

From the recurrence formula

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

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

$$x_\nu - \zeta > 0. \quad (5.15)$$

From (5.15) follows

$$p_\nu > 0,$$

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

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

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)

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

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

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

converges to α .

Now identify a_v with d_v . Then

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

and since

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

we have

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

$$\frac{1}{v q_{v+1}} \rightarrow \alpha.$$

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

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

and by (5.12)

$$\begin{aligned} \nu |\zeta - x_\nu|^\alpha &\rightarrow \frac{1}{\alpha A}, \\ \nu^{1/\alpha} |\zeta - x_\nu| &\rightarrow \frac{1}{|\alpha A|^{1/\alpha}}, \end{aligned} \quad \text{Q.E.D.} \quad (5.22)$$

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 ν which is at most 1 and $|\zeta - x_\nu|$ 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*:

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

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

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

and

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

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

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

* See, for example, Chapter 8 and Appendix E.

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.

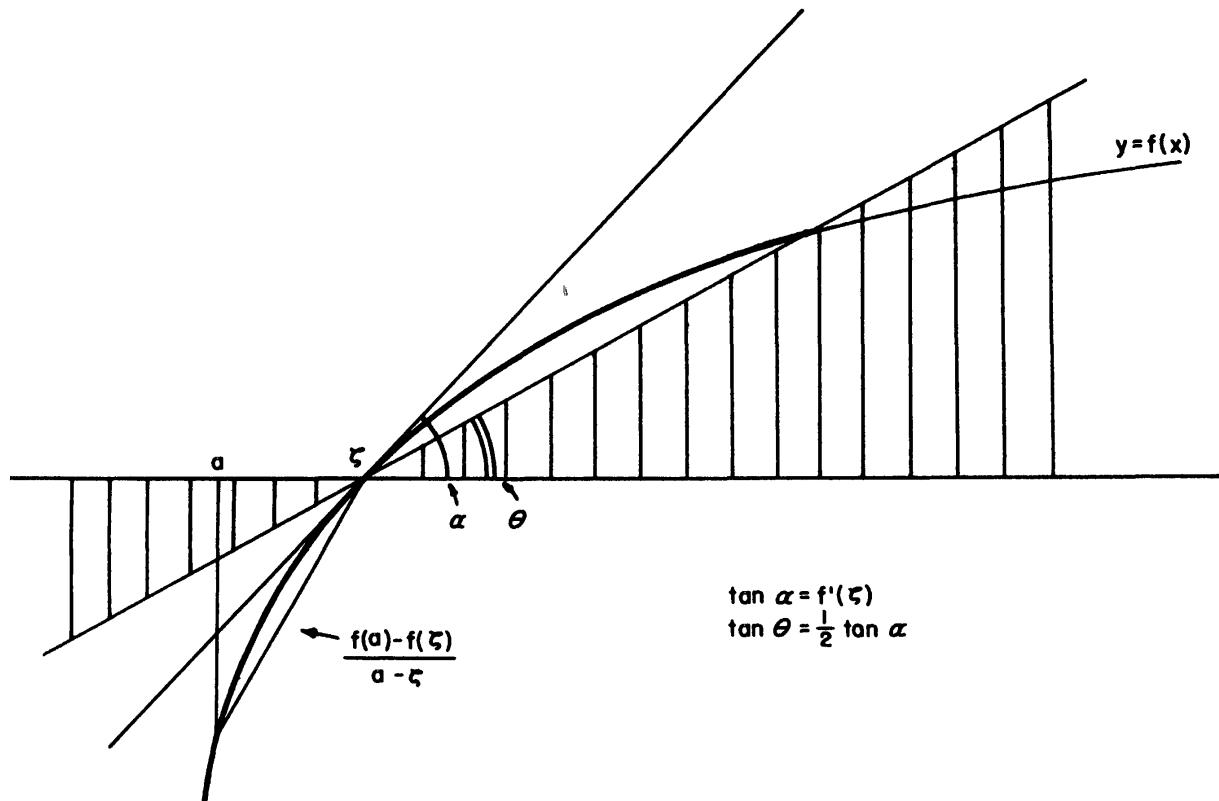


Fig. 2

In Fig. 2, our point $(a, f(a))$ must so be chosen that the chord from it to ζ will not fall within the cross-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 cross-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

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,

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

In this way we obtain a sequence of numbers x_{ν} ($\nu = 0, 1, \dots$). We consider the following problems: Can we conclude from the properties of this sequence that a root exists in a neighborhood of our approximations? Does the sequence tend to this zero, and if so, how good is the approximation by x_{ν} ?

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 (1.32), 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 $\varphi(0) = \zeta$,

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

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 bracket 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 $(x_1, x_1 + k)$. If the sign is negative, then ζ lies in the interval $(x_1 - k, x_1)$.

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 $f(x_0) - f(\zeta) = (x_0 - \zeta)f'(\xi_0)$, $\xi_0 \prec (x_0, \zeta)$. Substituting this in (6.6), we obtain

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

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

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 horner's; therefore, $\sqrt[3]{2} = 1.414\dots$ 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.

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 root 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 that 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 two parallel theorems for a real function of a real variable and an analytic function of a complex variable. The proofs, with the exception of one point (the proof of the uniqueness), are parallel. In the complex case, we replace the interval J_0 : $\langle x_0, x_0 + 2h_0 \rangle$ by the circle K_0 : $|z - z_1| \leq |h_0|$, z, z_1 being complex numbers. We prove our theorem explicitly for a real function $f(x)$ of a real variable (Theorem 7.2) and do not indicate the obvious changes which are necessary in the proof for the complex case (Theorem 7.1), with the exception of the uniqueness proof.

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

3. **Theorem 7.1.** *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_ν by the recurrence formula

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

Then all z_ν lie in K_0 and we have

$$z_\nu \rightarrow \zeta \quad (\nu \rightarrow \infty),$$

where ζ is the only zero in K_0 . Unless it lies on the boundary of K_0 , ζ is a simple zero. Further, we have the relations

$$\frac{|z_{\nu+1} - z_\nu|}{|z_\nu - z_{\nu-1}|^2} \leq \frac{M}{2|f'(z_\nu)|} \quad (\nu = 1, 2, \dots),$$

$$|\zeta - z_{\nu+1}| \leq \frac{M}{2|f'(z_\nu)|} |z_\nu - z_{\nu-1}|^2 \quad (\nu = 1, 2, \dots).$$

4. Theorem 7.2. 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 $\max_{J_0} |f''(x)| = M$ and

\

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

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

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

Then all x_ν lie in J_0 and we have

$$x_\nu \rightarrow \zeta \quad (\nu \rightarrow \infty), \quad (7.2)$$

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

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

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

5. Proof of Theorem 7.2. We have (as $f''(x)$ is integrable, being measurable and bounded)

$$f'(x) - f'(x_0) = \int_{x_0}^x f''(x) dx,$$

$$|f'(x) - f'(x_0)| \leq |x - x_0|M \quad (7.3)$$

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),

$$|f'(x_1)| \geq |f'(x_0)| - |f'(x_1) - f'(x_0)| \geq |f'(x_0)| - \frac{|f'(x_0)|}{2},$$

$$|f'(x_1)| \geq \frac{|f'(x_0)|}{2}. \quad (7.5)$$

Now, integrating by parts, we have

$$\begin{aligned} \int_{x_0}^{x_1} (x_1 - x)f''(x) dx &= -(x_1 - x_0)f'(x_0) + f(x_1) - f(x_0) \\ &= -h_0f'(x_0) + f(x_0) + f(x_1) = f(x_1), \\ f(x_1) &= \int_{x_0}^{x_1} (x_1 - x)f''(x) dx. \end{aligned} \quad (7.6)$$

We introduce here a new variable of integration t , putting $x = x_0 + th_0$,

$$x_1 - x = x_1 - x_0 - th_0 = h_0 - th_0 = h_0(1 - t), \quad dx = h_0 dt:$$

$$f(x_1) = h_0^{-1} \int_0^1 (1 - t)f''(x_0 + th_0) dt.$$

6. It now follows, since $1 - t \geq 0$, that

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

and

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

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

$$|h_1| \leq \frac{|h_0|^2 M}{|f'(x_0)|}, \quad (7.8)$$

$$\frac{M|h_1|}{|f'(x_1)|} \leq \frac{|h_0|^2 M^2}{|f'(x_0)| |f'(x_1)|} \leq \frac{|h_0|^2 M^2}{|f'(x_0)| \frac{1}{2} |f'(x_0)|},$$

$$\frac{2M|h_1|}{|f'(x_1)|} \leq \frac{2^2 |h_0|^2 M^2}{|f'(x_0)|^2} = \left(\frac{2|h_0|M}{|f'(x_0)|} \right)^2.$$

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 $J_1: \langle x_1, x_1 + 2h_1 \rangle$ lies in J_0 .

7. 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_ν and h_ν , respectively, and consequently the hypotheses remain true for x_ν and h_ν ($\nu = 0, 1, \dots$).

Let J_ν ($\nu = 0, 1, \dots$) be the interval $\langle x_\nu, x_\nu + 2h_\nu \rangle$. We have a sequence of intervals $J_{\nu+1} \prec J_\nu$ ($J_{\nu+1}$ is contained in J_ν) with the radius of $J_{\nu+1}$ (the radius of an interval is half its length) at the most equal to one-half the radius of J_ν . 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 :

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

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

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

8. 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 $|x - x_0| < 2|h_0|$, 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). Therefore ζ is a simple zero of $f(x)$ if it lies in (J_0).

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

9. *In the complex case* we must proceed differently to prove that in this case too, ζ is the only zero of $f(z)$ in K_0 .

Let ζ^* be a zero of $f(z)$ inside K_0 . Integrating by parts, we have

$$\begin{aligned} \int_{z_\nu}^{\zeta^*} (z - \zeta^*)f''(z) dz &= \left[(z - \zeta^*)f'(z) \right]_{z_\nu}^{\zeta^*} - \int_{z_\nu}^{\zeta^*} f'(z) dz \\ &= f'(z_\nu) \left[\zeta^* - z_\nu + \frac{f(z_\nu)}{f'(z_\nu)} \right] = f'(z_\nu)(\zeta^* - z_{\nu+1}), \\ f'(z_\nu)(\zeta^* - z_{\nu+1}) &= \int_{z_\nu}^{\zeta^*} (z - \zeta^*)f''(z) dz. \end{aligned}$$

In this integral introduce $z = z_\nu + t(\zeta^* - z_\nu)$. Then

$$z - \zeta^* = z_\nu - \zeta^* + t(\zeta^* - z_\nu) = (z_\nu - \zeta^*)(1 - t),$$

$$f'(z_\nu)(\zeta^* - z_{\nu+1}) = -(z_\nu - \zeta^*)^2 \int_0^1 (1-t)f''(z_\nu + t(\zeta^* - z_\nu)) dt,$$

$$|f'(z_\nu)| |\zeta^* - z_{\nu+1}| \leq \frac{1}{2} |\zeta^* - z_\nu|^2 M,$$

$$2 \frac{|\zeta^* - z_{\nu+1}|}{|\zeta^* - z_\nu|} \leq |\zeta^* - z_\nu| \frac{M}{|f'(z_\nu)|}.$$

We shall now prove that the expression on the left side tends to zero and hence $z_\nu \rightarrow \zeta^*$.

10. We introduce

$$q_\nu = 2 \frac{|\zeta^* - z_{\nu+1}|}{|\zeta^* - z_\nu|}. \quad (7.15)$$

Then

$$q_n \leq \frac{|\zeta^* - z_n|M}{|f'(z_n)|}. \quad (7.16)$$

We have from (7.15)

$$\left(\prod_{\nu=0}^{n-1} q_\nu \right) \frac{|\zeta^* - z_0|}{|h_0|} \frac{|h_0|M}{2^n |f'(z_n)|} = \frac{|\zeta^* - z_n|M}{|f'(z_n)|}.$$

(For $n = 0$ the product in parentheses must be considered as *one*.) Indeed, the left-hand side is

$$\left(2^n \frac{|\zeta^* - z_n|}{|\zeta^* - z_0|} \right) \frac{|\zeta^* - z_0|M}{2^n |f'(z_n)|} = \frac{|\zeta^* - z_n|M}{|f'(z_n)|}.$$

Therefore, by (7.16)

$$q_n \leq \left(\prod_{\nu=0}^{n-1} q_\nu \right) \frac{|\zeta^* - z_0|}{|h_0|} \left[\frac{|h_0|M}{2^n |f'(z_n)|} \right]. \quad (7.17)$$

From (7.5) we have $|f'(z_1)| \geq \frac{1}{2} |f'(z_0)|$. Since our assumptions hold in each point z_ν ($\nu = 0, 1, \dots$), we have generally

$$2|f'(z_{\nu+1})| \geq |f'(z_\nu)|.$$

Using this result repeatedly, we obtain

$$\begin{aligned} 2^{n-1}[2|f'(z_n)|] &\geq 2^{n-1}|f'(z_{n-1})| \geq \dots \geq 2|f'(z_1)| \geq |f'(z_0)|, \\ 2^n|f'(z_n)| &\geq |f'(z_0)|, \end{aligned}$$

and from (7.17)

$$q_n \leq \left(\prod_{\nu=0}^{n-1} q_\nu \right) \frac{|\zeta^* - z_0|}{|h_0|} \left[\frac{|h_0|M}{|f'(z_0)|} \right].$$

By hypothesis, the expression in brackets is $\leq \frac{1}{2}$ and we get

$$q_n \leq \left(\prod_{\nu=0}^{n-1} q_\nu \right) \frac{1}{2} \frac{|\zeta^* - z_0|}{|h_0|}. \quad (7.18)$$

Now the distance from ζ^* to z_0 is at most equal to the diameter of K_0 , i.e., $|\zeta^* - z_0| \leq 2|h_0|$. We obtain therefore from (7.18) finally

$$q_n \leq \prod_{\nu=0}^{n-1} q_\nu. \quad (7.19)$$

For $n = 0$, the right-hand side of (7.19) is equal to one. Hence $q_0 \leq 1$ and by induction we see that generally $q_n \leq 1$. From (7.15) we now have

$$|\zeta^* - z_{\nu+1}| \leq \frac{1}{2} |\zeta^* - z_\nu| \leq \dots \leq \frac{1}{2^{\nu+1}} |\zeta^* - z_0|, \quad z_\nu \rightarrow \zeta^* (\nu \rightarrow \infty).$$

11. The assertions (a) and (b) of Theorem 7.2 can now be easily deduced: (a) is equivalent to

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

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

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

and by (7.11)

$$|h_\nu| \leq \frac{M|h_{\nu-1}|^2}{2|f'(x_\nu)|},$$

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 theorems are completely proved.

In the case $\zeta = x_0 + 2h_0$ it can be shown easily that $f(x)$ is a quadratic polynomial with ζ as a double zero.

It is now 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 certainly will 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

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 root of *exact multiplicity p*. We then replace the Newton-Raphson formula by

$$x_{v+1} = x_v - p \frac{f(x_v)}{f'(x_v)} \quad (v = 0, 1, 2, \dots). \dagger \quad (8.1)$$

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

Theorem 8.1. *Let $f(x)$ have a root ζ of exact multiplicity p , and assume that $f^{(p+1)}(x)$ exists and is continuous in a neighborhood of ζ . Compute the sequence x_v ($v = 1, 2, \dots$) by (8.1). Then, if x_1 is sufficiently close to ζ , all x_v exist and we have $x_v \rightarrow \zeta$ and*

$$\frac{\zeta - x_{v+1}}{(\zeta - x_v)^2} \rightarrow - \frac{f^{(p+1)}(\zeta)}{p(p+1)f^{(p)}(\zeta)} \quad (v \rightarrow \infty). \quad (8.2)$$

2. Notice that if $p = 1$, we have the familiar result (6.9) for the Newton-Raphson formula.

Proof of Theorem 8.1. We have from (8.1)

$$\begin{aligned} \zeta - x_{v+1} &= \zeta - x_v + p \frac{f(x_v)}{f'(x_v)}, \\ (\zeta - x_{v+1})f'(x_v) &= pf(x_v) + (\zeta - x_v)f''(x_v) = F(x_v), \end{aligned} \quad (8.3)$$

If we define $F(x)$ by

$$F(x) = pf(x) - (x - \zeta)f'(x). \quad (8.4)$$

* Numbers in square brackets refer to the bibliography on p. 327.

† Equation (8.1) is obtained applying (6.8) to $f(x)^{1/p}$.

In our proof we shall use the familiar Taylor expansion

$$\Phi(x) = \sum_{\nu=0}^n \frac{(x-a)^\nu}{\nu!} \Phi^{(\nu)}(a) + R_{n+1}, \quad (8.5)$$

where

$$R_{n+1} = \frac{(x-a)^{n+1}}{(n+1)!} \Phi^{(n+1)}(\xi), \quad \xi \prec (a, x) \quad (8.6)$$

or

$$R_{n+1} = \int_a^x \frac{(x-t)^n}{n!} \Phi^{(n+1)}(t) dt. \quad (8.7)$$

We will also use Leibniz' formula for the n th derivative of a product of two functions:

$$(uv)^{(n)} = uv^{(n)} + \binom{n}{1} u' v^{(n-1)} + \dots \quad (8.8)$$

In our application we have $u = x - \zeta$ and, since in the terms omitted in (8.8) u is differentiated at least twice, they will all vanish. From (8.4) and (8.8)

$$F^{(\nu)}(x) = (\phi - \nu) f^{(\nu)}(x) - (x - \zeta) f^{(\nu+1)}(x), \quad (8.9)$$

$$F^{(\nu)}(\zeta) = 0 \quad (\nu = 0, 1, \dots, \phi), \quad F^{(\phi)}(x) = (\zeta - x) f^{(\phi+1)}(x). \quad (8.10)$$

3. From (8.5) applied to $f'(x)$, we have

$$f'(x) = \sum_{\nu=0}^{\phi-2} \frac{(x-\zeta)^\nu}{\nu!} f^{(\nu+1)}(\zeta) + R_{\phi-1} = R_{\phi-1}$$

and by (8.6)

$$f'(x_\nu) = \frac{(x_\nu - \zeta)^{\phi-1}}{(\phi-1)!} f^{(\phi)}(\xi_1), \quad \xi_1 \prec (x_\nu, \zeta). \quad (8.11)$$

By (8.5) and (8.10), it follows further that

$$F(x) = \sum_{\nu=0}^{\phi-1} \frac{(x-\zeta)^\nu}{\nu!} F^{(\nu)}(\zeta) + R_\phi = R_\phi$$

and by (8.7)

$$F(x) = \frac{1}{(p-1)!} \int_{\zeta}^x (x-t)^{p-1} F^{(p)}(t) dt;$$

therefore by (8.9)

$$(p-1)! F(x) = - \int_{\zeta}^x [(x-t)^{p-1} (t-\zeta)] f^{(p+1)}(t) dt.$$

4. The bracketed expression in the integrand does not change its sign in the interval of integration. Applying the generalized mean value theorem of the integral calculus, we obtain

$$(p-1)! F(x) = - f^{(p+1)}(\xi_2) \int_{\zeta}^x (x-t)^{p-1} (t-\zeta) dt, \quad \xi_2 \prec (x, \zeta).$$

Integrating by parts, we have

$$\begin{aligned} \int_{\zeta}^x (x-t)^{p-1} (t-\zeta) dt &= - \left[\frac{(x-t)^p (t-\zeta)}{p} \right]_{\zeta}^x + \int_{\zeta}^x \frac{(x-t)^p}{p} dt = \frac{(x-\zeta)^{p+1}}{p(p+1)}, \\ F(x) &= - \frac{f^{(p+1)}(\xi_2)(x-\zeta)^{p+1}}{p(p+1)(p-1)!}. \end{aligned}$$

Using (8.3) and (8.11), we finally get

$$\zeta - x_{p+1} = - \frac{(\zeta - x_p)^2 f^{(p+1)}(\xi_2)}{p(p+1)f^{(p)}(\xi_1)}, \quad \xi_1, \xi_2 \prec (x_p, \zeta). \quad (8.12)$$

Now if $x \rightarrow \zeta$, we have $\xi_1 \rightarrow \zeta$, $\xi_2 \rightarrow \zeta$ and obtain (8.2) immediately.

If we now put $\sup_{(x_p, \zeta)} |f^{(p+1)}(x)| = M_{p+1}$, $\inf_{(x_p, \zeta)} |f^{(p)}(x)| = m_p$ we have from (8.12)

$$|\zeta - x_{p+1}| \leq \frac{(\zeta - x_p)^2}{p(p+1)} \frac{M_{p+1}}{m_p}. \quad (8.13)$$

Remark. In many cases it may be difficult to obtain m_p . One can usually obtain a good estimate of this quantity if one knows M_{p+1} and the p th derivative at one point x_0 . Then an estimate for m_p is easily obtained from

$$|f^{(p)}(x)| \geq |f^{(p)}(x_0)| - M_{p+1} |x - x_0|. \quad (8.14)$$

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., $\zeta < J_0$, where $f(\zeta) = 0$, $f'(x)f''(x) \neq 0$ for $x < J_0$, $f(x_0)f(y_0) < 0$, and $f(x_0)f''(x_0) > 0$. Assume without loss of generality

$$f(x_0) < 0 < f(y_0), \quad f'(x) > 0, \quad f''(x) < 0 \quad (x < J_0) \quad (9.0)$$

(see Fig. 3).

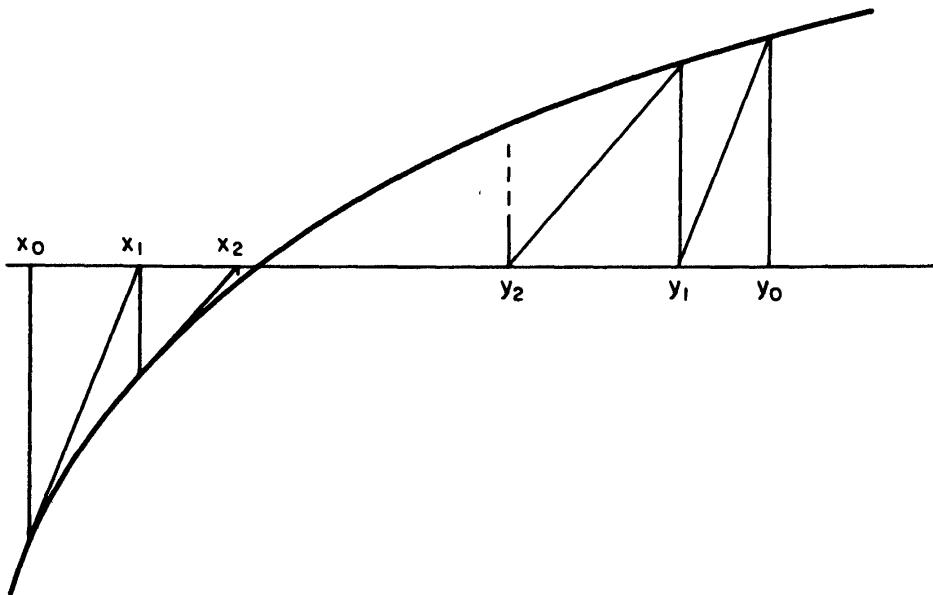


Fig. 3.

The sequence x_ν ($\nu = 0, 1, \dots$) is defined by the Newton-Raphson formula (6.3). The sequence y_ν , introduced by Fourier, is defined by

$$y_{\nu+1} = y_\nu - \frac{f(y_\nu)}{f'(x_\nu)} \quad (\nu = 0, 1, \dots). \quad (9.1)$$

Notice that $y_{\nu+1}$ will be the point of intersection of the line through the point $[y_\nu, f(y_\nu)]$ with the slope $f'(x_\nu)$, 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 ,

$$-\frac{f(\xi)}{f'(\xi)} = 0, \quad -\frac{f(\eta)}{f'(\xi)} = 0, \quad \xi = \eta = \zeta,$$

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

$$f'[y_v + t(\zeta - y_v)] - f'(x_v) = \int_0^{y_v - x_v + t(\zeta - y_v)} f''(x_v + w) dw. \quad (9.10)$$

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 $[(1 + \lambda_\nu)/\lambda_\nu] - t > 0$ for $0 \leq t \leq 1$, we can apply the generalized mean value theorem of the integral calculus to (9.14) and obtain

$$\frac{\zeta - y_{\nu+1}}{(\zeta - y_\nu)^2} = \frac{f''(\xi)}{f'(x_\nu)} \int_0^1 \left(\frac{1 + \lambda_\nu}{\lambda_\nu} - t \right) du dt, \quad \xi \prec (x_\nu, y_\nu), \quad (9.15)$$

$$\frac{\zeta - y_{\nu+1}}{(\zeta - y_\nu)^2} = \frac{f''(\xi)}{f'(x_\nu)} \left(\frac{1 + \lambda_\nu}{\lambda_\nu} - \frac{1}{2} \right) = \frac{1}{2} \frac{f''(\xi)}{f'(x_\nu)} \left(1 + \frac{2}{\lambda_\nu} \right). \quad (9.16)$$

By our assumptions

$$\kappa = -\frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)} \quad (9.17)$$

is $\neq 0$; then by (9.16)

$$\frac{\lambda_\nu}{2 + \lambda_\nu} \frac{\zeta - y_{\nu+1}}{(\zeta - y_\nu)^2} \rightarrow -\kappa \quad (\nu \rightarrow \infty). \quad (9.18)$$

From (6.9) we have

$$\frac{\zeta - x_{\nu+1}}{(\zeta - x_\nu)^2} \rightarrow \kappa \quad (\nu \rightarrow \infty), \quad (9.19)$$

and dividing (9.18) by (9.19) gives

$$\frac{-(\zeta - y_{\nu+1})/(\zeta - x_{\nu+1})}{[(\zeta - y_\nu)/(\zeta - x_\nu)]^2} \sim 1 + \frac{2}{\lambda_\nu} \quad (\nu \rightarrow \infty). \quad (9.20)$$

But here the left-hand expression is by (9.4) equal to $\lambda_{\nu+1}/\lambda_\nu^2$. Therefore, we get

$$\frac{\lambda_{\nu+1}/\lambda_\nu}{2 + \lambda_\nu} \rightarrow 1 \quad (\nu \rightarrow \infty). \quad (9.21)$$

By (9.21) we have in any case $\underline{\lambda_{\nu+1}/\lambda_\nu} \geq 2$ and therefore

$$\lambda_\nu \rightarrow \infty. \quad (9.22)$$

Hence from (9.21)

$$\frac{\lambda_{\nu+1}}{\lambda_\nu^2} \rightarrow 1. \quad (9.23)$$

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

$$\frac{y_{v+1} - x_{v+1}}{(y_v - x_v)^2} \rightarrow \kappa = -\frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)}, \quad \text{Q.E.D.}$$

1. We assume as in Chapter 9 replacing there y_0 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_ν ($\nu = 0, 1, \dots$) is defined by the Newton-Raphson formula (6.3) and the sequence z_ν is obtained by applying the *regula falsi* to $x_{\nu-1}$ and $z_{\nu-1}$; i.e.,

$$z_{\nu+1} = z_\nu - f(z_\nu) \frac{z_\nu - x_\nu}{f(z_\nu) - f(x_\nu)} \quad (\nu = 0, 1, \dots). \quad (10.1)$$

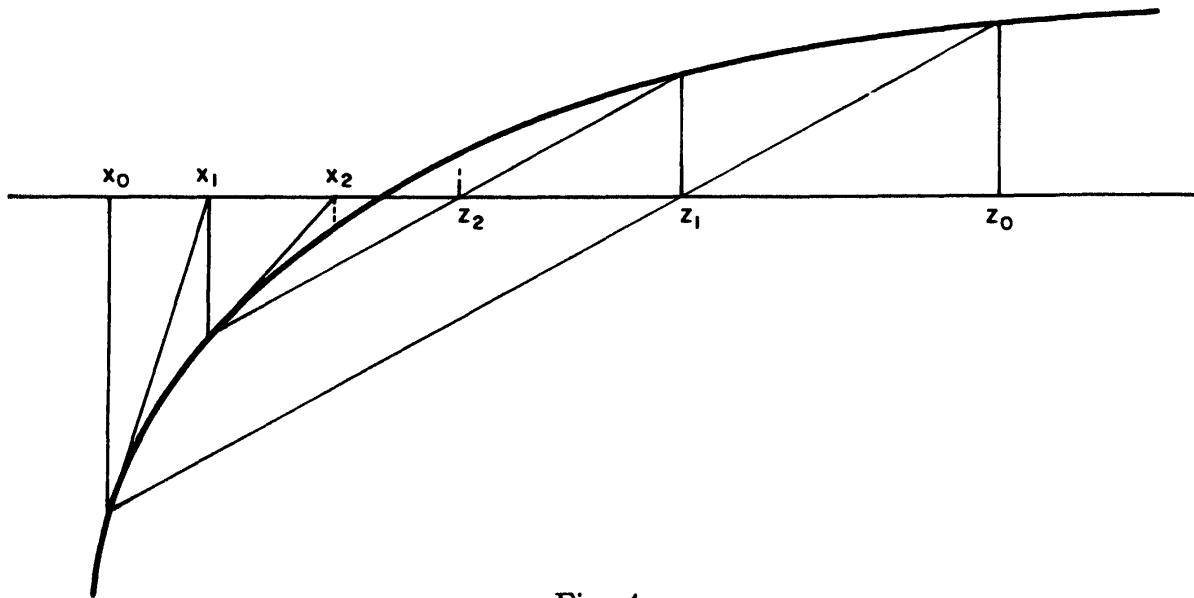


Fig. 4.

We see from Fig. 4 immediately that the z_ν decrease monotonically to a certain number η and it follows now from (10.1) that

$$\frac{f(z_\nu)}{[f(z_\nu) - f(x_\nu)]/(z_\nu - x_\nu)}$$

tends to zero. But here the denominator remains between two positive bounds $[\min f'(x), \max f'(x)]$, 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

$$\begin{aligned} \frac{z_{v+1} - \zeta}{z_v - \zeta} &= \frac{N}{D}, \\ N &= \frac{f(x_v) - f(z_v)}{x_v - z_v} - \frac{f(\zeta) - f(z_v)}{\zeta - z_v}, \quad D = \frac{f(z_v) - f(x_v)}{z_v - x_v}. \end{aligned} \quad (10.4)$$

The 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

$$\begin{aligned} hS &= \int_0^1 \int_{x_v}^{x_v + th} f''(y) dy dt = \int_0^1 [f'(x_v + th) - f'(x_v)] dt, \\ h^2S &= \int_0^1 f'(x_v + th) h dt - hf'(x_v). \end{aligned}$$

Since the integrand is $d[f(x_v + th)]/dt$, and $x_v + h = \zeta$, we have

$$\begin{aligned} h^2S &= f(\zeta) - f(x_v) - hf'(x_v) = -[f(x_v) + hf'(x_v)], \\ -\frac{h^2S}{f'(x_v)} &= \frac{f(x_v)}{f'(x_v)} + h, \\ -\frac{h^2S}{f'(x_v)} &= x_v - x_{v+1} + \zeta - x_v = \zeta - x_{v+1}, \\ -\frac{S}{f'(x_v)} &= \frac{\zeta - x_{v+1}}{(\zeta - x_v)^2}. \end{aligned} \quad (10.11)$$

Dividing (10.9) by (10.11), we have from (10.2)

$$\frac{(z_{\nu+1} - \zeta)/(\zeta - x_{\nu+1})}{(z_{\nu} - \zeta)/(\zeta - x_{\nu})} = -\frac{N^*}{S} \frac{f'(x_{\nu})}{D} = \frac{\mu_{\nu+1}}{\mu_{\nu}}. \quad (10.12)$$

4. We will have to show that μ_{ν} 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 $\sum_{\nu=0}^{\infty} |c_{\nu} - 1|$ converges and no c_{ν} is $= 0$, then $\prod_{\nu=0}^{\infty} c_{\nu}$ converges to a limit $\neq 0$.

Theorem B. If the factors c_{ν} are continuous functions of a point and if $\sum_{\nu=0}^{\infty} |c_{\nu} - 1|$ converges uniformly for all points in a domain, then $\prod_{\nu=0}^{\infty} c_{\nu}$ converges to a continuous function in this domain.

We write now

$$\mu_n = \mu_0 \prod_{\nu=0}^{n-1} \frac{\mu_{\nu+1}}{\mu_{\nu}} = \mu_0 \frac{\prod_{\nu=0}^{n-1} Q_{\nu}}{\prod_{\nu=0}^{n-1} q_{\nu}}, \quad (10.13)$$

where by (10.12)

$$Q_{\nu} = -\frac{N^*}{S}, \quad q_{\nu} = \frac{D}{f'(x_{\nu})} \quad (\nu = 0, 1, \dots). \quad (10.14)$$

5. Consider first

$$Q_{\nu} - 1 = \frac{-N^* - S}{S}. \quad (10.15)$$

From (10.8), (10.9), and (10.10) follows

$$-N^* - S$$

$$= \int_0^1 \int_0^1 \{ f''[z_{\nu} + t(x_{\nu} - z_{\nu}) + ut(\zeta - x_{\nu})] - f''[x_{\nu} + ut(\zeta - x_{\nu})] \} t dt du.$$

Applying (10.5) to the integrand, we have

$$-N^* - S$$

$$= (z_{\nu} - x_{\nu}) \int_0^1 \int_0^1 \int_0^1 t(1-t)f^{(3)}[x_{\nu} + ut(\zeta - x_{\nu}) + w(1-t)(z_{\nu} - x_{\nu})] dw dt du.$$

Since $t(1-t)$ is nonnegative in $\langle 0, 1 \rangle$, we apply the generalized mean value theorem of the integral calculus and obtain

$$-N^* - S = (z_v - x_v) f^{(3)}(\xi) \int_0^1 \int_0^1 \int_0^1 (1-t)t dt dw du, \quad \xi \prec (x_v, z_v).$$

Integrating, we have

$$-N^* - S = \frac{f^{(3)}(\xi)}{6} (z_v - x_v), \quad \xi \prec (x_v, z_v). \quad (10.16)$$

From (10.10) the generalized mean value theorem leads to

$$S = f''(\eta) \int_0^1 \int_0^1 t dt du, \quad S = \frac{f''(\eta)}{2}, \quad \eta \prec (x_v, z_v). \quad (10.17)$$

From (10.15), (10.16), and (10.17) follows

$$Q_v - 1 = \frac{z_v - x_v}{3} \frac{f^{(3)}(\xi)}{f''(\eta)}, \quad \xi, \eta \prec (x_v, z_v). \quad (10.18)$$

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

$$\begin{aligned} D - f'(x_v) &= (z_v - x_v) \int_0^1 \int_0^1 f'' [x_v + ut(z_v - x_v)] t dt du \\ &= (z_v - x_v) f''(\xi_1) \int_0^1 \int_0^1 t dt du, \quad \xi_1 \prec (x_v, z_v). \end{aligned}$$

$$D - f'(x_v) = \frac{z_v - x_v}{2} f''(\xi_1), \quad \xi_1 \prec (x_v, z_v),$$

and from (10.19)

$$q_\nu - 1 = \frac{z_\nu - x_\nu}{2} \frac{f''(\xi_1)}{f'(x_\nu)}, \quad \xi_1 < (x_\nu, z_\nu). \quad (10.21)$$

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

$$|Q_\nu - 1| \leq |z_\nu - x_\nu|K, \quad |q_\nu - 1| \leq |z_\nu - x_\nu|K, \quad (10.23)$$

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

$$\frac{\zeta - x_{\nu+1}}{(\zeta - x_\nu)^2} \rightarrow \kappa, \quad \frac{\zeta - x_{\nu+1}}{\zeta - x_\nu} \sim \kappa(\zeta - x_\nu) \rightarrow 0.$$

Hence $\sum_{\nu=0}^{\infty} |\zeta - x_\nu|$ is convergent.

Similarly, from (3.11) and (9.17) we have

$$\frac{\zeta - z_{\nu+1}}{(\zeta - z_\nu)(\zeta - x_\nu)} \rightarrow \kappa, \quad \frac{\zeta - z_{\nu+1}}{\zeta - z_\nu} \sim \kappa(\zeta - x_\nu) \rightarrow 0,$$

and $\sum_{\nu=0}^{\infty} |\zeta - z_\nu|$ is convergent. From

$$z_\nu - x_\nu = z_\nu - \zeta + \zeta - x_\nu, \quad |z_\nu - x_\nu| \leq |\zeta - x_\nu| + |\zeta - z_\nu|$$

the convergence of $\sum_{\nu=0}^{\infty} |z_\nu - x_\nu|$ now follows.

9. We now show that $\sum_{\nu=0}^{\infty} |z_\nu - x_\nu|$ 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_ν do not depend on z_ν . We will keep x_0 fixed and discuss what happens as $z_0 \downarrow \zeta$. Now all z_ν are strictly increasing functions of z_0 . The initial value of z_0 is $y_0 = z_0'$; denote corresponding values of z_ν by z_ν' . Then

$$z_\nu \leq z_\nu' \quad (\nu = 0, 1, \dots),$$

and hence

$$|z_\nu - x_\nu| \leq |z_\nu' - x_\nu|.$$

Therefore by (10.23)

$$|Q_\nu - 1| \leq K |z_\nu' - x_\nu|, \quad |q_\nu - 1| \leq K |z_\nu' - x_\nu|.$$

This holds for $\zeta < z_0 \leq y_0$. Hence $\sum_{\nu=0}^{\infty} |Q_\nu - 1|$, $\sum_{\nu=0}^{\infty} |q_\nu - 1|$ converge *uniformly* if z_0 varies between ζ and y_0 . We see by Theorem B that both products $\prod_{\nu=0}^{\infty} Q_\nu$ and $\prod_{\nu=0}^{\infty} q_\nu$ converge uniformly for these z_0 , and from (10.13) it follows now that $\lim_{\nu \rightarrow \infty} \mu_\nu$ exists uniformly for all z_0 between ζ and y_0 . Now the z_ν are rational and continuous functions of z_0 . By (10.2) the μ_ν are also continuous functions of z_0 and we see that $A(z_0)$ is a continuous function of z_0 for $\zeta < z_0 \leq y_0$.

10. If $z_0 = \zeta$, then each $z_\nu = \zeta$ ($\nu = 0, 1, \dots$) and it follows that for $z_0 \downarrow \zeta$ each $\mu_\nu(z_0)$ tends to 0. As each $\mu_\nu(z_0)$ is a monotonically increasing function of z_0 , so is $A(z_0)$ in the interval $\zeta < z_0 \leq y_0$. Therefore, as $A(z_0) \geq 0$, it tends to limit A_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

$$|A(z_0) - \mu_k(z_0)| \leq \varepsilon \quad (\zeta < z_0 \leq y_0).$$

We have therefore

$$A(z_0) \leq \varepsilon + \mu_k(z_0)$$

and, as $z_0 \downarrow \zeta$, $A_0 \leq \varepsilon$. We see that $A_0 = 0$, i.e., $A(z_0) \downarrow 0$ ($z_0 \downarrow \zeta$). It follows that the error from the z_ν side can be made an arbitrarily small part of the error from the x_ν side if we start with a sufficiently close point z_0 .

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*

$$x_\nu \prec J_x, \quad f(x_\nu) = f_\nu \quad (\nu = 0, 1, 2). \quad (11.1)$$

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 $w(x_0) = f_0$, $w(x_1) = f_1$. In 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 $\phi(w)$ given by

$$\frac{w - f_1}{w - f_0} = \Delta^* \frac{\Phi - x_1}{\Phi - x_0}. \quad (11.9)$$

Solving (11.9), we have

$$\Phi(w) = \frac{(x_0 - x_1 \Delta^*)w + x_1 f_0 \Delta^* - x_0 f_1}{(1 - \Delta^*)w + f_0 \Delta^* - f_1}. \quad (11.10)$$

It is easily verified that generally for constant $\alpha, \beta, \gamma, \delta$

$$\left(\frac{\alpha w + \beta}{\gamma w + \delta} \right)^{(8)} = 6\gamma^2 \frac{\alpha \delta - \beta \gamma}{(\gamma w + \delta)^4}. \quad (11.11)$$

From (11.10) and (11.11) it follows that

$$\Phi^{(3)}(w) = 6(1 - \Delta^*)^2 \frac{\Delta^*(f_1 - f_0)(x_1 - x_0)}{N^4}, \quad (11.12)$$

where

$$N = (1 - \Delta^*)w + f_0\Delta^* - f_1. \quad (11.13)$$

ERROR ESTIMATES

4. Let $\varphi(w) = \Phi(w)$. If we apply (1.32), our interpolating function is $\Phi(w)$; consequently the factor $f^{(n)}(\xi) - \gamma^{(n)}(\xi)$ in (1.32) must now be replaced by $[\Phi(\eta) - \varphi(\eta)]^{(n)}$. Then we have

$$\Phi(f) - \varphi(f) = \frac{1}{6}(f - f_0)^2(f - f_1)[\Phi^{(3)}(\eta) - \varphi^{(3)}(\eta)], \quad \eta \prec (f, f_0, f_1). \quad (11.14)$$

For $f = 0$, denoting by x_2 the new approximation $\varphi(0)$ to our root ζ , we have from (11.14)

$$\zeta - x_2 = -\frac{f_0^2 f_1}{6} [\Phi^{(3)}(\eta) - \varphi^{(3)}(\eta)], \quad \eta \prec (0, f_0, f_1). \quad (11.15)$$

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 $\frac{1}{2}(x_1 - x_0)^2 f''(\xi_1)$, $\xi_1 \prec (x_0, x_1)$, and we have

$$\frac{\Delta^* - 1}{x_1 - x_0} = \frac{1}{2} \frac{f''(\xi_1)}{f'(x_0)}, \quad \xi_1 \prec (x_0, x_1). \quad (11.19)$$

If $x_0, x_1 \rightarrow \zeta$, then

$$\frac{\Delta^* - 1}{x_1 - x_0} \rightarrow \frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)}. \quad (11.20)$$

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),

$$\Phi^{(3)}(\eta) = \frac{6\Delta^*[(\Delta^* - 1)/(x_1 - x_0)]^2(f_1 - f_0)/(x_1 - x_0)}{[N/(x_1 - x_0)]^4} \rightarrow \frac{3}{2} \frac{f''(\zeta)^2 f'(\zeta)}{f'(\zeta)^2 f'(\zeta)^4}, \quad (11.24)$$

$$\Phi^{(3)}(\eta) \rightarrow \frac{3}{2} \frac{f''(\zeta)^2}{f'(\zeta)^5}. \quad (11.25)$$

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

$$f(x_0) = f(x_0) - f(\zeta) = f'(\xi_0)(x_0 - \zeta), \quad \xi_0 \prec (x_0, \zeta), \quad (11.28)$$

$$f(x_1) = f'(\xi_1)(x_1 - \zeta), \quad \xi_1 \prec (x_1, \zeta). \quad (11.29)$$

Substituting these results in (11.15), we obtain

$$\frac{\zeta - x_2}{(x_0 - \zeta)^2(x_1 - \zeta)} \rightarrow \frac{1}{6} f^{(3)}(\zeta) f'(\zeta)^{-1} - \frac{1}{4} f''(\zeta)^2 f'(\zeta)^{-2}. \quad (11.30)$$

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 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	0.0 ₂ 6045	1977	4011	2994	3	$1.4937 \cdot 10^{-3}$	2.759
3	0.0 ₂ 4551	4320	3811	0802	6	$4.95 \cdot 10^{-8}$	2.585
4	0.0 ₂ 4551	4815	4232	6592	3	$9 \cdot 10^{-19}$	2.47
5	0.0 ₂ 4551	4815	4232	6591	4	—	—

* Computed by Mr. Allen Reiter in the Mathematics Research Center of U.S.A., Madison, Wisconsin.

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

$$a_0 z_{\mu+n} + a_1 z_{\mu+n-1} + \dots + a_n z_\mu = k_{\mu+n} \quad (a_0 = 1; \mu = 0, 1, \dots), \quad (12.1)$$

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 $z_\mu (\mu \geq 0)$ 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 :

$$a_0 y_{\mu+n} + a_1 y_{\mu+n-1} + \dots + a_n y_\mu = 0 \quad (a_0 = 1; \mu = 0, 1, \dots). \quad (12.2)$$

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):

$$\phi(x) = x^n + a_1 x^{n-1} + \dots + a_n. \quad (12.3)$$

We consider simultaneously with $\phi(x)$ the polynomial

$$\psi(x) = x^n \phi\left(\frac{1}{x}\right) = 1 + a_1 x + \dots + a_n x^n. \quad (12.4)$$

If the constants k_ν are given, we can consider the power series

$$K(x) = \sum_{\nu=-n}^{\infty} k_\nu x^\nu \quad (12.5)$$

as given, and the problem of determination of the z_ν by (12.1) can be considered as the problem of determination of the corresponding *generating series*

$$Z(x) = \sum_{\nu=0}^{\infty} z_\nu x^\nu. \quad (12.6)$$

3. If we then multiply $Z(x)$ by $\psi(x)$, the coefficients of x^ν in the product are for $\nu \geq n$, in virtue of (12.1), just the corresponding k_ν . 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_ν are obtained in developing the right-side 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 $\phi(x)$ 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_ν 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 $\phi(x)$ has multiple zeros and, e.g., u is a zero of $\phi(x)$ 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 the 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 $\phi(x)$, we obtain finally, if v_1, \dots, v_k are the *different* zeros of $\phi(x)$ with the corresponding multiplicities m_1, \dots, m_k , as the complete solution of (12.2) the expression

$$z_\mu = \sum_{\kappa=1}^k v_\kappa^\mu Q_\kappa(u), \quad (12.10)$$

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 $|\xi| < s$, 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_k 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 $\gamma_1 = s/(s - |\xi|)$:

$$\frac{1}{(1 - \xi x)(1 - sx)^m} \ll \frac{s/(s - |\xi|)}{(1 - sx)^m} \quad (m = 1, 2, \dots). *$$

* 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.

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 the 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 zeros u_1 and u_2 of $\phi(x)$*

$$|u_1| > 1 > |u_2| \quad (12.11)$$

and for a positive constant $s < |u_1|$

$$k_\nu = O(s^\nu) \quad (\nu \rightarrow \infty). \quad (12.12)$$

Then we have for a convenient constant α ,

$$\frac{z_\nu}{u_1^\nu} \rightarrow \alpha \quad (\nu \rightarrow \infty). \quad (12.13)$$

Further we have, if $s > |u_2|$,

$$z_\nu = \alpha u_1^\nu + O(s^\nu) \quad (\nu \rightarrow \infty). \quad (12.14)$$

If $s = |u_2|$ and m is the maximal multiplicity of the zeros of $\phi(x)$ with modulus $= |u_2|$, we have

$$z_\nu = \alpha u_1^\nu + O(\nu^m u_2^\nu) \quad (\nu \rightarrow \infty). \quad (12.15)$$

If $s < |u_2|$, we have, m having the same meaning as above,

$$z_\nu = \alpha u_1^\nu + O(\nu^{m-1} u_2^\nu) \quad (\nu \rightarrow \infty). \quad (12.16)$$

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 (12.3), we have

$$Z(x) = \frac{K_1(x)}{\prod_{\kappa=1}^k (1 - u_\kappa x)}, \quad (12.17)$$

$$K_1(x) = \sum_{\nu=0}^{\infty} k_\nu' x^\nu,$$

with $k_\nu' = O(s^\nu)$ ($\nu \rightarrow \infty$). 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

$$P(x) = \frac{K_2(x)}{(1 - u_1 x) \prod_{\kappa=2}^k (1 - u_\kappa x)}, \quad (12.19)$$

$$K_2(x) = \sum_{\nu=0}^{\infty} k_\nu'' x^\nu = K_1(x) - \alpha \prod_{\kappa=2}^k (1 - u_\kappa x), \quad (12.20)$$

where we have again $k_\nu'' = O(s^\nu)$ and obviously $K_2(1/u_1) = 0$. Therefore, by Lemma 12.1 for $\xi = u_1$, the ν th coefficient of

$$\frac{K_2(x)}{1 - u_1 x} = K_3(x) = \sum_{\nu=0}^{\infty} k_\nu^{(3)} x^\nu$$

is $O(s^\nu)$, and we have

$$Z(x) - \frac{\alpha}{1 - u_1 x} = \frac{K_3(x)}{\prod_{\kappa=2}^k (1 - u_\kappa x)}. \quad (12.21)$$

10. If now $s > |u_2|$, we can apply here to the right-hand quotient the Corollary 1 to the above lemma and see that the ν th coefficient of the development of this quotient is $O(s^\nu)$. 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_k 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,

$$\frac{1}{\prod_{\kappa=2}^k (1 - u_\kappa x)} = \sum_{\tau=1}^t \frac{p_\tau(x)}{(1 - u^{(\tau)} x)^{m_\tau}},$$

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

$$k_{\tau,v} = O(s^v) \quad (v \rightarrow \infty; \tau = 1, \dots, t).$$

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 $\gamma / [(1 - sx)(1 - |u_2| x)^m]$ 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 $\leq \gamma^{(m+v)} |u_2|^v$, 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 apply now 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 there x_0 by x_{v-1} ,

$$y_{v+1} - y_v - y_{v-1} = \ln \left| \frac{2f'(\xi)^3}{f''(\xi)f'(\xi_1)f'(\xi_2)} \right|. \quad (12.25)$$

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

$$k_{v+1} = O(\delta_{v-1} + \delta_v).$$

On the other hand, it follows from the discussions in Sections 9 and 10 of Chapter 3 that

$$\delta_v = O(d^{t_1 v + 1} / \sqrt{5}) \quad (12.28)$$

* We have to apply here the mean value theorem either to $\ln f'(x)$ or to $\ln [-f'(x)]$ and correspondingly in the case of $f''(x)$.

where $0 < d < 1$ and $t_1 > 1$. Then we have obviously $\sqrt[d]{\delta} \rightarrow 0$, and we obtain finally from (12.25), putting $y_\nu = v_\nu - \delta$,

$$v_{\nu+1} - v_\nu - v_{\nu-1} = k_{\nu+1}, \quad (12.29)$$

where $\sum_{\nu=0}^{\infty} k_\nu x^\nu$ 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,

$$y_\nu + \delta - \alpha t_1^\nu = v_\nu - \alpha t_1^\nu = O(t_2^\nu) \quad (12.30)$$

where, since with $x_\nu \rightarrow \zeta$, $y_\nu \rightarrow \infty$, α is a positive constant. It follows further by (12.24), (12.27), and (12.29),

$$|x_\nu - \zeta| = \left| \frac{2f'(\zeta)}{f''(\zeta)} \right| \exp(-\alpha t_1^\nu) [1 + O(t_2^\nu)]. \quad (12.31)$$

We raise this formula to the power t_1 , rewrite (12.31) for $\nu + 1$, and divide; then we obtain

$$\frac{|x_{\nu+1} - \zeta|}{|x_\nu - \zeta|^{t_1}} = \left| \frac{2f'(\zeta)}{f''(\zeta)} \right|^{t_1} + O(t_2^\nu), \quad (12.32)$$

from which (3.19) follows immediately.

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, assuming that the right-hand expression in (11.30) does not vanish, 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

$$x^n - b_1 x^{n-1} - \dots - b_n = 0 \quad (12.33)$$

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 the equation (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

$$s_1 k_1 + s_2 k_2 + \dots + s_m k_m = 1. \quad (12.34)$$

The 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.*

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

$$\alpha = \min_{\kappa=1, \dots, n} \frac{u_\kappa}{\sigma^\kappa}. \quad (12.40)$$

Proof. The 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)

$$\begin{aligned} u_{n+N} &\geq \sum_{\kappa=0}^{n-1} p_\kappa u_{N+\kappa} \geq \alpha \sum_{\kappa=0}^{n-1} p_\kappa \sigma^{N+\kappa} \\ &= \alpha \sigma^N \sum_{\kappa=0}^{n-1} p_\kappa \sigma^\kappa. \end{aligned}$$

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.

* The condition imposed in the Theorem 12.2 on the indices of the nonvanishing b , is obviously essential. For if the left-side polynomial in (12.38) is a polynomial in x^k , $k > 1$, then there are k roots of (12.38) with the modulus p .

ERROR ESTIMATES

1. Let $w = f(x)$ be defined in J_x and $f(\zeta) = 0$, $\zeta \prec J_x$, where $f'(x) \neq 0$ for all x in J_x . Assume that we are given n distinct interpolation points x_ν ($\nu = 1, \dots, n$) in J_x in a sufficiently close neighborhood of ζ , $f(x_\nu) = y_\nu$ ($\nu = 1, \dots, n$). 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 $T(y_\nu) = x_\nu$ ($\nu = 1, \dots, n$).

Let $T(0) = x_{n+1}$ and assume that $f^{(n)}(x)$ is continuous in J_x . We then apply (1.30) by replacing there t_ν by y_ν , $f(t_\nu)$ by x_ν , x by 0, and $F(x)$ by

$$W(w) = \prod_{\nu=1}^n (w - y_\nu).$$

Then

$$x_{n+1} = T(0) = (-1)^{n-1} \prod_{\nu=1}^n y_\nu \sum_{\nu=1}^n \frac{x_\nu}{y_\nu} \frac{1}{W'(y_\nu)}. \quad (13.1)$$

We now apply (2.2) for $x = 0$ in replacing there ϕ by Φ and the x_ν by y_ν . We have then, putting

$$S = \frac{1}{n!} \Phi^{(n)}(\eta), \quad \eta \prec (0, y_1, \dots, y_n), \quad (13.2)$$

$$\zeta - x_{n+1} = (-1)^n S \prod_{\nu=1}^n y_\nu. \quad (13.3)$$

Since

$$y_\nu = f(x_\nu) = (x_\nu - \zeta) f'(\xi_\nu), \quad \xi_\nu \prec (x_\nu, \zeta),$$

we have from (13.3)

$$\begin{aligned}\zeta - x_{n+1} &= S \prod_{\nu=1}^n (\zeta - x_\nu) \prod_{\nu=1}^n f'(\xi_\nu), \\ \frac{\zeta - x_{n+1}}{\prod_{\nu=1}^n (\zeta - x_\nu)} &= S \prod_{\nu=1}^n f'(\xi_\nu),\end{aligned}\tag{13.4}$$

and therefore in particular, if the x_ν ($\nu = 1, \dots, n$) tend to ζ ,

$$\frac{\zeta - x_{n+1}}{\prod_{\nu=1}^n (\zeta - x_\nu)} \rightarrow \frac{1}{n!} \Phi^{(n)}(0) f'(\zeta)^n \quad (x_\nu \rightarrow \zeta, \nu = 1, \dots, n). \tag{13.5}$$

2. Let k be an upper bound of the modulus of the right-hand side of (13.4). Then

$$|\zeta - x_{n+1}| \leq k \prod_{\nu=1}^n |\zeta - x_\nu|.$$

Put $K = k^{1/(n-1)}$. Then we have further

$$K|\zeta - x_{n+1}| \leq \prod_{\nu=1}^n (K|\zeta - x_\nu|). \tag{13.6}$$

Taking all x_ν ($\nu = 1, \dots, n$) sufficiently close to ζ , we make all factors on the right-hand side of (13.6) < 1 . We then introduce ε_ν by

$$K|\zeta - x_\nu| = u^{\varepsilon_\nu} \quad (\nu \geq 1), \quad u = \max(K|\zeta - x_1|, K|\zeta - x_2|), \tag{13.7}$$

where $0 < u < 1$. From (13.6) and (13.7) we have

$$u^{\varepsilon_{n+1}} \leq u^{\varepsilon_1 + \dots + \varepsilon_n}. \tag{13.8}$$

ITERATION WITH n DISTINCT POINTS OF INTERPOLATION

3. Since in (13.7) $u < 1$ we have

$$\varepsilon_{n+1} \geq \varepsilon_1 + \dots + \varepsilon_n. \tag{13.9}$$

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

$$s_n = \sum_{\nu=1}^n \varepsilon_\nu \quad (n \geq 2).$$

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 $\min(\varepsilon_1, \varepsilon_2) = 1$. 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 $\varepsilon_\mu \rightarrow \infty$ ($\mu \rightarrow \infty$),

$$x_\nu - \zeta \rightarrow 0. \quad (13.12)$$

* This procedure appears to present another disadvantage, since it seems that the first f_μ must be calculated with arbitrarily great precision. However, as follows from Appendix H, this disadvantage is not very essential.

6. Let

$$\ln \frac{1}{|\zeta - x_\nu|} = y_\nu. \quad (13.13)$$

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.

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

$$f_{n,p}(x) \equiv px^n - x^{n-1} - \dots - x - 1 = 0. \quad (13.16)$$

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 $\mu_{n,p}$ 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 $m \geq n$:

$$U_m(n, p) \equiv (n+1)^{m+1} - (p+1) \left(1 + \frac{1}{p} \right)^m n^m < 0 \quad (1 \leq p < n \leq m). \quad (13.18)$$

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 prove now 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

$$\frac{f_{n,p}(x)}{x - \mu} = c_0 x^{n-1} + \dots + c_\nu x^{n-\nu+1} + \dots + c_{n-1}, \quad (13.22)$$

where

$$c_0 = p, \quad c_\nu = p\mu^\nu - \mu^{\nu-1} - \dots - \mu - 1 \quad (1 \leq \nu \leq n-1).$$

If we bring the last $n - \nu$ terms to the right in the equation

$$p\mu^n - \mu^{n-1} - \dots - \mu - 1 = 0$$

and divide by $\mu^{n-\nu}$, we obtain

$$c_\nu = \frac{1}{\mu} + \frac{1}{\mu^2} + \dots + \frac{1}{\mu^{n-\nu}} = \xi \frac{1 - \xi^{n-\nu}}{1 - \xi} \quad (\nu = 1, \dots, n-1).$$

We see that all coefficients c_ν of (13.22) are positive.

11. We now prove

$$\frac{c_{\nu+1}}{c_\nu} < \frac{c_\nu}{c_{\nu-1}} \quad (\nu = 1, \dots, n-2). \quad (13.23)$$

For $\nu = 1$ 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 $2 \leq \nu \leq n-2$, putting $k = n-\nu$, $2 \leq k \leq n-2$, (13.23) is equivalent to

$$\frac{1 - \xi^{k+1}}{1 - \xi^k} < \frac{1 - \xi^k}{1 - \xi^{k-1}},$$

$$(1 - \xi^k)^2 > (1 - \xi^{k+1})(1 - \xi^{k-1}),$$

$$2\xi^k < \xi^{k-1} + \xi^{k+1},$$

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

$$g(x) = a_0x^n + \dots + a_n = 0 \quad (13.24)$$

all coefficients a_v are positive, then we have for each root ξ of (13.24)

$$|\xi| \leq \gamma = \max_{1 \leq v \leq n} \frac{a_v}{a_{v-1}}. \quad (13.25)$$

*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,

$$q_n < \max \left(\frac{c_1}{c_0}, \frac{c_2}{c_1}, \dots, \frac{c_{n-1}}{c_{n-2}} \right) = \frac{c_1}{c_0} = \frac{1}{p} c_1. \quad (13.26)$$

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.

* 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.

We have

$$(x - \gamma)g(x) = a_0 x^{n+1} - (\gamma a_0 - a_1)x^n - \dots - (\gamma a_{v-1} - a_v)x^{n-v} - \dots - a_n\gamma.$$

Since by definition of γ all expressions $\gamma a_{v-1} - a_v$ ($v = 1, \dots, n$) are ≥ 0 , γ is a simple zero and the only positive zero of $(x - \gamma)g(x)$, 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 $(x - \gamma)g(x)$ 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)

$$\frac{|\zeta - x_1| \dots |\zeta - x_n|}{|\zeta - x_{n+1}|} = \frac{n!}{|\Phi^{(n)}(\eta)|} \prod_{v=1}^n \frac{1}{|f'(\xi_v)|}$$

and, therefore, assuming that

$$\Phi^{(n)}(0) \neq 0, \quad f'(\zeta) \neq 0, \quad (13.29)$$

$$y_{n+1} - y_1 - \dots - y_n = \ln \left| n! \frac{f'(\zeta)^{-n}}{\Phi^{(n)}(0)} \right| - \ln \left| \frac{\Phi^{(n)}(\eta)}{\Phi^{(n)}(0)} \right| - \sum_{v=1}^n \ln \left| \frac{f'(\xi_v)}{f'(\zeta)} \right|.$$

If now, as $v \rightarrow \infty$,

$$x_v \rightarrow \zeta, \quad y_v \rightarrow \infty,$$

we can write

$$\begin{aligned} y_{\mu+n} - y_\mu - \dots - y_{\mu+n-1} &= \ln \left| n! \frac{f'(\zeta) - n}{\Phi^{(n)}(0)} \right| \\ &\quad - \ln \left| \frac{\Phi^{(n)}(\eta^{(\mu)})}{\Phi^{(n)}(0)} \right| - \sum_{\nu=1}^n \ln \left| \frac{f'(\xi_{\nu}^{(\mu)})}{f'(\zeta)} \right|. \end{aligned} \quad (13.30)$$

16. As to the $\xi_{\nu}^{(\mu)}$, they lie between the greatest and the smallest of the $n+2$ numbers

$$x_\mu, \dots, x_{\mu+n}, \zeta$$

and tend to ζ as $\mu \rightarrow \infty$. And we easily see from (13.2) that $|\eta^{(\mu)}|$ is $\leq \max(|f(x_\mu)|, \dots, |f(x_{\mu+n-1})|)$. We have, therefore, since, from a certain index μ on, the $|x_\mu - \zeta|$ decrease,

$$\eta^{(\mu)} = O(x_\mu - \zeta), \quad \xi_{\nu}^{(\mu)} - \zeta = O(x_\mu - \zeta).$$

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

$$\begin{aligned} \frac{\Phi^{(n)}(\eta^{(\mu)})}{\Phi^{(n)}(0)} &= 1 + O(|x_\mu - \zeta|), \\ \frac{f'(\xi_{\nu}^{(\mu)})}{f'(\zeta)} &= 1 + O(|x_\mu - \zeta|), \\ k_{n+\mu} \equiv -\ln \left| \frac{\Phi^{(n)}(\eta^{(\mu)})}{\Phi^{(n)}(0)} \right| - \sum_{\nu=1}^n \ln \left| \frac{f'(\xi_{\nu}^{(\mu)})}{f'(\zeta)} \right| &= O(|x_\mu - \zeta|). \end{aligned} \quad (13.31)$$

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

$$y_\mu = v_\mu + \beta. \quad (13.34)$$

The 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 $u_1 = \mu_n$, $m = 1$,

$$v_\nu = \alpha \mu_n^\nu + O(q_n^\nu),$$

$$y_\nu = \beta + \alpha \mu_n^\nu + O(q_n^\nu),$$

$$|x_\nu - \zeta| = \exp(-\beta) \exp(-\alpha \mu_n^\nu) [1 + O(q_n^\nu)], \quad (13.36)$$

$$\frac{|x_{\nu+1} - \zeta|}{|x_\nu - \zeta|^{\mu_n}} = \exp[(\mu_n - 1)\beta] + O(q_n^\nu), \quad (13.37)$$

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 worth while 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.

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 :

$$z = \Phi(w) = z_0 + \sum_{\nu=1}^{\infty} \frac{(w - w_0)^{\nu}}{\nu!} \Phi^{(\nu)}(w_0). \quad (14.1)$$

For $w = 0$ it follows that

$$\zeta - z_0 = \sum_{\nu=1}^{\infty} \frac{(-1)^{\nu}}{\nu!} w_0^{\nu} \Phi^{(\nu)}(w_0). \quad (14.2)$$

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 (1.23b), if the expression of $\Phi^{(n+1)}$ 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 $f^{(\nu)}(z_0)$ 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

$$z = \Phi(w) = z_0 + \sum_{v=1}^{\infty} \frac{(w - w_0)^v}{v!} \Phi^{(v)}(w_0) \quad (14.4)$$

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 satisfies there the inequality

$$|\Phi(w) - z_0| < r. \quad (14.6)$$

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) we have

$$T(z) = z^2 \int_0^1 (1-t)f''(tz) dt. \quad (14.9)$$

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 $z = \phi(w)$ is analytic in K_w and satisfies there the 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 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 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

$$2M \left[\frac{1}{2} \frac{|f'_0|^2}{M} - \left(|f'_0|r - \frac{Mr^2}{2} \right) \right] = (|f'_0| - rM)^2 \quad (14.15)$$

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 deduce now 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

$$r \sum_{\nu=1}^{\infty} \frac{|w_0|^{\nu}}{R(r)^{\nu}}.$$

We have in particular, putting

$$E_n = \zeta - z_0 - \sum_{\nu=1}^n \frac{(-1)^{\nu}}{\nu!} w_0^{\nu} \Phi^{(\nu)}(w_0), \quad (14.17)$$

$$|E_n| \leq \left[\frac{|w_0|}{R(r)} \right]^{n+1} \frac{r}{1 - |w_0|/R(r)}. \quad (14.18)$$

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

$$\left| \frac{1}{\nu!} \Phi^{(\nu)}(w_0) \right| \leq \frac{r}{R^{\nu}(r)} \quad (\nu = 1, 2, \dots), \quad (14.19)$$

which gives us the majorant for (14.2) and in particular

$$\begin{aligned} \left| \zeta - z_0 - \sum_{\nu=1}^n \frac{(-1)^{\nu}}{\nu!} w_0^{\nu} \Phi^{(\nu)}(w_0) \right| &\leq r \sum_{\nu=n+1}^{\infty} \left[\frac{|w_0|}{R(r)} \right]^{\nu} \\ &= \left[\frac{|w_0|}{R(r)} \right]^{n+1} \frac{r}{1 - |w_0|/R(r)}, \end{aligned} \quad \text{Q.E.D.}$$

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 now 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 7.1,

$$2\rho M \leq f_0', \quad (14.25)$$

where ρ is given by (14.20), while $M = \max|f''(z)|$ is taken in the circle K_0 , $|z - z_1| = |z - z_0 - h_0| \leq \rho$. Since $M(2\rho)$ is the $\max|f''(z)|$ 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).

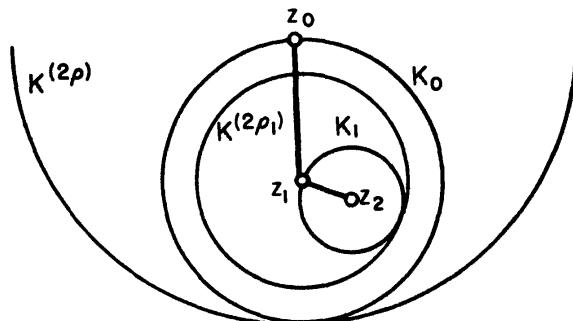


FIG. 5.

It follows from the proof of 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 (7.10) 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 7.1 as a bound for $M(2\rho_1)$ computed in the circle around z_1 . As a matter of fact, we obtain then from (7.9) the inequality corresponding to (14.25) with \leq . But deriving this relation, we use (7.5) and (7.7), and in these inequalities our proof excludes the equality sign, 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 7.1 are satisfied, the Taylor development (14.2) is always convergent if we replace there $z_0, f(z_0)$ by $z_v, f(z_v)$ for any $v > 0$, 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

$$|E_n| \leq \left[\frac{|w_0|}{R(r)} \right]^{m+1} \frac{r}{1 - |w_0|/R(r)} + \sum_{v=n+1}^m \frac{|w_0|^v}{v!} |\Phi^v(w_0)|. \quad (14.26)$$

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 $\phi(w)$ in w_0 are

$$\phi'(w_0) = 0.1, \quad \phi''(w_0) = -0.012, \quad \phi'''(w_0) = 0.00372,$$

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

$$\begin{aligned} \phi'_0 &= 0.089047195, \quad \phi''_0 = -0.00889674773, \quad \phi'''_0 = 0.002289382342, \\ \zeta &= 2.094551482097. \end{aligned} \quad (14.28)$$

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.

1. Consider a polynomial

$$f(x) = \prod_{v=1}^n (x - \zeta_v) \quad (15.1)$$

of exact degree n with n real zeros ordered monotonically:

$$\zeta_1 \leq \zeta_2 \leq \dots \leq \zeta_n. \quad (15.2)$$

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,

$$\zeta_1 \leq \zeta'_1 \leq \zeta_2 \leq \zeta'_2 \leq \dots \leq \zeta'_{n-1} \leq \zeta_n, \quad (15.3)$$

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 $\operatorname{sgn}(f(x)/f'(x)) = -1$. 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),

$$\frac{f'(x)}{f(x)} = \sum_{\nu=1}^n \frac{1}{x - \zeta_\nu}, \quad (15.6)$$

and differentiating this again, we obtain after multiplication by -1

$$H(x) \equiv \frac{f'^2 - ff''}{f^2} = \sum_{\nu=1}^n \frac{1}{(x - \zeta_\nu)^2}. \quad (15.7)$$

The formulas (15.6) and (15.7) hold, of course, also if the ζ_ν are not necessarily real. But since in our case all ζ_ν are real we have obviously from (15.7)

$$\frac{1}{(x - \zeta(x))^2} \leq H(x), \quad |x - \zeta(x)| \geq \sqrt{H(x)}. \quad (15.8)$$

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_ν and ζ'_ν and form, starting with $x_0 = x$, the sequence x_ν by the iteration rule:

$$x_{\nu+1} = x_\nu - K(x_\nu) \quad (\nu = 0, 1, \dots, \quad x_0 = x). \quad (15.10)$$

5. Theorem 15.1. *The x_ν in (15.10) ($\nu = 1, 2, \dots$) lie all 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

$$\frac{f'(x)}{f(x)} < 0, \quad K(x) < 0.$$

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 $k = -f(x)/f'(x) \rightarrow 0$,

$$x - x_1 = K(x) = -k + \frac{1}{2}k^2 \frac{f''(x)}{f'(x)} - \frac{1}{8}k^3 \frac{f''(x)^2}{f'(x)^2} (1 + O(k)).$$

On the other hand, we have from (2.20)

$$\zeta - x = k - \frac{1}{2}k^2 \frac{f''(x)}{f'(x)} + \frac{3f''(x)^2 - f'(x)f'''(x)}{6f'(x)^2} k^3 (1 + O(k)).$$

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 here x by x , (15.11) follows immediately.

7. Assume now that ζ is a multiple zero of $f(x)$ of multiplicity $p > 1$. Then we have $f(x) = (x - \zeta)^p \phi(x)$, $\phi(\zeta) \neq 0$, and therefore for $x \rightarrow \zeta$

$$f(x) \sim (x - \zeta)^p \phi(\zeta),$$

and further, by differentiation,

$$f'(x) \sim p(x - \zeta)^{p-1} \phi(\zeta), \quad f''(x) \sim p(p-1)(x - \zeta)^{p-2} \phi(\zeta).$$

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),

$$K(x) \sim \frac{x - \zeta}{\sqrt[p]{p}}.$$

But then it follows from (15.10) with $v = 0$, $x_0 = x$,

$$x_1 - \zeta = x - \zeta - K(x),$$

and, dividing on both sides by $x - \zeta$,

$$\frac{x_1 - \zeta}{x - \zeta} \rightarrow 1 - \frac{1}{\sqrt[p]{p}}.$$

Replacing here x by x_ν , we obtain (15.12) and Theorem 15.2 is proved.

8. In the case that $\zeta = \zeta(x)$ is a multiple zero of the multiplicity $p > 1$ and if p is known, we can still obtain cubic convergence to ζ modifying slightly the rule (15.10). Put $x = x_0$ and

$$x_{\nu+1} = x_\nu - \sqrt[p]{pK(x_\nu)} \quad (x_0 = x; \quad \nu = 0, 1, \dots). \quad (15.15)$$

Then we have

Theorem 15.3. If under the conditions of the 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_ν converge monotonically to ζ and we have

$$\frac{x_{\nu+1} - \zeta}{(x_\nu - \zeta)^3} \rightarrow \frac{(p+2)f^{(p+1)}(\zeta)^2 - 2(p+1)f^{(p)}(\zeta)f^{(p+2)}(\zeta)}{2p(p+1)^2(p+2)f^{(p)}(\zeta)^2}. \quad (15.16)$$

9. Proof. We have $f(x) = (x - \zeta)^p \phi(x)$ and, denoting by $\Phi_0, \Phi_0', \Phi_0''$, respectively, the values of $\phi(\zeta), \phi'(\zeta), \phi''(\zeta)$,

$$\begin{aligned} \phi(x) &= \Phi_0 + (x - \zeta)\Phi_0' + \frac{1}{2}(x - \zeta)^2\Phi_0'' + O((x - \zeta)^3), \\ \phi'(x) &= \Phi_0' + (x - \zeta)\Phi_0'' + O((x - \zeta)^2), \\ \phi''(x) &= \Phi_0'' + O((x - \zeta)), \end{aligned}$$

and therefore, differentiating $(x - \zeta)^p \phi(x)$,

$$\begin{aligned} f(x) &= (x - \zeta)^p \Phi_0 + (x - \zeta)^{p+1} \Phi_0' + \frac{1}{2}(x - \zeta)^{p+2} \Phi_0'' + \dots, \\ f'(x) &= p(x - \zeta)^{p-1} \Phi_0 + (p+1)(x - \zeta)^p \Phi_0' \\ &\quad + \frac{1}{2}(p+2)(x - \zeta)^{p+1} \Phi_0'' + \dots, \\ f''(x) &= (p-1)p(x - \zeta)^{p-2} \Phi_0 + p(p+1)(x - \zeta)^{p-1} \Phi_0' \\ &\quad + \frac{1}{2}(p+1)(p+2)(x - \zeta)^p \Phi_0'' + \dots. \end{aligned} \quad (15.17)$$

From these formulas we have

$$\begin{aligned} p \left(\frac{f(x)}{(x - \zeta)^p} \right)^2 &= p\Phi_0^2 + 2p\Phi_0\Phi_0'(x - \zeta) \\ &\quad + p(\Phi_0'^2 + \Phi_0\Phi_0'')(x - \zeta)^2 + \dots, \\ \left(\frac{f'(x)}{(x - \zeta)^{p-1}} \right)^2 &= p^2\Phi_0^2 + 2p(p+1)\Phi_0\Phi_0'(x - \zeta) + [(p+1)^2\Phi_0'^2 \\ &\quad + p(p+2)\Phi_0\Phi_0''](x - \zeta)^2 + \dots, \end{aligned}$$

$$\begin{aligned} \frac{f(x)f''(x)}{(x-\zeta)^{2p-2}} &= p(p-1)\Phi_0^2 + 2p^2\Phi_0\Phi_0'(x-\zeta) + [p(p+1)\Phi_0'^2 \\ &\quad + (p^2+p+1)\Phi_0\Phi_0''](x-\zeta)^2 + \dots \end{aligned}$$

Adding the first and the third of these formulas and subtracting the second one, we obtain on the right-hand side

$$(\Phi_0\Phi_0'' - \Phi_0'^2)(x-\zeta)^2 + O((x-\zeta)^3),$$

and therefore, multiplying by $(x-\zeta)^{2p}$,

$$\begin{aligned} pf(x)^2 - (x-\zeta)^2(f'(x)^2 - f(x)f''(x)) \\ = (\Phi_0\Phi_0'' - \Phi_0'^2)(x-\zeta)^{2p+2} + O((x-\zeta)^{2p+3}). \end{aligned}$$

10. It follows now from (15.7) that

$$\begin{aligned} 1 - \frac{(x-\zeta)^2 H(x)}{p} &= \frac{pf(x)^2 - (x-\zeta)^2(f'(x)^2 - f(x)f''(x))}{pf(x)^2} \\ &= \frac{\Phi_0\Phi_0'' - \Phi_0'^2}{p\Phi_0^2}(x-\zeta)^2 + O((x-\zeta)^3), \end{aligned}$$

and further

$$\frac{p}{(x-\zeta)^2 H(x)} = 1 + \frac{\Phi_0\Phi_0'' - \Phi_0'^2}{p\Phi_0^2}(x-\zeta)^2 + O((x-\zeta)^3).$$

Therefore, since $K(x)^2 = 1/H(x)$,

$$K(x)^2 = \frac{(x-\zeta)^2}{p} \left(1 + \frac{\Phi_0\Phi_0'' - \Phi_0'^2}{p\Phi_0^2}(x-\zeta)^2 + O((x-\zeta)^3) \right).$$

If we take the square root, it follows that

$$K(x) = \frac{x-\zeta}{\sqrt{p}} \left(1 + \frac{\Phi_0\Phi_0'' - \Phi_0'^2}{2p\Phi_0^2}(x-\zeta)^2 \right) + O((x-\zeta)^4). \quad (15.18)$$

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

$$x - \zeta - \sqrt{p}K(x) = \frac{\Phi_0'^2 - \Phi_0\Phi_0''}{2p\Phi_0^2}(x-\zeta)^3 + O((x-\zeta)^4),$$

and therefore, by virtue of (15.15),

$$x_{v+1} - \zeta = \frac{\Phi'(x_v)^2 - \Phi(x_v)\Phi''(x_v)}{2\phi\Phi_0^2} (x_v - \zeta)^3 + O((x_v - \zeta)^4).$$

From this we have finally

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^3} \rightarrow \frac{\Phi_0'^2 - \Phi_0\Phi_0''}{2\phi\Phi_0^2}. \quad (15.19)$$

If we observe that, by (15.17),

$$\Phi_0 = \frac{f^{(\phi)}(\zeta)}{\phi!}, \quad \Phi_0' = \frac{f^{(\phi+1)}(\zeta)}{(\phi+1)!}, \quad \Phi_0'' = 2 \frac{f^{(\phi+2)}(\zeta)}{(\phi+2)!},$$

(15.16) now follows immediately from (15.19). Theorem 15.3 is proved.

12. It can be expected *a priori* that the 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 the 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 $f(x)$, the 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

$$q_0 \leq \frac{1}{10}. \quad (15.21)$$

Denoting $\max|f^{(3)}(x)|$ in U by M_3 , assume further that we have throughout U

$$|q(x)| \leq \frac{1}{5}, \quad 2M_3|h(x)|^2 \leq |f'(x)| \quad (|x - x_0| \leq 6h_0). \quad (15.22)$$

Then $f(x)$ has exactly one zero ζ_0 in the neighborhood U_0 , $|x - x_0| \leq 2h_0$, of x_0 and the 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)

$$\psi'(x) = 1 - (1 - q)(1 - q)^{-1/2} - \frac{1}{2}h(1 - q)^{-3/2}q',$$

$$\psi'(x) = 1 - \frac{h^2}{2}(1 - q)^{-3/2} \frac{f^{(3)}}{f'} - (1 - q)^{-3/2} \left[(1 - q)^2 + \frac{h}{2} \frac{f'^2f'' - 2ff''^2}{f'^3} \right].$$

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

$$\begin{aligned} \psi'(x) &= 1 - \frac{1 - \frac{3}{2}q}{(1 - q)^{3/2}} - \frac{h^2}{2(1 - q)^{3/2}} \frac{f^{(3)}}{f'}, \\ \psi'(x) &= Q(q) - \frac{h^2}{2(1 - q)^{3/2}} \frac{f^{(3)}}{f'}, \quad Q(u) = 1 - \frac{1 - \frac{3}{2}u}{(1 - u)^{3/2}}. \end{aligned} \quad (15.25)$$

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 $|u| < 1$, $|Q'(u)| < Q'(|u|)$. It follows, therefore, that

$$|Q(u)| \leq \left| \int_0^u Q'(t) dt \right| \leq \left| \int_0^u Q'(|t|) dt \right|,$$

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

$$|Q(q)| \leq Q(1/5) = 1 - \frac{7/5}{2(4/5)^{3/2}} = \frac{16 - 7\sqrt{5}}{16} = \frac{11}{16(16 + 7\sqrt{5})} < \frac{1}{40}.$$

As to the second term in the 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 by (15.21)

$$|\psi(x_0) - x_0| = \frac{h_0}{(1-q_0)^{1/2}} \leq \sqrt{\frac{5}{4}} h_0.$$

We will apply now 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 is here $2h_0$, 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 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 proofs in Chapter 7.

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

$$f_0 = |f(z_0)|, \quad f_0' = |f'(z_0)|, \quad f_0'' = \max(1, |f''(z_0)|),$$

$$q_0 = h_0 \frac{f''(z_0)}{f'(z_0)} = \frac{f(z_0)f''(z_0)}{f'(z_0)^2}, \quad p_0 = \frac{M|h_0|^2}{f_0'},$$

$$Q_0 = \frac{f_0 f_0''}{f_0'^2}, \quad k_0 = \frac{h_0}{\sqrt{1 - q_0}},$$

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.

Assume further that we have

$$Q_0 \leq \frac{1}{4}, \tag{16.1}$$

$$p_0 \leq \frac{1}{8}, \tag{16.2}$$

$$|h_0|M \leq \frac{1}{8}f_0''. \tag{16.3}$$

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

$$\frac{1}{|w|} = \frac{1}{|\sqrt{1 - q_0}|} \leq \frac{1}{\sqrt{\frac{3}{4}}} = \frac{2\sqrt{3}}{3}, \quad (16.6)$$

$$\frac{1}{|w(1 + w)|} \leq \frac{2}{3}\sqrt{3} \cdot 2(2 - \sqrt{3}) = \frac{4}{3}(2\sqrt{3} - 3) < 0.7. \quad (16.7)$$

Further, from (16.5) and (16.6)

$$\left| \frac{w+2}{2w(w+1)^2} \right| \leq 4 \left(\frac{1}{2} + \frac{2\sqrt{3}}{3} \right) (7 - 4\sqrt{3}) = \frac{4}{3}(16\sqrt{3} - 27) < 0.5. \quad (16.8)$$

3. We put now, by (15.23),

$$k_0 = z_0 - z_1 = \frac{h_0}{w}, \quad \eta_0 \equiv k_0 - h_0 = \frac{1-w}{w} h_0 = \frac{h_0 q_0}{w(w+1)}. \quad (16.9)$$

Then, by (16.7) and (16.4),

$$|\eta_0| = \frac{|h_0 q_0|}{|w(1+w)|} < 0.7 |h_0 q_0| \leq (7/40) |h_0| < \frac{1}{5} |h_0|. \quad (16.10)$$

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),

$$\begin{aligned} \delta_0 &= \frac{q_0 h_0}{w(1+w)} - \frac{h_0 q_0}{2} = \frac{(w+2)(1-w)}{2w(1+w)} h_0 q_0 = \frac{w+2}{2w(1+w)^2} h_0 q_0^2, \\ |\delta_0| &< \frac{1}{2} |h_0 q_0|^2, \end{aligned}$$

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

$$f_1 = |f(z_1)|, \quad f_1' = |f'(z_1)|, \quad f_1'' = \max(1, |f''(z_1)|), \quad (16.14)$$

$$h_1 = \frac{f(z_1)}{f'(z_1)}, \quad Q_1 = \left| \frac{f_1 f_1''}{f_1'^2} \right|, \quad p_1 = h_1^2 \frac{M}{f_1'}.$$

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),

$$f''(z_1) - f''(z_0) = \theta k_0 M, \quad |\theta| \leq 1,$$

$$|f''(z_1) - f''(z_0)| \leq 1.2 |h_0| M = 1.2 f_0'' \left(|h_0| \frac{M}{f_0''} \right),$$

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 here k_0 by $h_0 + \eta_0$ and k_0^2 by $(h_0 + \eta_0)^2$; 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),

$$|\eta_0| \leq 0.7 |h_0 q_0| \leq 0.7 |h_0|^2 \frac{f''_0}{f'_0},$$

$$\frac{|2h_0 + \eta_0|}{2} \leq 1.1 |h_0|$$

and therefore

$$\begin{aligned} f_1 &\leq |h_0|^3 \left(0.5 \frac{f''_0^2}{f'_0} + 0.77 \frac{f''_0^2}{f'_0} + 0.288 M \right) \\ &< |h_0|^3 \left(1.3 \frac{f''_0^2}{f'_0} + 0.3 M \right). \end{aligned} \tag{16.17}$$

6. Dividing (16.17) on both sides by f'_1 we have by (16.14), (16.16), (16.1), (16.2)

$$\begin{aligned} |h_1| &= \frac{f_1}{f'_1} < |h_0| \left(2.6 \left(|h_0| \frac{f''_0}{f'_0} \right)^2 + 0.6 |h_0|^2 \frac{M}{f'_0} \right) \\ &< |h_0| \left(2.6 Q_0^2 + \frac{0.6}{8} \right) \leq \frac{19}{80} |h_0|, \\ |h_1| &< \frac{1}{4} |h_0|. \end{aligned} \tag{16.18}$$

On the other hand, multiplying (16.17) on both sides by f''_1/f'^2_1 and observing that this is, by (16.15) and (16.16), $\leq 4 \cdot 1.2 f''_0/f'^2_0 < 5(f''_0/f'^2_0)$, we have from the definitions of Q_1 , Q_0 , and p_0

$$Q_1 < 5 |h_0|^3 \left(1.3 \frac{f''_0^3}{f'^3_0} + 0.3 \frac{M f''_0}{f'^2_0} \right) = 6.5 Q_0^3 + 1.5 p_0 Q_0.$$

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),

$$p_1 = |h_1|^2 \frac{M}{f'_1} \leq \frac{2}{16} |h_0|^2 \frac{M}{f'_0} = \frac{1}{8} p_0 < \frac{1}{8}. \quad (16.20)$$

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 . We can therefore form indeed the whole infinite sequence z_ν by the iteration function (15.23) and have in particular

$$|z_{\nu+1} - z_\nu| \leq 1.2 \left| \frac{f(z_\nu)}{f'(z_\nu)} \right|, \quad \left| \frac{f(z_{\nu+1})}{f'(z_{\nu+1})} \right| < \frac{1}{4} \left| \frac{f(z_\nu)}{f'(z_\nu)} \right|.$$

From these inequalities we have now

$$\left| \frac{f(z_\nu)}{f'(z_\nu)} \right| \leq \frac{|h_0|}{4^\nu}, \quad |z_{\nu+1} - z_\nu| \leq \frac{1.2}{4^\nu} |h_0| \quad (\nu = 0, 1, 2, \dots),$$

and we see that $\zeta = \lim_{\nu \rightarrow \infty} z_\nu$ 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

$$0 = \lim_{\nu \rightarrow \infty} \frac{f(z_\nu)}{f'(z_\nu)} = \frac{f(\zeta)}{f'(\zeta)}$$

that in any case $f(\zeta) = 0$. Theorem 16.1 is proved.

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 by z in (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_n , 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 $\zeta(z_0)$.*

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

$$z_{v+1} = z_v + \sin z_v.$$

* The result of Theorem 16.2 is far less general than that of Theorem 15.1 in so far 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 $f'(x)$ can lie outside an interval containing all zeros of $f(x)$.

† The symbol $\operatorname{sgn} \alpha$ has the meaning $\alpha/|\alpha|$. It is of course defined only for $\alpha \neq 0$.

Replacing here $\sin z_v$ by $\sin(\pi - z_v)$ and developing this in powers of $(\pi - z_v)$ we obtain for our iteration

$$\pi - z_{v+1} = \pi - z_v - \left[(\pi - z_v) - \frac{(\pi - z_v)^3}{6} + \dots \right]$$

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).

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

$$f(z) = T(z) + \varphi(z), \quad T(z) = \sum_{\nu=p}^{n-1} a_\nu (z - \zeta)^\nu, \quad a_p = \frac{f^{(p)}(\zeta)}{p!} \neq 0, \quad (17.5)$$

where $\varphi(z)$ has in ζ a zero of the order $\geq n$. Put

$$A = |a_{p+1}| + \dots + |a_{n-1}|. \quad (17.6)$$

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**

$$\varepsilon_0 < \frac{r}{3}, \quad \varepsilon_0 \leq \frac{1}{3}, \quad \varepsilon_0 \leq \frac{|a_p|}{6A}, \quad \varepsilon_0^{n-p} \frac{K}{n!} \leq \frac{|a_p|}{7^n}. \quad (17.7)$$

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 , and given by the formula (1.21). We have in our notations

$$\begin{aligned} L_{n-1}(f, z) &= f(z_1) + (z - z_1)[z_1, z_2]f + (z - z_1)(z - z_2)[z_1, z_2, z_3]f \\ &\quad + \dots + (z - z_1)\dots(z - z_{n-1})[z_1, \dots, z_n]f. \end{aligned} \quad (17.8)$$

This is a linear operator applied to $f(z)$ and we have therefore from (17.5)

$$L(z) = L_{n-1}(T, z) + L_{n-1}(\Phi, z).$$

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

$$L(z) = T(z) + L^*(z), \quad L^*(z) = L_{n-1}(\Phi, z). \quad (17.9)$$

4. We have from (17.5)

$$\frac{T(z)}{z^p} = \sum_{\nu=p}^{n-1} a_\nu z^{\nu-p}, \quad \left| \frac{T(z)}{z^p} \right| \geq |a_p| - \sum_{\nu=p+1}^{n-1} |a_\nu| \cdot |z|^{\nu-p}.$$

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 bounds for $L^*(z)$, for both $|z| = \varepsilon$ and $|z| = 3\varepsilon$.

Put, for $\nu = 0, 1, \dots$,

$$\max_{|z|=\varepsilon} |\Phi^{(\nu)}(z)| = M_\nu, \quad \max_{|z|=3\varepsilon} |\Phi^{(\nu)}(z)| = M'_\nu. \quad (17.11)$$

* If $A = 0$, the third of the conditions (17.7) is to be disregarded.

Then we have, replacing $f(z)$ by $\phi(z)$ in (17.8) and estimating the divided differences by (1.11),

$$\begin{aligned} |L^*(z)| &\leq \sum_{\nu=0}^{n-1} \frac{(2\varepsilon)^\nu}{\nu!} M_\nu \quad (|z| \leq \varepsilon), \\ |L^*(z)| &\leq \sum_{\nu=0}^{n-1} \frac{(4\varepsilon)^\nu}{\nu!} M'_\nu \quad (|z| \leq 3\varepsilon), \end{aligned} \quad (17.12)$$

and we now have to obtain the estimates of M_ν and M'_ν .

6. Differentiating (17.5) ν times, we have

$$f^{(\nu)}(z) = T^{(\nu)}(z) + \phi^{(\nu)}(z),$$

where $\phi^{(\nu)}(z)$ is the remainder of the Maclaurin development of $f^{(\nu)}(z)$ if we stop at $z^{n-1-\nu}$. But then we have for a positive $\delta < r$ and $|z| \leq \delta$

$$|\phi^{(\nu)}(z)| \leq \frac{\delta^{n-\nu}}{(n-\nu)!} \max_{|z|=\delta} |(f^{(\nu)}(z))^{(n-\nu)}|.$$

From (17.4) we have, therefore, putting $\delta = \varepsilon$ and $\delta = 3\varepsilon$,

$$M_\nu \leq \frac{\varepsilon^{n-\nu}}{(n-\nu)!} K, \quad M'_\nu \leq \frac{(3\varepsilon)^{n-\nu}}{(n-\nu)!} K \quad (\nu = 0, 1, \dots, n-1). \quad (17.13)$$

7. Putting this into (17.12), we have for $|z| \leq \varepsilon$

$$|L^*(z)| \leq \sum_{\nu=0}^{n-1} \frac{(2\varepsilon)^\nu}{\nu!} \frac{\varepsilon^{n-\nu}}{(n-\nu)!} K = \frac{K}{n!} \varepsilon^n \sum_{\nu=0}^{n-1} \binom{n}{\nu} 2^\nu < \frac{(3\varepsilon)^n}{n!} K \quad (17.14)$$

and for $|z| \leq 3\varepsilon$

$$|L^*(z)| \leq \sum_{\nu=0}^{n-1} \frac{(4\varepsilon)^\nu}{\nu!} \frac{(3\varepsilon)^{n-\nu}}{(n-\nu)!} K = \frac{K\varepsilon^n}{n!} \sum_{\nu=0}^{n-1} \binom{n}{\nu} 4^\nu 3^{n-\nu} < \frac{(7\varepsilon)^n}{n!} K. \quad (17.15)$$

If we use the last inequality (17.7), the relations (17.14) and (17.15) become, since $n > p \geq 1$,

$$\begin{aligned} |L^*(z)| &\leq 3^n \varepsilon^p \cdot \frac{|a_p|}{7^n} < \frac{1}{2^{n-1}} \left(\frac{|a_p|}{2} \varepsilon^p \right) \\ &\leq \frac{|a_p|}{2} \varepsilon^p \quad (|z| = \varepsilon), \end{aligned} \tag{17.16}$$

$$\begin{aligned} |L^*(z)| &\leq 7^n \varepsilon^p \frac{|a_p|}{7^n} = 2 \left(\frac{1}{3} \right)^p \left(\frac{|a_p|}{2} (3\varepsilon)^p \right) \\ &< \frac{|a_p|}{2} (3\varepsilon)^p \quad (|z| = 3\varepsilon). \end{aligned} \tag{17.17}$$

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,

$$K_{p+1} = \max_{|z-\zeta|=r} |f^{(p+1)}(z)|, \quad K_{n+1} = \max_{|z-\zeta|=r} |f^{(n+1)}(z)|, \tag{17.18}$$

$$\gamma = \frac{p! f^{(n)}(\zeta)}{n! f^{(p)}(\zeta)}, \tag{17.19}$$

$$C_1 = \frac{K_{p+1}}{(p+1)|f^{(p)}(\zeta)|}, \quad C_2 = \frac{2K_{n+1}}{|f^{(n)}(\zeta)|}, \tag{17.20}$$

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*

$$\varepsilon C_1 < 1. \tag{17.21}$$

Then we have, denoting any zero of $L(z)$ situated in (17.2) by z_{n+1} ,

$$(z_{n+1} - \zeta)^p = \gamma \prod_{\nu=1}^n (z_{n+1} - z_\nu) \frac{1 + \theta_2 \varepsilon C_2}{1 + \theta_1 \varepsilon C_1}, \quad |\theta_1| \leq 1, \quad |\theta_2| \leq 1. \quad (17.22)$$

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 nonvanishing 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 (1.23c).

Since $L(z_{n+1}) = 0$, we have, using (1.23c),

$$f(z_{n+1}) = R(z_{n+1}) = \prod_{\nu=1}^n (z_{n+1} - z_\nu) \left(\frac{f^{(n)}(0)}{n!} + \theta_2 2\varepsilon \frac{K_{n+1}}{n!} \right), \quad |\theta_2| \leq 1.$$

Using the definition of C_2 in (17.20), we have

$$\frac{f(z_{n+1})}{\prod_{\nu=1}^n (z_{n+1} - z_\nu)} = \frac{f^{(n)}(0)}{n!} (1 + \theta_2 \varepsilon C_2). \quad (17.24)$$

From (17.23) and (17.24) the assertion (17.22) follows immediately, using (17.19) and (17.5).

Approximation of Equations by Algebraic Equations of a Given Degree. Asymptotic Errors for Simple Roots

1. Under the conditions of Theorems 17.1 and 17.2, denote by z_{n+1} that of the roots 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_1, \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 $\nu = 0, 1, \dots$, from $z_{\nu+1}, \dots, z_{\nu+n}$ the $z_{\nu+n+1}$. 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*

$$\delta_0 \equiv C\varepsilon < 1, \quad C_1\varepsilon \leq \frac{1}{2}, \quad (18.2)$$

the sequence z_ν can be formed indefinitely and is convergent to ζ in such a way that

$$\sqrt[n]{|z_\nu - \zeta|} \rightarrow 0 \quad (\nu \rightarrow \infty). \quad (18.3)$$

2. Proof. Put

$$\xi_\nu = C|z_\nu - \zeta| \quad (\nu = 1, 2, \dots). \quad (18.4)$$

Without loss of generality we can and will assume $\zeta = 0$.

From (17.22) and (17.7) we have then

$$|z_{n+1}|^p < |\gamma|(2\varepsilon)^n \cdot \frac{1 + rC_2}{1/2} < C^{n-p} \cdot \varepsilon^n = C^{-p}\delta_0^n,$$

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

$$\xi_{n+1}, \xi_{n+2}, \dots, \xi_{2n} \leq \delta_0^{n/p}. \quad (18.5)$$

3. The corresponding z_{n+1}, \dots, z_{2n} have the distances from the origin

$$\leq C^{-1} \delta_0^{n/p} = \varepsilon \delta_0^{(n-p)/p} \equiv \varepsilon_1. \quad (18.6)$$

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 $\xi_\nu \rightarrow 0$, $z_\nu \rightarrow 0$. Further, for $\nu \rightarrow \infty$, $\nu = kn + \kappa$, $\kappa = 1, 2, \dots, n$,

$$\frac{1}{\nu} \ln \xi_\nu \leq \frac{1}{(k+1)n} \left(\frac{n}{p} \right)^k \ln \delta_0 \rightarrow -\infty, \quad \sqrt[p]{Cz_\nu} \rightarrow 0,$$

and (18.3) is proved.

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

$$|z_{n+1} - \zeta|^p \leq |\gamma| \prod_{\nu=1}^n |z_{\nu+1} - z_\nu| \frac{1+rC_2}{1/2} < 2^{-n} C^{n-p} \prod_{\nu=1}^n |z_{\nu+1} - z_\nu|.$$

Multiplying this on both sides by C^p and using (18.4) we get

$$\xi_{n+1}^p \leq 2^{-n} \prod_{\nu=1}^n (\xi_{\nu+1} + \xi_\nu). \quad (18.10)$$

We will assume from now on that none of the ξ_ν 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 $\min(\xi_1, \dots, \xi_n) = u$. 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

$$\xi_{n+1} \leq 2^{-n} \prod_{\nu=1}^n (2\xi_\nu) = \prod_{\nu=1}^n \xi_\nu. \quad (18.11)$$

6. Since the ξ_ν tend to 0 and are < 1 by (18.2), we see that, for $\nu \geq 1$,

$$\xi_{\nu+n} \leq \xi_{\nu+n-1} \xi_{\nu+n-2} \dots \xi_\nu \quad (\nu \geq 1), \quad (18.12)$$

$$\xi_{\nu+1} < \xi_\nu \quad (\nu \geq n), \quad \frac{\xi_{\nu+1}}{\xi_\nu} \rightarrow 0 \quad (\nu \rightarrow \infty). \quad (18.13)$$

Assume now that $\zeta = 0$ and apply (17.22) with $p = 1$, $\zeta = 0$ to $z_\nu, z_{\nu+1}, \dots, z_{\nu+n}$, $\nu \geq n$. Since we can, by (18.13), replace here ε by $|z_\nu|$, we have

$$z_{\nu+n} = \gamma \prod_{\kappa=0}^{n-1} (z_{\nu+n} - z_{\nu+\kappa}) \frac{1 + \theta_2 z_\nu C_2}{1 + \theta_1 z_\nu C_1},$$

$$\frac{z_{\nu+n}}{\prod_{\kappa=0}^{n-1} z_{\nu+\kappa}} = (-1)^n \gamma \prod_{\kappa=0}^{n-1} \left(1 - \frac{z_{\nu+n}}{z_{\nu+\kappa}}\right) \frac{1 + \theta_2 z_\nu C_2}{1 + \theta_1 z_\nu C_1}. \quad (18.14)$$

By (18.12) and (18.13) we have for the single factors in the right-hand product in (18.14)

$$\left| \frac{z_{\nu+n}}{z_{\nu+\kappa}} \right| < |z_\nu|, \quad 1 - \frac{z_{\nu+n}}{z_{\nu+\kappa}} = 1 + O(z_\nu).$$

We have therefore from (18.14)

$$z_{\nu+n} = (-1)^n (\gamma + O(z_\nu)) \prod_{\kappa=0}^{n-1} z_{\nu+\kappa}. \quad (18.15)$$

7. Take the moduli on both sides of (18.15) and put

$$\ln \frac{1}{|z_\nu|} = y_\nu \quad (\nu \geq 1), \quad \ln |\gamma| = (n - 1)\beta. \quad (18.16)$$

Then we obtain

$$y_{\nu+n} - \sum_{\kappa=0}^{n-1} y_{\nu+\kappa} = (1 - n)\beta + k_{\nu+n}, \quad k_{\nu+n} = O(z_\nu). \quad (18.17)$$

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,

$$y_\nu = \beta + \alpha \mu_n^\nu + O(q_n^\nu)$$

and, going back to (18.16) and replacing z_ν by $z_\nu - \zeta$, similarly to (13.36) and (13.37),

$$|z_\nu - \zeta| = \exp(-\beta) \exp(-\alpha \mu_n^\nu) (1 + O(q_n^\nu)), \quad (18.18)$$

$$\frac{z_{\nu+1} - \zeta}{|z_\nu - \zeta|^{\mu_n}} = \exp[(\mu_n - 1)\beta] + O(q_n^\nu), \quad (18.19)$$

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_ν in our case is of the same type as that of the x_ν in Chapter 13. But here we have to solve an algebraic equation of the degree n at every step, while in Chapter 13 each new approximation is obtained by rational operations. On the other hand, the 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 .

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

$$|\xi|_p \equiv (|x_1|^p + \dots + |x_n|^p)^{1/p} \quad (p \geq 1). \quad (19.2)$$

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 $|\xi|_\infty$:

$$|\xi|_\infty \equiv \max_v |x_v|. \quad (19.3)$$

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*.

Clearly $|c\xi|_p = |c| |\xi|_p$ for any constant c . We shall now prove for $p = 1$ and $p = \infty$

$$|\xi + \eta|_p \leq |\xi|_p + |\eta|_p \quad (p = 1, \infty). \quad \dagger \quad (19.4)$$

Indeed, let $\eta = (y_1, \dots, y_n)$; then we have

$$\begin{aligned} |\xi + \eta|_1 &= |x_1 + y_1| + \dots + |x_n + y_n| \leq |x_1| + \dots + |x_n| \\ &\quad + |y_1| + \dots + |y_n| = |\xi|_1 + |\eta|_1, \end{aligned}$$

$$|\xi + \eta|_\infty = \max_v |x_v + y_v| \leq \max_v |x_v| + \max_v |y_v| = |\xi|_\infty + |\eta|_\infty.$$

* Both names will be used in the following discussion.

† Relation (19.4) is true for all $p \geq 1$, but we need it here only for $p = 1$ and $p = \infty$.

2. Let $A = (a_{ij})$ denote an $n \times n$ matrix. We use $|A|$ to denote the determinant of A and ξ' to denote the transpose of ξ , i.e., the corresponding column vector. We write

$$A\xi' = \eta', \quad (19.5)$$

where the components of $\eta = (y_1, \dots, y_n)$ are given by

$$y_i = \sum_{j=1}^n a_{ij}x_j \quad (i = 1, \dots, n). \quad (19.6)$$

From (19.6) we have

$$|y_i| \leq \sum_{j=1}^n |a_{ij}| |x_j| \leq \sum_{j=1}^n |a_{ij}| |\xi|_\infty \quad (i = 1, \dots, n). \quad (19.7)$$

If we introduce a measure of the “size of A ” by

$$|A|_\infty \equiv \max_i \sum_{j=1}^n |a_{ij}|, \quad (19.8)$$

we have from (19.7)

$$|y_i| \leq |A|_\infty |\xi|_\infty \quad (i = 1, \dots, n),$$

and therefore

$$|\eta|_\infty \leq |A|_\infty |\xi|_\infty. \quad (19.9)$$

3. We now show that in (19.9) $|A|_\infty$ cannot be replaced by a smaller constant; i.e., for each A it is possible to construct a ξ so that the equality in (19.9) holds. Without loss of generality, let $A \neq 0$. Let

$$|A|_\infty = |a_{m1}| + |a_{m2}| + \dots + |a_{mn}|.$$

We then define ξ by

$$x_v = \begin{cases} \frac{|a_{mv}|}{a_{mv}}, & a_{mv} \neq 0 \\ 0, & a_{mv} = 0 \end{cases} \quad (v = 1, \dots, n). \quad (19.10)$$

Then from (19.6)

$$y_m = a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = \sum_{v=1}^n |a_{mv}| = |A|_\infty$$

and hence

$$|\eta|_\infty \geq |A|_\infty. \quad (19.11)$$

Now, not all x_v ($v = 1, \dots, n$) are zero, for otherwise $A = 0$. Hence $|\xi|_\infty = 1$ and from (19.9) and (19.11) we have $|\eta|_\infty = |A|_\infty |\xi|_\infty$.

4. In order to obtain the corresponding relations for $p = 1$, observe that from (19.6) it follows that

$$|y_i| \leq \sum_{j=1}^n |a_{ij}| |x_j| \quad (i = 1, \dots, n),$$

and by summation

$$|\eta|_1 \leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}| |x_j| = \sum_{j=1}^n \left(\sum_{i=1}^n |a_{ij}| \right) |x_j|. \quad (19.12)$$

Putting

$$t_j = \sum_{i=1}^n |a_{ij}| \quad (j = 1, \dots, n),$$

we can write (19.12) in the form

$$|\eta|_1 \leq \sum_{j=1}^n t_j |x_j|. \quad (19.13)$$

If we now introduce another measure of the “size of A ” by

$$|A|_1 \equiv \max_j \sum_{i=1}^n |a_{ij}| = \max_j t_j, \quad (19.14)$$

we have

$$\begin{aligned} |\eta|_1 &\leq |A|_1 \sum_{j=1}^n |x_j|, \\ |\eta|_1 &\leq |A|_1 |\xi|_1. \end{aligned} \quad (19.15)$$

As before, we show that in (19.15) $|A|_1$ 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

$$|\eta|_1 = |A|_1 |\xi|_1.$$

5. We now deduce some properties of $|A|_p$ ($p = 1, \infty$). Clearly, for any constant c ,

$$|cA|_p = |c||A|_p \quad (p = 1, \infty). \quad (19.16)$$

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 (19.9) and (19.15) we have

$$\begin{aligned} |\eta'_1|_p &\leq |A|_p |\xi|_p, & |\eta'_2|_p &\leq |B|_p |\xi|_p \quad (p = 1, \infty), \\ \eta'_1 + \eta'_2 &= (A + B)\xi', \\ |\eta'_1 + \eta'_2|_p &\leq |A + B|_p |\xi|_p \quad (p = 1, \infty). \end{aligned} \quad (19.17)$$

We choose $\xi \neq 0$ so that equality holds in (19.17). Then from (19.4) 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

$$|A + B|_p \leq |A|_p + |B|_p \quad (p = 1, \infty). \quad (19.18)$$

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 $|\eta'|_p = |AB|_p |\xi|_p$,

$$|AB|_p \leq |A|_p |B|_p \quad (p = 1, \infty). \quad (19.19)$$

6. 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

$$|\lambda I - A| = 0 \quad (19.20)$$

are called the *fundamental* (or *characteristic*) *roots* or *eigenvalues* of A and correspondingly (19.20) is called the *fundamental* (or *characteristic*) *equation* of A . We introduce the following notation:

$$\lambda_A = \max_v |\lambda_v| \quad (v = 1, \dots, n). \quad (19.21)$$

Notice that if λ_v is a root of (19.20), then we can find a vector $\xi_v \neq 0$ which is a solution of the system

$$(\lambda_v I - A)\xi_v' = 0, \quad \text{i.e., } A\xi_v' = \lambda_v \xi_v'. \quad (19.22)$$

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*

$$\lambda_A \leqq |A|_p \quad (p = 1, \infty).$$

Proof. Let m be such that $|\lambda_m| = \lambda_A$ and let ξ_m be a fundamental vector corresponding to λ_m . Then

$$A\xi_m' = \lambda_m \xi_m', \quad |\lambda_m| |\xi_m|_p \leqq |A|_p |\xi_m|_p,$$

and, since $|\xi_m|_p \neq 0$, $\lambda_A = |\lambda_m| \leqq |A|_p$, Q.E.D.

7. Let S be a *regular* $n \times n$ matrix, i.e., one such that $|S| \neq 0$. 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 $(\lambda I - SAS^{-1}) = S(\lambda I - A)S^{-1}$,

$$|\lambda I - SAS^{-1}| = |S(\lambda I - A)S^{-1}| = |\lambda I - A|.$$

Remark. If we transform A by S , λ_A does not change, but $|A|_p$ may change.

8. 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*

$$SAS^{-1} = D + J, \quad (19.23)$$

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, (19.23) can be written as

$$SAS^{-1} = \begin{pmatrix} U_1 & & & 0 \\ & U_2 & & \\ & & \ddots & \\ 0 & & & U_m \end{pmatrix} \quad (19.24)$$

$$U_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & & \\ & & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{pmatrix} \quad (19.25)$$

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 (19.23) is reduced to the "diagonal form." This may also happen if the fundamental roots are multiple.

9. We have obviously

$$|J|_p \leq 1 \quad (p = 1, \infty). \quad (19.26)$$

For any $\varepsilon \neq 0$ it follows from (19.23), applied to $(1/\varepsilon)A$, that

$$\begin{aligned} S \frac{1}{\varepsilon} AS^{-1} &= \frac{1}{\varepsilon} D + J, \\ SAS^{-1} &= D + \varepsilon J. \end{aligned} \quad (19.27)$$

Hence we see that in Jordan's canonical form the value *one* of the nonvanishing elements of J is not essential; it can be replaced by any number $\varepsilon \neq 0$. Now by (19.26)

$$|\varepsilon J|_p = |\varepsilon| |J|_p \leq |\varepsilon|. \quad (19.28)$$

Theorem 19.3. Given an $\varepsilon > 0$, there exists an S such that

$$\lambda_A \leq |SAS^{-1}|_p \leq \lambda_A + \varepsilon \quad (p = 1, \infty). \quad (19.29)$$

Indeed, from Theorem 19.1 and (19.28) we have

$$\lambda_A \leq |SAS^{-1}|_p \leq |D|_p + |\varepsilon J|_p \leq \lambda_A + \varepsilon.$$

As the fundamental roots of the matrix A are roots of the algebraic equation (19.20), 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 (19.20) in terms of the variations of the elements of A . We give such estimates in Appendix K.

10. If $\xi = (x_\nu)$ and $\eta = (y_\nu)$ are two n -dimensional vectors, we call the expression

$$\bar{x}_1 y_1 + \dots + \bar{x}_n y_n = (\xi, \eta), \quad (19.30)$$

the inner product of ξ and η (in this order) and denote it by the symbol (ξ, η) or also by the symbol $\bar{\xi} \eta_c$, 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

$$|(\xi, \eta)| \leq \sqrt{\sum_{\nu=1}^n |x_\nu|^2} \sqrt{\sum_{\nu=1}^n |y_\nu|^2} \quad (19.31)$$

or, what amounts to the same,

$$\left| \sum_{\nu=1}^n x_\nu y_\nu \right| \leq \sqrt{\sum_{\nu=1}^n |x_\nu|^2} \sqrt{\sum_{\nu=1}^n |y_\nu|^2}. \quad (19.32)$$

11. If the equation (19.20) 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

$$\sum_{\nu=1}^n \lambda_\nu = \sum_{\nu=1}^n a_{\nu\nu}. \quad (19.33)$$

The expression on the right is called the trace of the matrix A and is sometimes denoted by $\text{tr}(A)$.

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 $\phi(z)$, the eigenvalues of $\phi(A)$ are given by the expressions

$$\phi(\lambda_1), \phi(\lambda_2), \dots, \phi(\lambda_n),$$

provided all these numbers are finite, that is, that the denominator of $\phi(z)$ 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} , $1/\lambda_v$.

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)

$$|\prod_m|_\infty \leq \sigma(\lambda_A + \varepsilon)^m \quad (m = 1, 2, \dots). \quad (20.3)$$

If in particular $\lambda_A < 1$ 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

$$\sigma = |S|_\infty |S^{-1}|_\infty, \quad \eta_1 = \frac{\varepsilon}{2\sigma}, \quad (20.6)$$

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

$$\begin{aligned}
 S \prod_m S^{-1} &= [S(A + U_1)S^{-1}] \dots [S(A + U_m)S^{-1}] \\
 &= [SAS^{-1} + SU_1 S^{-1}] [SAS^{-1} + SU_2 S^{-1}] \\
 &\quad \dots [SAS^{-1} + SU_m S^{-1}], \\
 S \prod_m S^{-1} &= \prod_{\mu=1}^m (B + V_\mu).
 \end{aligned} \tag{20.8}$$

From (20.1) and (20.5)–(20.8) it follows by (19.19) that

$$\begin{aligned}
 |V_\mu|_\infty &\leq \sigma \eta_1 = \frac{\varepsilon}{2}, \quad |B + V_\mu|_\infty \leq s + \sigma \eta_1 < \lambda_A + \varepsilon, \\
 |S \prod_m S^{-1}|_\infty &\leq (\lambda_A + \varepsilon)^m.
 \end{aligned} \tag{20.9}$$

Hence, since $|\Pi_m|_\infty = |S^{-1}(S \prod_m S^{-1})S|_\infty$, it follows from (20.9) and (20.6) that

$$|\prod_m|_\infty \leq \sigma(\lambda_A + \varepsilon)^m,$$

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:*

$$A_1 = A, \quad A_{\mu+1} = A_\mu A + W_\mu. \tag{20.10}$$

Let $\varepsilon > 0$ be such that $\lambda_A + \varepsilon < 1$. There exist two positive constants $\eta_2 > 0$, $\sigma > 0$ such that, if

$$|W_\mu|_\infty \leq \eta_2 |A_\mu|_\infty \quad (\mu = 1, 2, \dots), \tag{20.11}$$

then

$$|A_m|_\infty < \sigma(\lambda_A + \varepsilon)^m \quad (m = 1, 2, \dots), \quad A_m \rightarrow 0 \quad (m \rightarrow \infty). \tag{20.12}$$

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

$$\eta_2 = \frac{\varepsilon}{2\sigma^2}. \tag{20.13}$$

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 $A_\mu = S^{-1}B_\mu S$,

$$|A_\mu|_\infty \leq \sigma |B_\mu|_\infty \quad (\mu = 1, 2, \dots), \quad (20.17)$$

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.

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 ($v = 1, 2, \dots$)*

$$\prod_m \equiv \prod_{v=1}^m (A + U_v). \quad (21.1)$$

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

$$(\lambda_A - \varepsilon)^{-m} |\prod_m \zeta'|_1 \rightarrow \infty \quad (m \rightarrow \infty), \quad (21.2)$$

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

$$\lambda_A = |\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_k| = s > 1 \geq |\lambda_{k+1}| \geq \dots \geq |\lambda_n|. * \quad (21.3)$$

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 (19.27):

$$SAS^{-1} = B = D + \tau J. \quad (21.5)$$

* If $k = n$, the expression σ_m defined by (21.18) is $= 0$, but the whole discussion remains valid.

If we introduce

$$\sigma = |S|_1 |S^{-1}|_1, \quad \delta = \frac{\eta - \tau}{\sigma} \quad (21.6)$$

and put

$$V_\nu = SU_\nu S^{-1}, \quad W_\nu = V_\nu + \tau J, \quad (21.7)$$

we have from (21.5)

$$B + V_\nu = D + W_\nu. \quad (21.8)$$

3. But then we obviously have

$$\begin{aligned} S(A + U_\nu)S^{-1} &= SAS^{-1} + SU_\nu S^{-1} = B + V_\nu, \\ S \left[\prod_{\nu=1}^m (A + U_\nu) \right] S^{-1} &= \prod_{\nu=1}^m (D + W_\nu). \end{aligned} \quad (21.9)$$

Hence we have replaced A by the simpler matrix D . Now from (21.6) and the hypothesis $|U_\nu|_1 \leq \delta$ follows

$$|V_\nu|_1 \leq |S|_1 |U_\nu|_1 |S^{-1}|_1 \leq \sigma \delta = \eta - \tau$$

and from (21.7)

$$|W_\nu|_1 \leq \eta. \quad (21.10)$$

On the other hand, for any vector ζ ,

$$S \left[\prod_{\nu=1}^m (A + U_\nu) \right] S^{-1} (S\zeta') = \left[\prod_{\nu=1}^m (D + W_\nu) \right] S\zeta', \quad (21.11)$$

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

$$\left| \prod_{\nu=1}^m (D + W_\nu) \xi_0' \right|_1 \rightarrow \infty \quad (m \rightarrow \infty). \quad (21.13)$$

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 such a vector ξ the vector $\zeta' = S^{-1}\xi'$ satisfies the relation

$$|\prod_m \zeta'|_1 \rightarrow \infty$$

and ζ lies in the solid angle obtained from (21.12) 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 $|X_\nu|_1 \leq \delta_\epsilon$, we have for any vector ξ from L_ϵ

$$(\lambda_A - \epsilon)^{-m} \left| \prod_m \zeta' \right|_1 = \left| \prod_{\nu=1}^m (C + X_\nu) \zeta' \right|_1 \rightarrow \infty,$$

and this holds as long as

$$\underbrace{|U_\nu|_1}_{\text{---}} \leq (\lambda_A - \epsilon) \delta_\epsilon = \delta(A, \epsilon) \quad (\nu = 1, 2, \dots).$$

Our theorem is proved.

8. Corollary. *Under the conditions of Theorem 21.1, we have obviously*

$$(\lambda_A - \epsilon)^{-m} \left| \prod_m \right|_1 \rightarrow \infty \quad (m \rightarrow \infty). \quad (21.28)$$

Characterization of Points of Attraction and Repulsion for Iterations with Several Variables

1. Let $\xi = (x_1, \dots, x_n)$ be a point in the n -dimensional real or complex space S and put

$$y_i = f_i(x_1, \dots, x_n) \equiv f_i(\xi) \quad (i = 1, \dots, n). \quad (22.1)$$

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

$$A = J(\zeta) = \left(\frac{\partial(f_i(\zeta))}{\partial(z_j)} \right). \quad (22.7)$$

2. Theorem 22.1. Assume that $f_i(\xi)$ ($i = 1, \dots, n$) are totally differentiable at the center ζ . Further, assume that for the matrix (22.7) we have (cf. Chapter 19, Section 6)

$$\lambda_A < 1. \quad (22.8)$$

Then ζ is a point of attraction, i.e., there exist two neighborhoods V, V_0 of ζ such that if $\xi_0 \prec V$ and the sequence ξ_κ is obtained by (22.5), we have $\xi_\kappa \prec V_0$ ($\kappa = 1, 2, \dots$), and

$$\xi_\kappa \rightarrow \zeta \quad (\kappa \rightarrow \infty). \quad (22.9)$$

3. Proof. Without loss of generality, we can assume that ζ 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 $\xi_0 \prec V$, $\xi_1, \dots, \xi_k \prec V_0$. 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)

$$\xi_{k+1} \prec \frac{\sigma}{1+\sigma} V_0 \prec V_0;$$

our assumptions hold for ξ_{k+1} too, and consequently for all ξ_v . It follows now that for all κ , if $\xi_0 \prec V$,

$$\xi'_{\kappa+1} = \prod_{v=0}^{\kappa} [A + U(\xi_v)]\xi'_0 \quad (\kappa = 0, 1, \dots),$$

and since $\lambda_A < 1$, 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 $\xi_0 \prec LV$ (region common to L and V) the sequence ξ_ν 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 $\xi \prec V$ we have

$$|u_{ij}(\xi)| < \frac{\delta}{n+1} \quad (i, j = 1, \dots, n) \quad (22.21)$$

Suppose now that for a point ξ_0 from LV all points ξ_ν in (22.5) stay in V and that none of them lies in the origin. If we define $U(\xi_\nu)$ and $A(\xi_\nu)$ by (22.11) and (22.12) we have by (22.21)

$$|U(\xi_\nu)|_1 < \delta \quad (\nu = 0, 1, \dots),$$

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

$$\lambda_{A(\xi)} < 1 \quad (22.22)$$

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_n 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:

$$x' = \frac{x}{2} (1 - (x^2 + y^2)), \quad (22.26)$$

$$y' = \frac{y}{2} (3 + (x^2 + y^2)).$$

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 LV , and convergence for every starting

point ξ_0 from V outside of L . From (22.25) we have, putting $z = re^{i\theta}$, and $z' = r'e^{i\theta'}$,

$$|z'|^2 = z'\bar{z}' = r^2(1 + \frac{1}{4}(1 + r^2)^2 - (1 + r^2)\cos 2\theta),$$

$$\frac{|z'|^2}{|z|^2} = 1 + \frac{1}{4}(1 + r^2)^2 - (1 + r^2)\cos 2\theta, \quad (22.28)$$

and further

$$\tan \theta' = \frac{y'}{x'} = \frac{3y + yr^2}{x(1 - r^2)} = \frac{3 + r^2}{1 - r^2} \tan \theta. \quad (22.29)$$

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

$$\cos(2 \arg z_\mu) = \frac{1 - t_\mu^2}{1 + t_\mu^2} \rightarrow -1.$$

But now it follows from (22.28), since $\cos 2\theta \rightarrow -1$, that

$$\liminf \frac{|z_{\mu+1}|^2}{|z_\mu|^2} = \liminf (1 + \frac{1}{2}(1 + |z_\mu|^2)^2) \geq 9/4,$$

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

$$0 < \arg z < \pi, \quad \pi < \arg z < 2\pi.$$

TRIANGLE INEQUALITY

1. While in Chapter 19 we dealt only with the vector norm $|\xi|_p$ for $p = 1$ and ∞ , from now on we shall need also the case $p = 2$.

Again, as in Chapter 19, let $A = (a_{ij})$, $B = (b_{ij})$, denote $(n \times n)$ matrices and put

$$\Delta_2(A) = \left(\sum_{i,j=1}^n |a_{ij}|^2 \right)^{1/2}, \quad \Delta_1(A) = \sum_{i,j} |a_{ij}|, \quad \Delta_\infty(A) = \max_{i,j} |a_{ij}|. \quad (23.1)$$

For a row vector (19.1) we have, according to (19.2), the definition

$$|\xi|_2 = (x_1^2 + x_2^2 + \dots + x_n^2)^{1/2}. \quad (23.2)$$

This is the so-called *Euclidean norm* of ξ . With this notation, (19.31) can be written as

$$|(\xi, \eta)| \leq |\xi|_2 |\eta|_2. \quad (23.3)$$

We have now to prove that the norm $|\xi|_2$ also satisfies the relation (10.4), the so-called *triangle inequality*. This follows at once from the general inequality

$$\sqrt{\sum_{\kappa=1}^k |u_\kappa + v_\kappa|^2} \leq \sqrt{\sum_{\kappa=1}^k |u_\kappa|^2} + \sqrt{\sum_{\kappa=1}^k |v_\kappa|^2}. \quad (23.4)$$

To prove (23.4), we square it:

$$\sum_{\kappa=1}^k |u_\kappa + v_\kappa|^2 \leq \sum_{\kappa=1}^k |u_\kappa|^2 + \sum_{\kappa=1}^k |v_\kappa|^2 + 2 \sqrt{\sum_{\kappa=1}^k |u_\kappa|^2} \sqrt{\sum_{\kappa=1}^k |v_\kappa|^2}.$$

This becomes, using the identity

$$|u_\kappa + v_\kappa|^2 = (u_\kappa + v_\kappa)(\bar{u}_\kappa + \bar{v}_\kappa) = |u_\kappa|^2 + |v_\kappa|^2 + u_\kappa \bar{v}_\kappa + \bar{u}_\kappa v_\kappa,$$

$$\sum_{\kappa=1}^k u_\kappa \bar{v}_\kappa + \sum_{\kappa=1}^k \bar{u}_\kappa v_\kappa \leq 2 \sqrt{\sum_{\kappa=1}^k |u_\kappa|^2} \sqrt{\sum_{\kappa=1}^k |v_\kappa|^2},$$

and this follows now immediately by the Cauchy-Schwarz inequality.

2. We have from (19.5) and (19.6) by the Cauchy-Schwarz inequality

$$|y_i|^2 \leq \sum_{j=1}^n |a_{ij}|^2 \sum_{j=1}^n |x_j|^2$$

and, summing this over i ,

$$\sum_{i=1}^n |y_i|^2 \leq \sum_{i,j=1}^n |a_{ij}|^2 \sum_{j=1}^n |x_j|^2.$$

This can be written by (23.1) and (23.2) as

$$|\eta|_2 \leq \Delta_2(A)|\xi|_2, \quad (23.5)$$

a relation completely analogous to (19.9) and (19.15). However, here it is no longer true that $\Delta_2(A)$ in (23.5) cannot be replaced by a smaller constant for a given A . For instance, if $A = I$, we have $\Delta_2(I) = \sqrt{n}$, while in this case $\eta = \xi$.

From (23.1) it follows by the Cauchy-Schwarz inequality, if we put

$$C = (c_{ik}) = AB, \quad c_{ik} = \sum_{\lambda=1}^n a_{i\lambda} b_{\lambda k},$$

that

$$\begin{aligned} \Delta_2^2(C) &= \sum_{i,k} \left| \sum_{\lambda=1}^n a_{i\lambda} b_{\lambda k} \right|^2 \leq \sum_{i,k} \sum_{\nu} |a_{i\nu}|^2 \sum_{\mu} |b_{\mu k}|^2 = \sum_{i,\nu} |a_{i\nu}|^2 \sum_{\mu,k} |b_{\mu k}|^2 \\ &= \Delta_2^2(A) \cdot \Delta_2^2(B), \quad \Delta_2(AB) \leq \Delta_2(A) \cdot \Delta_2(B). \end{aligned} \quad (23.6)$$

3. We have for any constant c immediately the relation

$$\Delta_2(cA) = |c|\Delta_2(A).$$

As to the “triangle inequality”

$$\Delta_2(A + B) \leq \Delta_2(A) + \Delta_2(B) \quad (23.7)$$

corresponding to (19.18), it is equivalent to

$$\sqrt{\sum_{i,j=1}^n |a_{ij} + b_{ij}|^2} \leq \sqrt{\sum_{i,j=1}^n |a_{ij}|^2} + \sqrt{\sum_{i,j=1}^n |b_{ij}|^2}$$

and follows at once from relation (23.4), considering A and B as n^2 -dimensional vectors.

BILINEAR AND QUADRATIC FORMS OF SYMMETRIC MATRICES

4. If $\xi = (x_\nu)$ and $\eta = (y_\nu)$ are two n -dimensional vectors, the expression

$$(\xi, A\eta) = \bar{\xi}_i A_{ij} \eta_j = \sum_{\mu, \nu=1}^n \bar{x}_\mu a_{\mu\nu} y_\nu$$

is the *bilinear form* corresponding to the matrix A . We have then from (23.3) and (23.5)

$$|\xi, A\eta| \leq \Lambda_2(A) |\xi|_2 |\eta|_2. \quad (23.8)$$

If an $(n \times n)$ matrix $A(a_{\mu\nu})$ is real and *symmetric*, that is, if $a_{\mu\nu} = a_{\nu\mu}$ ($\mu, \nu = 1, \dots, n$), then, as is well known, all eigenvalues of A are real. The greatest and the smallest of these eigenvalues will be denoted, respectively, by $\Lambda(A)$ and $\lambda(A)$.

To such a matrix A corresponds a *quadratic form* in the components of a real vector $\xi(x_\nu)$, defined by $A(\xi) = (\xi, A\xi')$:

$$A(\xi) = \sum_{\mu, \nu=1}^n a_{\mu\nu} x_\mu x_\nu. \quad (23.9)$$

Then, according to a familiar result from the theory of quadratic forms, $\Lambda(A)$ and $\lambda(A)$ are respectively the *maximum* and the *minimum* of $A(\xi)$ on the unit hypersphere $|\xi|_2 = 1$:

$$\Lambda(A) = \max A(\xi), \quad \lambda(A) = \min A(\xi) \quad (|\xi|_2 = 1). \quad (23.10)$$

It follows that

$$\Lambda(A)|\xi|_2^2 \geq A(\xi) \geq \lambda(A)|\xi|_2^2. \quad (23.11)$$

5. If $\lambda(A) > 0$, the form $A(\xi)$ is called *positive definite*, while if $\lambda(A) = 0$, $A(\xi)$ is *positive semidefinite*. A positive definite form $A(\xi)$ has for all $\xi \neq 0$ positive values $\geq \lambda(A)|\xi|_2^2$. A positive semidefinite form has for all ξ nonnegative values, but then $A(\xi)$ can also be equal to 0 for some $\xi \neq 0$.

If A is a symmetric matrix with the eigenvalues $\lambda_1, \dots, \lambda_n$, it follows from Chapter 19, Section 11, that A^2 has the eigenvalues $\lambda_1^2, \dots, \lambda_n^2$. On the other hand, since A is symmetric, the diagonal elements of A^2 are

$$\sum_{\nu=1}^n a_{1\nu}^2, \quad \sum_{\nu=1}^n a_{2\nu}^2, \dots, \quad \sum_{\nu=1}^n a_{n\nu}^2.$$

Applying (19.33) to A^2 we see that $\sum_{\nu=1}^n \lambda_\nu^2 = \sum_{\mu, \nu=1}^n a_{\mu\nu}^2 = \Delta_2^2(A)$.

Since $\Lambda(A)$ is the greatest of the λ_ν , we have

$$\Delta_2(A) \leq \sqrt{n}\Lambda(A). \quad (23.12)$$

Consider again the relation (19.5) assuming now A as a real symmetric matrix and ξ as a real vector. Squaring (19.6) and summing over i we get

$$|\eta|^2 = \sum_{i=1}^n y_i^2 = \sum_{j, k=1}^n \sum_{i=1}^n a_{ij} a_{ik} x_j x_k = \sum_{j, k=1}^n a_{jk}^* x_j x_k,$$

where $a_{jk}^* = \sum_{i=1}^n a_{ij} a_{ik} = \sum_{i=1}^n a_{ji} a_{ik}$ is the general element of the matrix $(a_{jk}^*) = A^2$. It follows now from (23.11) that we have $|\eta|_2^2 \leq \Lambda(A^2)|\xi|_2^2 = \max(\lambda(A)^2, \Lambda(A)^2)|\xi|_2^2$ and therefore generally for a symmetric matrix A

$$\begin{aligned} |\eta|_2 &= |A\xi|_2 \leq \max(|\lambda(A)|, |\Lambda(A)|)|\xi|_2, \\ |\xi_i A \eta_c|_2 &\leq |\xi_i|_2 |A \eta_c|_2 \leq \max(|\Lambda(A)|, |\lambda(A)|)|\xi|_2 |\eta|_2. \end{aligned} \quad (23.13)$$

ESTIMATE OF $\Delta_p(ABC)$

6. We have from (23.1)

$$\Delta_1(A) = \sum_{i, j=1}^n |a_{ij}|, \quad \Delta_\infty(A) = \max_{i, j} |a_{ij}| \quad (23.14)$$

and have then, replacing in (23.1) $|a_{ij}|^2$ by $|a_{ij}| \Delta_\infty(A)$,

$$\Delta_\infty(A) \leq \Delta_2(A) \leq \Delta_\infty^{1/2}(A) \Delta_1^{1/2}(A) \leq \Delta_1(A) \quad (23.15)$$

and obviously (considering A as an n^2 -dimensional vector), by (19.4),

$$\Delta_p(A + B) \leq \Delta_p(A) + \Delta_p(B) \quad (\phi = 1, \infty). \quad (23.16)$$

The above definitions of $\Delta(A)$ are special cases of the general definition

$$\Delta_\phi(A) = \left(\sum |a_{ij}|^\phi \right)^{1/\phi} \quad (\phi \geq 1). \quad (23.17)$$

From now on we will deal with three couples of indices:

$$p = 1, q = \infty; \quad p = q = 2; \quad p = \infty, q = 1. \quad (23.18)$$

Then we will prove that if $\eta = A\xi$, we have

$$|\eta|_p \leq \Delta_p(A)|\xi|_q \quad (p = 1, 2, \infty). \quad (23.19)$$

For $p = 2$ this is (23.5). For $p = 1$ we have

$$|\eta|_1 = \sum_i \left| \sum_j a_{ij}x_j \right| \leq \sum_i \sum_j |a_{ij}| |\xi|_\infty = \Delta_1(A)|\xi|_\infty,$$

and for $p = \infty$

$$|\eta|_\infty = \max_i \left| \sum_j a_{ij}x_j \right| \leq \max_i \max_j |a_{ij}| \sum_j |x_j| = \Delta_\infty(A)|\xi|_1.$$

If A , B , and $C = (c_{ij})$ are $(n \times n)$ matrices, it is easy to see that we have

$$\Delta_p(ABC) \leq \Delta_p(A) \Delta_q(B) \Delta_p(C). \quad (23.20)$$

Indeed, for $p = 2$ this follows immediately from (23.6). For $p = 1$ we have

$$\begin{aligned} \Delta_1(ABC) &= \sum_{i,j} \left| \sum_{\lambda, \mu} a_{i\lambda} b_{\lambda\mu} c_{\mu j} \right| \leq (\max_{\lambda, \mu} |b_{\lambda\mu}|) \sum_{i,j} |a_{i\lambda}| |c_{\mu j}| \\ &= \Delta_\infty(B) \Delta_1(A) \Delta_1(C), \end{aligned}$$

and for $p = \infty$,

$$\begin{aligned} \Delta_\infty(ABC) &= \max_{\mu, \lambda} \left| \sum_{i,j} a_{i\lambda} b_{\lambda\mu} c_{\mu j} \right| \leq \Delta_\infty(A) \Delta_\infty(C) \max_{i,j} \sum_{\lambda, \mu} |b_{\lambda\mu}| \\ &= \Delta_\infty(A) \Delta_1(B) \Delta_\infty(C). \end{aligned}$$

VARIATION OF $\Delta_p(A^{-1})$

7. From the identity

$$A(B^{-1} - A^{-1})B = A - B$$

it follows that

$$B^{-1} - A^{-1} = A^{-1}(A - B)B^{-1}, \quad (23.21)$$

If both A and B are regular.

Applying (23.20) and (19.19) to the second formula we have

$$\Delta_p(B^{-1} - A^{-1}) \leq \Delta_q(B - A) \Delta_p(A^{-1}) \Delta_p(B^{-1}), \quad (23.22)$$

$$|B^{-1} - A^{-1}|_p \leq |B - A|_p |A^{-1}|_p |B^{-1}|_p \quad (\phi = 1, \infty). \quad (23.23)$$

On the other hand, by the triangle inequality we have $\Delta_p(B^{-1}) \leq \Delta_p(A^{-1}) + \Delta_p(B^{-1} - A^{-1})$ and, since A, B can be interchanged,

$$|\Delta_p(B^{-1}) - \Delta_p(A^{-1})| \leq \Delta_p(B^{-1} - A^{-1}). \quad (23.24)$$

Using now (23.22) we can write

$$\Delta_p(B^{-1}) = \Delta_p(A^{-1}) - \theta \Delta_q(B - A) \Delta_p(A^{-1}) \Delta_p(B^{-1}), \quad |\theta| \leq 1, * \quad (23.25)$$

and, solving this with respect to $\Delta_p(B^{-1})$,

$$\frac{\Delta_p(B^{-1})}{\Delta_p(A^{-1})} = \frac{1}{1 + \theta \Delta_p(A^{-1}) \Delta_q(B - A)}, \quad |\theta| \leq 1. \quad (23.26)$$

Proceeding in the same way with the norms $|A|_p$ we have similarly

$$\frac{|B^{-1}|_p}{|A^{-1}|_p} = \frac{1}{1 + \theta |A^{-1}|_p |B - A|_p}, \quad |\theta| \leq 1. \quad (23.27)$$

8. The formula (23.26) 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 the formulas (23.21)–(23.27) we *assumed* that B^{-1} existed. It is easy, however, to prove the corresponding assertion correctly.

Theorem 23.1. *If we have two $(n \times n)$ matrices A, B , with A assumed as regular, and either*

$$\Delta_p(A^{-1}) \Delta_q(B - A) < 1 \quad (23.28)$$

or

$$|A^{-1}|_p |B - A|_p < 1, \quad (23.29)$$

then B^{-1} exists.

* This relation is easily seen to be equivalent to

$$\left| \frac{1}{\Delta_p(B^{-1})} - \frac{1}{\Delta_p(A^{-1})} \right| \leq \Delta_q(B - A).$$

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 in the case of the assumption (23.28) the formula (23.26) and obtain

$$\Delta_p(B_t^{-1}) \leq \frac{\Delta_p(A^{-1})}{1 - \Delta_p(A^{-1}) \Delta_q(B - A)} \quad (0 \leq t < t_0),$$

and 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 . The completely analogous argument holds in the case of the assumption (23.29).

LENGTH OF ARC IN THE $|\xi|_p$ METRIC

9. The definition of $|\xi|_p$ for $p = 2$ is the usual (Euclidean) length (and distance). We will therefore from now on usually drop the subscript 2 of the vector norm $|\dots|_2$ unless different norms are considered simultaneously. For other values of p we will have to use a new definition of length (and distance) which has many properties in common with the Euclidean concept.

We will need in particular for $p = 1, \infty$ the definition of the length of arc. Assume that we have in the n -dimensional space R_n an arc C expressed in a parameter t :

$$\xi(t) = (x_1(t), \dots, x_n(t)) \quad (t_0 \leq t \leq T),$$

where $x_1(t), \dots, x_n(t)$ have continuous first derivatives in $\langle t_0, T \rangle$. Decompose the interval $\langle t_0, T \rangle$ in $N + 1$ intervals by

$$t_0 < t_1 < \dots < t_N < T = t_{N+1}.$$

Then we are going to prove that, for $p = 1$ or ∞ , as $N \rightarrow \infty$, $\max_v (t_{v+1} - t_v) \rightarrow 0$, we have, as is well known for $p = 2$,

$$\lim_{N \rightarrow \infty} \sum_{v=0}^N |\xi(t_{v+1}) - \xi(t_v)|_p = \int_{t_0}^T |\xi'(t)|_p dt. \quad (23.30)$$

We will call this limit the *p length of the arc C*.

For $p = 1$ we have by the mean value theorem

$$\begin{aligned} \sum_{v=0}^N |\xi(t_{v+1}) - \xi(t_v)|_1 &= \sum_{i=1}^n \left[\sum_{v=0}^N |x_i(t_{v+1}) - x_i(t_v)| \right] \\ &= \sum_{i=1}^n \sum_{v=0}^N |x_i'(\theta_{v(i)})| (t_{v+1} - t_v), \end{aligned}$$

where each $\theta_{v(i)}$ lies in the interval $\langle t_v, t_{v+1} \rangle$.

But here, since the function $|x_i'(t)|$ is continuous, we have by the definition of the integral

$$\sum_{v=0}^N |x_i'(\theta_{v(i)})| (t_{v+1} - t_v) \rightarrow \int_{t_0}^T |x_i'(t)| dt,$$

and the limit in (23.30) is

$$\int_{t_0}^T \sum_i |x_i'(t)| dt = \int_t^T |\xi'(t)|_1 dt.$$

10. For $p = \infty$, putting $F(t) = \max_i |x_i'(t)|$, we have to prove that

$$\sum_{v=0}^N |\xi(t_{v+1}) - \xi(t_v)|_\infty - \int_{t_0}^T F(t) dt \equiv S$$

tends to 0 if $\max_v |t_{v+1} - t_v|$ goes to 0. Denote by δ a positive number arbitrarily small. Then there exists a positive ϵ such that we have for $i = 1, \dots, n$

$$|x_i'(t'') - x_i'(t')| \leq \delta \quad (|t'' - t'| \leq \epsilon, \quad t_0 \leq t' \leq t'' \leq T).$$

We first have to prove that $F(t)$ is *continuous*, that is, that if $u_v \rightarrow u$ and all u_v lie in $\langle t_0, T \rangle$, we have $F(u_v) \rightarrow F(u)$. Indeed, if we consider only the u_v with $|u_v - u| \leq \epsilon$ we have certainly, for $i = 1, \dots, n$,

$$|x_i'(u_v)| \leq F(u) + \delta, \quad F(u_v) \leq F(u) + \delta.$$

On the other hand, if $F(u) = |x_j'(u)|$, we have $F(u) - \delta \leq |x_j'(u_v)|$ and therefore

$$F(u_v) = \max_i |x_i'(u_v)| \geq |x_j'(u_v)| \geq F(u) - \delta,$$

which proves, as δ is arbitrarily small, that $F(u_v) \rightarrow F(u)$.

We assume now that all $t_{v+1} - t_v$ are $\leq \varepsilon$. We can write, by the mean value theorem for integrals,

$$\int_{t_0}^T F(t) dt = \sum_{v=0}^N \int_{t_v}^{t_{v+1}} F(t) dt = \sum_{v=0}^N F(\theta_v)(t_{v+1} - t_v),$$

where each θ_v is a number from the interval $\langle t_v, t_{v+1} \rangle$. On the other hand, applying the mean value theorem to the differences $x_i(t_{v+1}) - x_i(t_v)$, we have

$$|\xi(t_{v+1}) - \xi(t_v)|_\infty = \max_i |x_i'(\theta_v^{(i)})| (t_{v+1} - t_v),$$

where all $\theta_v^{(i)}$ lie in $\langle t_v, t_{v+1} \rangle$. We obtain then for S

$$S = \sum_{v=0}^N (t_{v+1} - t_v) [\max_i |x_i'(\theta_v^{(i)})| - \max_i |x_i'(\theta_v)|].$$

But if we replace here $\theta_v^{(i)}$ by θ_v , each $x_i'(\theta_v^{(i)})$ moves at the most by δ and the same holds for $\max_i |x_i'(\theta_v^{(i)})|$. Therefore, the modulus of the expression in the brackets is $\leq 2\delta$ and $|S| \leq 2\delta(T - t_0)$. This proves (23.30) also for $p = \infty$.

11. Applying the triangle inequality (19.18) repeatedly we have

$$|\xi(T) - \xi(t_0)|_p \leq \sum_{v=0}^N |\xi(t_{v+1}) - \xi(t_v)|_p$$

and therefore, using (23.30) and replacing T by β and t_0 by α ,

$$|\xi(\beta) - \xi(\alpha)|_p \leq \int_{\alpha}^{\beta} |\xi'(t)|_p dt \quad (p = 1, 2, \infty; \quad \alpha \leq \beta). \quad (23.31)$$

FORMULATION OF THE THEOREM

1. Our aim in this chapter will be the generalization of Theorem 2.2 to systems of equations. For this generalization a convenient choice of a suitable concept of “neighborhood” in R_n corresponding to the interval $\langle x_0 - \eta, x_0 + \eta \rangle$ in Theorem 2.2 is fundamental.

From the point of view of elementary geometry the most natural definition would be: The set of all points of the space in question with the distance $\leq \eta$ from the “center” ξ_0 of the neighborhood. However, this definition, although satisfactory for most analytic discussions, is not very convenient from the point of view of the numerical analysis. From the numerical point of view it is, rather, preferable to use the generalized distance as defined in the preceding chapters and corresponding to the values $p = 1, \infty$. We will therefore define generally the neighborhood $U_\eta^{(p)}(\xi_0)$ of ξ_0 as the set of all points with

$$|\xi - \xi_0|_p \leq \eta \quad (p = 1, 2, \infty). \quad (24.1)$$

2. The set defined by (24.1) is a closed set, the “closed η - p neighborhood of ξ_0 .” It is convex; that is to say, if two points ξ_1 and $\xi_1 + \xi_2$ belong to $U_\eta^{(p)}(\xi_0)$, the same is true for all points of the “segment”

$$\xi_1 + h\xi_2 \quad (0 \leq h \leq 1)$$

connecting ξ_1 with $\xi_1 + \xi_2$. This follows immediately by the triangle inequality from

$$\begin{aligned} |\xi_1 + h\xi_2 - \xi_0|_p &= |h(\xi_1 + \xi_2 - \xi_0) + (1 - h)(\xi_1 - \xi_0)|_p \\ &\leq h|\xi_1 + \xi_2 - \xi_0|_p + (1 - h)|\xi_1 - \xi_0|_p \\ &\leq h\eta + (1 - h)\eta = \eta. \end{aligned}$$

3. In order to generalize Theorem 2.2 to the case of n equations with n unknowns we will replace the function $f(x)$ by n functions $f_\nu(x_1, \dots, x_n) \equiv f_\nu(\xi)$, where the n coordinates x_ν are components of the vector (or point) ξ . The approximate solution x_0 of $f(x) = 0$ is replaced by the vector $\xi_0 = (x_1^{(0)}, \dots, x_n^{(0)})$. The derivative $f'(x)$ is replaced by the Jacobian matrix $J = (\partial f_\mu / \partial x_\nu)$ and instead of $|f'(x)|$ we will consider an appropriate norm of the matrix J^{-1} . Our theorem is now

Theorem 24.1. *Let $f_\nu(\xi) = f_\nu(x_1, \dots, x_n)$ ($\nu = 1, \dots, n$) be n real functions of the real point ξ , defined and continuous with their first derivatives in the ρ - p neighborhood $U = U_\rho^{(p)}(\xi_0)$ 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_ν with respect to the x_μ and put*

$$f_\nu(\xi) = y_\nu, \quad \eta = (y_1, \dots, y_n)', \quad f_\nu(\xi_0) = y_\nu^{(0)}, \quad \eta_0 = (y_1^{(0)}, \dots, y_n^{(0)})' \quad (24.2)$$

If we have for one of the values $p = 1, 2, \infty$ and the corresponding $q = \infty, 2, 1$ either

$$|\eta_0|_q |\Delta_p(J^{-1})| \leq \rho, \quad p = 1, 2, \infty, \quad (24.3)$$

or

$$|\eta_0|_p |J^{-1}|_p \leq \rho, \quad p = 1, \infty, \quad (24.4)$$

throughout the whole neighborhood $U_\rho^{(p)}(\xi_0)$, then this neighborhood contains at least one point at which all $f_\nu(\xi)$ ($\nu = 1, \dots, n$) vanish.

PROOF OF THEOREM 24.1

4. Consider for $t \geq 0$ the system of n equations in n unknowns x_1, \dots, x_n

$$f_\mu(x_1, \dots, x_n) \equiv f_\mu(\xi) = y_\mu = (1 - t)y_\mu^{(0)} \quad (\mu = 1, \dots, n). \quad (24.5)$$

As by (24.3) or (24.4) the determinant $|J|$ does not vanish at ξ_0 , it follows by the classical existence theorem that for sufficiently small $\tau > 0$ there exist solutions $\xi(t)$ of (24.5) for $0 \leq t < \tau$, continuous and differentiable with respect to t , with $\xi(0) = \xi_0$ and forming a contin-

uous arc C_τ contained in $U = U_\rho^{(p)}(\xi_0)$. We have then along C_τ , using the notations (24.2),

$$\frac{d\eta}{dt} = -\eta_0, \quad \frac{d\eta}{dt} = J \frac{d\xi}{dt}, \quad \frac{d\xi}{dt} = -J^{-1}\eta_0.$$

It follows therefore from (23.19), (19.9), and (19.15) that

$$\left| \frac{d\xi}{dt} \right|_p \leq \Delta_p(J^{-1})|\eta_0|_q, \quad \left| \frac{d\xi}{dt} \right|_p \leq |J^{-1}|_p |\eta_0|_p,$$

and by virtue of (24.3) and (24.4) $|d\xi/dt|_p \leq \rho$. We have, therefore, integrating along C_τ ,

$$|\xi(t) - \xi_0|_p \leq \int_0^t \left| \frac{d\xi}{dt} \right|_p dt \leq t\rho \quad (0 \leq t < \tau). \quad (24.6)$$

5. Consider now the set M of all numbers $\tau > 0$ with the following property: There exists in U an arc C_τ ($\xi = \xi(t)$, $0 \leq t < \tau$) with the parameter t , such that we have along C_τ

$$f_\mu(\xi(t)) = (1-t)y_\mu^{(0)} \quad (\mu = 1, \dots, n), \quad \xi(0) = \xi_0 \quad (24.7)$$

and that $\xi'(t)$ exists and is continuous for $0 \leq t < \tau$. M is not empty.

Denote by τ^* the supremum of M and by t_v a sequence from the interval $(0, \tau^*)$ tending to τ^* . We can obviously assume $\tau^* \leq 1$, since otherwise the assertion of the theorem follows at once from (24.7) for $t = 1$. The solutions of (24.5) corresponding to the t_v may be denoted by ξ_v . These solutions lie all in U and have therefore at least one limiting point ξ^* which must also lie in the closed set U . We can then assume from the beginning that the ξ_v tend to ξ^* . Replacing in (24.5) t by t_v and ξ by ξ_v , we see for $v \rightarrow \infty$ that ξ^* satisfies (24.5) for $t = \tau^*$.

6. If τ^* is $= 1$ the assertion of the theorem follows from what has just been shown. We can hence assume from now on that $\tau^* < 1$.

Put, then, $\tau^*\rho = \rho^* < \rho$. Applying (24.6) to $t = t_v$, we see that $|\xi(t_v) - \xi_0|_p \leq t_v \rho$. For $v \rightarrow \infty$ it follows that $|\xi^* - \xi_0|_p \leq \rho^* < \rho$. This shows that ξ^* is an *inner point* of U . Applying the classical existence theorem to the point ξ^* , we see that there exists a neigh-

borhood $U_0 = U_{\rho_0}^{(p)}(\xi^*)$ of ξ^* contained in U and an arc C contained in U_0 , given by

$$\xi = \xi(t) \quad (\tau^* - d \leq t \leq \tau^* + d), \quad \xi(\tau^*) = \xi^*$$

with continuously differentiable $\xi(t)$ and a positive d , and such that (24.5) is satisfied along C . We can assume further by the classical existence theorem that for any t from $(\tau^* - d, \tau^* + d)$ there exists in U_0 only one solution of (24.5).

Denote by t' one of the t_ν lying in the interval $(\tau^* - \frac{1}{2}d, \tau^*)$ such that the corresponding $\xi' = \xi(t')$ lies in the interior of U_0 . The $t' = t_\nu$ belongs to a τ' from M such that $C' = C_{\tau'}$ satisfies the conditions of Section 5 and contains ξ' . Hence there is a t interval $(\tau' - \delta, \tau' + \delta)$ contained in $(\tau^* - \frac{1}{2}d, \tau^*)$ to which corresponds a partial arc C' of $C_{\tau'}$ containing ξ' and contained in U_0 . But then this arc must also belong to C . However, we can then begin along $C_{\tau'}$, proceed along C' , and finish along C , obtaining in this way an arc C^* satisfying the conditions of Section 5. Therefore, $\tau^* + d$ belongs also to M , which is impossible. We see that $\tau^* < 1$ is impossible and our theorem is proved.

A UNIQUENESS THEOREM

7. In practice it may be worth while to use simultaneously several criteria (24.3) and (24.4), each time choosing the corresponding p as small as possible—provided the first derivatives of the f_ν remain continuous in all these neighborhoods. 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 $f_\nu(\xi) = 0$. This can be secured in many cases using

Theorem 24.2. Consider a convex set of points S in the n -dimensional space R_n on which all functions

$$f_\nu(\xi) = f_\nu(x_1, \dots, x_n) \quad (\nu = 1, \dots, n)$$

are continuous with their first derivatives. Assume that the determinant

$$K(\xi_1, \dots, \xi_n) = \left| \frac{\partial f_\nu(\xi_\nu)}{\partial x_\mu} \right| = \begin{vmatrix} \frac{\partial f_1(\xi_1)}{\partial x_1} & \cdots & \frac{\partial f_1(\xi_1)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\xi_n)}{\partial x_1} & \cdots & \frac{\partial f_n(\xi_n)}{\partial x_n} \end{vmatrix} \quad (24.8)$$

remains $\neq 0$ if ξ_1, \dots, ξ_n run independently through S . Then if for two points of S , ξ', ξ'' , we have

$$f_\nu(\xi') - f_\nu(\xi'') = 0 \quad (\nu = 1, \dots, n), \quad (24.9)$$

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 (24.9)

$$\sum_{\mu=1}^n (x_\mu' - x_\mu'') \frac{\partial f_\nu}{\partial x_\mu}(\xi_\nu) = 0 \quad (\nu = 1, \dots, n),$$

where for each ν the point ξ_ν 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 24.2 is proved.

In the praxis, if the set S is sufficiently small, the values of the derivative $f'_\nu(\xi_\mu)$ 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 24.2 is immediate.

EXAMPLE

8. We will illustrate the use of our theorems in an example. Take

$$\begin{aligned} f_1 &= 4x_1 - x_2 - x_1 \sin x_2, & f_2 &= x_1 + x_2 - 4 \cot x_2, \\ \xi_0 &= (0.828, 3.849), & \eta_0 &= (0.021089, -0.021195). \end{aligned}$$

For $h = 10^{-3}$ we consider the neighborhood $U_h^{(\infty)}(\xi_0)$, the Jacobian matrix $J(\xi)$, the matrix $K(\xi_1, \xi_2)$, and $J^{-1}(\xi)$. We obtain throughout the whole neighborhood $U_h^{(\infty)}(\xi_0)$

$$\begin{aligned} f'_{1x_1} &= 4.650 (\pm 1), & f'_{1x_2} &= -0.370 (\pm 2), \\ f'_{2x_1} &= 1, & f'_{2x_2} &= 10.47 (\pm 2.5). \end{aligned}$$

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 $U_h^{(\infty)}(\xi_0)$.

J^{-1} is easily computed throughout the whole neighborhood $U_h^{(\infty)}(\xi_0)$ to be

$$J^{-1} = \frac{1}{100} \begin{pmatrix} 21.67 \pm 0.11 & -2.07 \pm 0.01 \\ -0.77 \pm 0.01 & 9.62 \pm 0.04 \end{pmatrix}.$$

9. We obtain now the following values of the norms of J^{-1} in the neighborhood $U_h^{(\infty)}(\xi_0)$:

p	$ J^{-1} _p$	$\Delta_p(J^{-1})$
1	$0.224 \pm 0.0_{22}$	$0.341 \pm 0.0_{22}$
2	—	$0.238 \pm 0.0_{22}$
∞	$0.237 \pm 0.0_{22}$	$0.217 \pm 0.0_{22}$

On the other hand, we have

$$|\eta_0|_1 = 0.0_{2228}, \quad |\eta_0|_\infty = 0.0_{2120}, \quad |\eta_0|_2 = 0.0_{2162}.$$

With these values we obtain the smallest values of ρ compatible with (24.3) and (24.4), putting $\delta = 10^{-4}$:

p	(24.3)	(24.4)
1	4.975δ	5.151δ
2	3.881δ	—
∞	4.1δ	2.852δ

We see that in our example the best value for $p = 1$ is obtained from (24.3) and for $p = \infty$ from (24.4). We have therefore only to consider the three neighborhoods corresponding to $\rho_1 = 4.975 \delta$, $\rho_2 = 3.881 \delta$, $\rho_\infty = 2.852 \delta$:

$$U_{\rho_1}^{(1)}, \quad U_{\rho_2}^{(2)}, \quad U_{\rho_\infty}^{(\infty)}. \quad (24.10)$$

These neighborhoods are contained in $U_h^{(\infty)}$. Therefore, our values of the norms hold in all neighborhoods (24.8) and Theorem 24.1 can be applied to any of these three neighborhoods. Since these neigh-

borhoods are also contained in $U_{h/2}^{(\infty)}(\xi_0)$, 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 (24.10). We represent the upper right quarters of the three neighborhoods (24.10) in Fig. 6. We see that the solution in question lies in the product $U_{\rho_1}^{(1)} U_{\rho_\infty}^{(\infty)}$. This is the pentagon OABCD together with the three pentagons obtained from it by symmetry with respect to OA and OD.

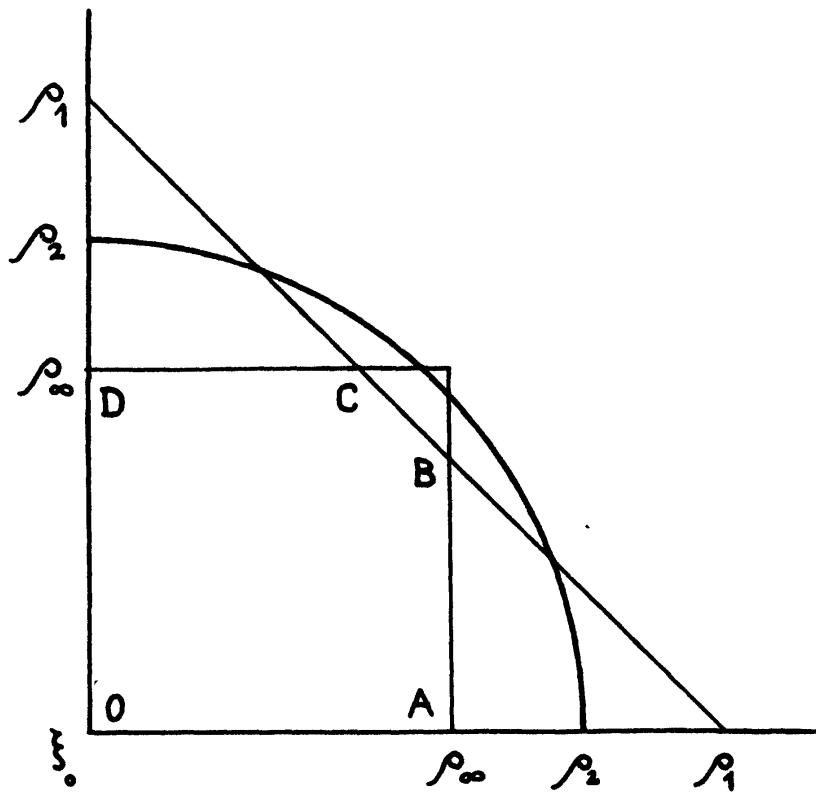


FIG. 6.

10. Another way to reduce the computation of J^{-1} , to its computation in one point only consists in the use of the formulas (23.22) and (23.23). In our example, if J^{-1} is computed in the point ξ_0 , we see easily that, if we denote the corresponding J by J_0 , for another point of $U_h^{(\infty)}(\xi_0)$ the difference $J - J_0$ is majorized by $h \begin{pmatrix} 1 & 1.4 \\ 0 & 4 \end{pmatrix}$

Therefore, we have

$$\Delta_\infty(J - J_0) \leq 4h, \quad \Delta_1(J - J_0) \leq 4h, \quad \Delta_2(J - J_0) \leq 4h,$$

$$|J - J_0|_\infty \leq 4h, \quad |J - J_0|_1 \leq 5.4h.$$

***n*-Dimensional Generalization of the Newton-Raphson Method. Statement of the Theorems**

1. We consider again, as in Chapter 22, *n* real functions f_i ,

$$f_i(\xi) = f_i(x_1, \dots, x_n) \quad (i = 1, 2, \dots, n),$$

of the *n*-dimensional real vector (or point) ξ with the components x_1, \dots, x_n , and assume that they are continuous with their first and second derivatives upon a certain closed *n*-dimensional set S . We denote now the Jacobian matrix of the f_i with respect to the x_j in a point ξ by

$$J(\xi) = \left(\frac{\partial f_i}{\partial x_j} (\xi) \right).$$

Consider a couple of numbers p, q satisfying the relations

$$\frac{1}{p} + \frac{1}{q} = 1, \quad p \geq 1, \quad q \geq 1,$$

where for $p = 1$ we put $q = \infty$, and for $q = 1$, $p = \infty$. Then put for an $n \times n$ matrix $A(a_{ij})$ generally

$$\Delta_q(A) = \left[\sum_{i,j} |a_{ij}|^q \right]^{1/q} \quad (q \geq 1) \quad (25.1)$$

and define $\Delta_q^*(\xi)$ by

$$\Delta_q^*(\xi) = \left[\sum_{\kappa, \mu, \nu} \left| \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_\nu} (\xi) \right|^q \right]^{1/q} \quad (p > 1, q > 1), \quad (25.2a)$$

$$\Delta_1^*(\xi) = \sum_{\kappa, \mu, \nu} \left| \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_\nu} (\xi) \right|, \quad (25.2b)$$

$$\Delta_\infty^*(\xi) = \max_{\kappa, \mu, \nu} \left| \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_\nu} (\xi) \right|. \quad (25.2c)$$

VARIATION OF THE JACOBIAN MATRIX

2. Lemma 1.1. *Assume that the conditions of Section 1 are satisfied along the segment S , $\xi + t\zeta$ ($0 \leq t \leq 1$), where ξ and ζ are two n -dimensional vectors. Then we have for $p \geq 1$*

$$\Delta_q(J(\xi + \zeta) - J(\xi)) \leq |\zeta|_p \Delta_q^*(\xi + \theta\zeta), \quad 0 \leq \theta \leq 1. \quad (25.3)$$

3. Proof. Assume first $p > 1$, $q > 1$. We write $\xi + t\zeta = \xi_t$, $\zeta = (z_1, \dots, z_n)$, and put

$$W(t) = \left(\sum_{\kappa, \mu} \left| \frac{d}{dt} \frac{\partial f_\kappa}{\partial x_\mu} (\xi_t) \right|^q \right)^{1/q} = \left(\sum_{\kappa, \mu} \left| \sum_v \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_v} (\xi_t) z_v \right|^q \right)^{1/q}, \quad (25.4)$$

$$D(t) = \Delta_q((J(\xi_t) - J(\xi)) = \sum_{\kappa, \mu} \left| \frac{\partial f_\kappa}{\partial x_\mu} (\xi_t) - \frac{\partial f_\kappa}{\partial x_\mu} (\xi) \right|^q. \quad (25.5)$$

In order to compute $D'(t)$ observe that we can write for any real function $g(t)$ differentiable in t , with $q > 1$:

$$\frac{d}{dt} |g(t)|^q = \frac{d}{dt} (g(t)^2)^{q/2} = \frac{q}{2} 2(g(t)^2)^{(q/2)-1} g(t) g'(t) = q |g(t)|^{q-2} g(t) g'(t),$$

as long as $g(t) \neq 0$. On the other hand, if $g(t) = 0$ and $g'(t)$ exists we have

$$|g(t')|^q = O(|t' - t|^q) \quad (t' \rightarrow t), \quad \frac{d|g(t)|^q}{dt} = 0.$$

We have therefore

$$\frac{1}{q} |D'(t)| = \sum_{\kappa, \mu} \left| \frac{\partial f_\kappa}{\partial x_\mu} (\xi_t) - \frac{\partial f_\kappa}{\partial x_\mu} (\xi) \right|^{q-2} \left(\frac{\partial f_\kappa}{\partial x_\mu} (\xi_t) - \frac{\partial f_\kappa}{\partial x_\mu} (\xi) \right) \left(\frac{d}{dt} \frac{\partial f_\kappa}{\partial x_\mu} (\xi_t) \right),$$

where, if a term of the sum (25.5) vanishes, the corresponding term in the above derivative has to be replaced by 0.

4. With this convention we can write

$$\frac{1}{q} |D'(t)| \leq \sum_{\kappa, \mu} \left| \frac{\partial f_\kappa}{\partial x_\mu} (\xi_t) - \frac{\partial f_\kappa}{\partial x_\mu} (\xi) \right|^{q-1} \left| \frac{d}{dt} \frac{\partial f_\kappa}{\partial x_\mu} (\xi_t) \right|.$$

To the right-hand sum in this inequality we apply the so-called Hölder's inequality

$$|\sum_\lambda a_\lambda b_\lambda| \leq \left(\sum_\lambda |a_\lambda|^p \right)^{1/p} \left(\sum_\lambda |b_\lambda|^q \right)^{1/q}. \quad (25.6)$$

Replacing here a_λ, b_λ by

$$\left| \frac{\partial f_\kappa}{\partial x_\mu}(\xi_t) - \frac{\partial f_\kappa}{\partial x_\mu}(\xi) \right|^{q-1}, \quad \left| \frac{d}{dt} \frac{\partial f_\kappa}{\partial x_\mu}(\xi_t) \right|,$$

since we have

$$p(q-1) = pq \left(1 - \frac{1}{q}\right) = q,$$

the first right-hand factor in (25.6) becomes $D(t)^{1/p}$ and the second $W(t)$.

We have therefore the relation

$$\frac{1}{q} |D'(t)| \leqq D(t)^{1/p} W(t),$$

and dividing on both sides by $D(t)^{1-(1/q)} = D(t)^{1/p}$ we get, if $D(t) \neq 0$,

$$\frac{1}{q} \left| \frac{D'(t)}{D(t)^{1-(1/q)}} \right| = |(D(t)^{1/q})'| \leqq W(t) \quad (D(t) \neq 0). \quad (25.7)$$

5. Assume now that $D(1) \neq 0$ and denote by t_0 the maximum of all t for which $D(t) = 0$ ($t_0 \geqq 0$ exists since $D(0) = 0$ and is continuous on $\langle 0, 1 \rangle$). Then we have for any t' with $t_0 < t' < 1$

$$D(1)^{1/q} - D(t')^{1/q} = \int_{t'}^1 (D(t)^{1/q})' dt,$$

and therefore, by (25.7), as $W(t) \geqq 0$,

$$|D(1)^{1/q} - D(t')^{1/q}| \leqq \int_{t'}^1 W(t) dt \leqq \int_0^1 W(t) dt.$$

For $t' \downarrow t_0$ we obtain

$$|D(1)^{1/q}| \leqq \int_0^1 W(t) dt = W(\theta), \quad 0 \leqq \theta \leqq 1,$$

or by (25.5)

$$\Delta_q(J(\xi + \zeta) - J(\xi)) \leqq W(\theta), \quad 0 \leqq \theta \leqq 1, \quad (25.7*)$$

and this is, of course, also true if $D(1) = 0$.

6. From (25.4) we have, applying to the inner sum over ν Hölder's inequality (25.6) and replacing there a_λ by z_ν and b_λ by $\partial^2 f_\kappa / \partial x_\mu \partial x_\nu$,

$$W(t)^q = \sum_{\kappa, \mu} \left| \sum_{\nu} \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_\nu} (\xi_t) z_\nu \right|^q \leq \sum_{\kappa, \mu} \left(\sum_{\nu} \left| \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_\nu} (\xi_t) \right|^q \right) |\zeta|_p^q,$$

and here the coefficient of $|\zeta|_p^q$ is, by the definition (25.2a), $A_q^*(\xi_t)^q$. We have therefore

$$W(t) \leq A_q^*(\xi_t) |\zeta|_p,$$

and putting this into (25.7*), finally the assertion (25.3) for $p > 1$, $q > 1$. But then the truth of (25.3) for $p = 1$ or $q = 1$ is obtained immediately for $p \downarrow 1$ or $q \downarrow 1$.

Lemma 1 is proved. From now on we consider only three couples of indices p, q , given by (23.18).

STATEMENT OF THE n -DIMENSIONAL ANALOG OF THE NEWTON-RAPHSON METHOD

7. In the Newton-Raphson method, applied to the system of equations

$$f_\kappa(x_1, \dots, x_n) \equiv f_\kappa(\xi) = 0 \quad (\kappa = 1, \dots, n), \quad (25.8)$$

we construct recurrently, starting from an initial point ξ_0 , a sequence of points $\xi_\lambda = (x_1^{(\lambda)}, \dots, x_n^{(\lambda)})$ in the following way:

Put

$$\begin{aligned} f_\kappa(\xi_\lambda) &= y_\kappa^{(\lambda)}, & \eta_\lambda &= (y_1^{(\lambda)}, \dots, y_n^{(\lambda)}), \\ \xi_{\lambda+1} - \xi_\lambda &= \zeta_\lambda = (z_1^{(\lambda)}, \dots, z_n^{(\lambda)}). \end{aligned}$$

Then ζ_λ is obtained from ξ_λ solving the linear system

$$\sum_{\nu=1}^n z_\nu^{(\lambda)} f'_{\kappa x_\nu}(\xi_\lambda) + y_\kappa^{(\lambda)} = 0 \quad (\kappa = 1, \dots, n). \quad (25.9)$$

8. In order to insure the possibility of this construction and the convergence of the ξ_λ to a solution of (25.8), we will assume that with the first step ζ_0 and ξ_1 have been computed from ξ_0 . We consider then the neighborhood U_0 of ξ_0 , given by

$$|\xi - \xi_0|_p \leq \sigma |\zeta_0|_p, \quad (25.10)$$

for a fixed value of $\sigma > 1$ and assume that the second derivatives of the f_κ are continuous in U_0 and that $\Delta_q^*(\xi)$ defined by (25.2) does not exceed for a general ξ in U_0 a certain positive constant M

$$\Delta_q^*(\xi) \leq M \quad (\xi \in U_0). \quad (25.11)$$

We denote again by $J(\xi)$ the Jacobian matrix of the f_κ in a point ξ , compute then for a convenient constant $\rho \geq 2$ the expression

$$\theta_0 = \rho M |\zeta_0|_\rho \Delta_p(J^{-1}(\xi_0)), \quad (25.12)$$

and will assume that we have $\theta_0 < 1$.

As to the constants σ and ρ , the simplest values of them are $\sigma = \rho = 2$. However, sometimes it may be more convenient to use different values of σ and ρ ; then they must satisfy the inequality

$$\sigma \geq \frac{2\rho - 2}{2\rho - 3} > 1, \quad \rho \geq 2, \quad (25.13)$$

or, if σ' is put equal to $\sigma - 1$,

$$\sigma' \geq \frac{1}{2\rho - 3}, \quad \rho \geq 2. \quad (25.13^\circ)$$

In what follows we denote the rectilinear segment connecting ξ_0 with ξ_1 , $\xi_0 + t\xi_0$ ($0 \leq t \leq 1$), by S_0 .

9. Our theorem will then be as follows:

Theorem 25.1. *Assume that the functions f_κ in (25.8) have continuous first derivatives and a nonvanishing Jacobian in a point ξ_0 . Form then, by (25.9) for $\lambda = 0$ the $z_\nu^{(0)}$, and then ζ_0 and ξ_1 . Take two positive constants σ, ρ satisfying (25.13) and assume that in the neighborhood U_0 (25.10) the functions f_κ have continuous second derivatives and that we have (25.11) for a certain M , and further, that the expression for θ_0 given by (25.12) is < 1 . Then the sequence ξ_λ can be formed as described in Section 7, remains in U_0 and converges to a solution ξ^* of the equations (25.8). Moreover, we have*

$$|\xi_{\lambda+1} - \xi^*|_\rho = O(|\xi_\lambda - \xi^*|_\rho^2) \quad (\lambda \rightarrow \infty). \quad (25.14)$$

If in the above theorem we have $\sigma \geq 2$, the conditions of the theorem can be somewhat weakened and the assertion improved.

We have

Theorem 25.2. *If we have in (25.13) $\sigma \geq 2$, the neighborhood U_0 in the assumptions in Theorem 25.1 can be replaced by the neighborhood V_0 :*

$$|\xi - \xi_1|_p \leq \sigma' |\zeta_0|_p, \quad \sigma' = \sigma - 1 \geq 1, \quad (25.15)$$

which is a subset of U_0 , and, beyond the assertions of Theorem 25.1, the sequence ξ_λ lies in V_0 .

1. Put, as long as ξ_λ lies in U_0 ,

$$\Delta_p(J^{-1}(\xi_\lambda)) = \Delta^{(\lambda)}, \quad (26.1)$$

$$\theta_\lambda = \rho M |\zeta_\lambda|_p \Delta^{(\lambda)}. \quad (26.2)$$

Developing $y_\kappa^{(1)} = f_\kappa(\xi_0 + \zeta_0)$ in a Taylor series with the remainder of the second order, we obtain, by virtue of (25.9) for $\lambda = 0$, denoting by α a convenient point of S_0 ,

$$y_\kappa^{(1)} = \frac{1}{2} \sum_{\mu, \nu} \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_\nu} (\alpha) z_\mu^{(0)} z_\nu^{(0)},$$

and therefore, by Hölder's inequality (25.6), replacing there a_λ by $z_\mu^{(0)} z_\nu^{(0)}$,

$$|y_\kappa^{(1)}| \leq \frac{1}{2} \left[\sum_{\mu, \nu} \left| \frac{\partial^2 f_\kappa}{\partial x_\mu \partial x_\nu} (\alpha) \right|^q \right]^{1/q} |\zeta_0|_p^2.$$

Raising this into the q th power and summing over κ , we have now, by (25.11), since S_0 lies both in U_0 and (for $\sigma \geqq 2$) in V_0 ,

$$|\eta_1|_q \leqq \frac{1}{2} M |\zeta_0|_p^2. \quad (26.3)$$

Further, applying (25.3) to ξ_0 and ζ_0 and using (25.11), we have

$$\Delta_q(J(\xi_1) - J(\xi_0)) \leqq M |\zeta_0|_p.$$

If in Theorem 23.1 we put $A = J(\xi_0)$, $B = J(\xi_1)$, the condition (23.28) of this theorem is satisfied since we have by (25.12) for $\rho \geqq 2$

$$\Delta_p(A^{-1}) \Delta_q(B - A) \leqq M |\zeta_0|_p \Delta_p(J^{-1}(\xi_0)) = \frac{\theta_0}{\rho} < \frac{1}{\rho} \leqq \frac{1}{2},$$

and it follows from (23.26) that

$$\Delta^{(1)} = \Delta_p(J^{-1}(\xi_1)) \leq \frac{1}{1 - \theta_0/\rho} \Delta_p(J^{-1}(\xi_0)) = \frac{\rho}{\rho - \theta_0} \Delta^{(0)}. \quad (26.4)$$

In particular, as $\rho \geq 2$, we have $\Delta^{(1)} < 2\Delta^{(0)}$.

2. From (25.9) for $\lambda = 1$, we have $\zeta_1 = -J^{-1}(\xi_1)\eta_1$, and therefore, by (23.19) and (26.1),

$$|\zeta_1|_p \leq \Delta^{(1)} |\eta_1|_q. \quad (26.5)$$

Introducing now (26.3) and (26.4) in (26.5), we have, as $\theta_0 < 1$, by (26.2),

$$|\zeta_1|_p \leq \frac{1}{2} \frac{\rho \Delta^{(0)} M |\zeta_0|_p^2}{\rho - \theta_0} = \frac{1}{2} \frac{\theta_0 |\zeta_0|_p}{\rho - \theta_0} < \frac{1}{2} \frac{|\zeta_0|_p}{\rho - 1} \leq \frac{1}{2} |\zeta_0|_p. \quad (26.6)$$

3. We consider now the neighborhoods

$$U_1: |\xi - \xi_1|_p \leq \sigma |\zeta_1|_p; \quad V_1: |\xi - \xi_2|_p \leq \sigma' |\zeta_1|_p$$

and are going to prove that U_1 is contained in U_0 , and, for σ' satisfying (25.13°), $V_1 \prec V_0$.

Indeed, for a point ξ in U_1 we have by (26.6)

$$|\xi - \xi_0|_p \leq |\xi - \xi_1|_p + |\zeta_0|_p \leq \sigma |\zeta_1|_p + |\zeta_0|_p \leq \left(\frac{\sigma}{2} \frac{1}{\rho - 1} + 1 \right) |\zeta_0|_p.$$

But here the coefficient of $|\zeta_0|_p$ is by (25.13) $\leq \sigma$. On the other hand, for a point ξ from V_1 we have, again by (26.6),

$$|\xi - \xi_1|_p \leq |\xi - \xi_2|_p + |\zeta_1|_p \leq (\sigma' + 1) |\zeta_1|_p \leq \frac{\sigma' + 1}{2(\rho - 1)} |\zeta_0|_p,$$

and here the coefficient of $|\zeta_0|_p$ is, by (25.13°), $\leq \sigma'$.

We verify finally that V_0 is a subset of U_0 . Indeed, if (25.15) is satisfied for a ξ , we have

$$|\xi - \xi_0|_p \leq |\xi - \xi_1|_p + |\zeta_0|_p \leq \sigma' |\zeta_0|_p + |\zeta_0|_p = \sigma |\zeta_0|_p.$$

4. Multiplying (26.6) by $\rho M \Delta^{(1)}$ and using (26.4) we have

$$\begin{aligned} \theta_1 &= \rho M |\zeta_1|_p \Delta^{(1)} \leq \frac{1}{2} \rho M \frac{\theta_0 |\zeta_0|_p}{\rho - \theta_0} \cdot \frac{\rho \Delta^{(0)}}{\rho - \theta_0} = \frac{\rho}{2} \frac{\theta_0^2}{(\rho - \theta_0)^2}, \\ \theta_1 &\leq \frac{\rho}{2} \frac{\theta_0^2}{(\rho - \theta_0)^2} < \frac{\rho}{2(\rho - 1)^2} \theta_0^2. \end{aligned} \quad (26.7)$$

As $\rho \geq 2$, the coefficient of θ_0^2 is here ≤ 1 and we have $\theta_1 < \theta_0^2 < 1$.

5. We see now that the conditions of Theorems 25.1 and 25.2 remain valid if we replace in them the index 0 by 1 (and 1 by 2). Therefore, our iteration is unlimited and we have generally, by (26.6),

$$|\zeta_{\lambda+1}|_p \leq \frac{1}{2} |\zeta_\lambda|_p \leq 2^{-\lambda-1} |\zeta_0|_p, \quad \theta_{\lambda+1} < \theta_\lambda^2.$$

It follows then for the components of ξ_λ that

$$|x_\kappa^{(\lambda+1)} - x_\kappa^{(\lambda)}| \leq 2^{-\lambda} |\zeta_0|_p \quad (\lambda = 1, 2, \dots; \kappa = 1, \dots, n),$$

so that each series

$$\sum_{\lambda=0}^{\infty} (x_\kappa^{(\lambda+1)} - x_\kappa^{(\lambda)})$$

is convergent; we have for any κ , $x_\kappa^{(\lambda)} \rightarrow x_\kappa^*$ ($\lambda \rightarrow \infty$), $\xi_\lambda \rightarrow \xi^* = (x_1^*, \dots, x_n^*)$, and this ξ^* lies also in U_0 and (for $\sigma \geq 2$) in V_0 . Further, it follows that

$$|x_\kappa^{(\lambda)} - x_\kappa^*| \leq \sum_{\nu=\lambda}^{\infty} |x_\kappa^{(\nu+1)} - x_\kappa^{(\nu)}| \leq |\zeta_\lambda|_p \sum_{\nu=0}^{\infty} 2^{-\nu} = 2 |\zeta_\lambda|_p$$

and hence

$$|\xi_\lambda - \xi^*|_p \leq 2n |\zeta_\lambda - \zeta_{\lambda+1}|_p = 2n |\zeta_\lambda|_p. \quad (26.7*)$$

We have from (26.6), by (26.2), putting

$$\delta = \frac{1}{2(\rho-1)},$$

$$|\zeta_{\lambda+1}|_p \leq \delta \theta_\lambda |\zeta_\lambda|_p = \frac{\rho M}{2(\rho-1)} \Delta^{(\lambda)} |\zeta_\lambda|_p^2 \quad (26.8)$$

and therefore

$$|\zeta_\lambda|_p \leq \delta \theta_{\lambda-1} |\zeta_{\lambda-1}|_p \leq \delta^2 \theta_{\lambda-1} \theta_{\lambda-2} |\zeta_{\lambda-2}|_p \leq \dots \leq \delta^\lambda \theta_{\lambda-1} \dots \theta_0 |\zeta_0|_p.$$

If we multiply this inequality on both sides by $\rho M \Delta^{(0)}$ and use (25.12), we have

$$\rho M \Delta^{(0)} |\zeta_\lambda|_p \leq \delta^\lambda \theta_{\lambda-1} \theta_{\lambda-2} \dots \theta_1 \theta_0^2.$$

On the other hand, for each $\kappa > 1$,

$$\theta_\kappa < \theta_{\kappa-1}^2 < \dots < \theta_0^{2^\kappa},$$

and therefore

$$\begin{aligned} \rho M A^{(0)} |\zeta_\lambda|_p &\leq \delta^\lambda \theta_0^{2^\lambda - 1 + 2^\lambda - 2 + \dots + 2 + 2}, \\ |\zeta_\lambda|_p &\leq \frac{1}{\rho M A^{(0)}} \left(\frac{1}{2(\rho - 1)} \right)^\lambda \theta_0^{2^\lambda} \end{aligned} \quad (26.9)$$

6. Applying (26.8) repeatedly we obtain for $\mu > \lambda > 0$

$$\begin{aligned} \frac{|\zeta_{\lambda+1}|_p}{|\zeta_\lambda|_p} &\leq \delta \theta_\lambda, \quad \frac{|\zeta_{\lambda+2}|_p}{|\zeta_\lambda|_p} \leq \delta^2 \theta_\lambda \theta_{\lambda+1} \leq \delta^2 \theta_\lambda^3, \dots, \\ \frac{|\zeta_\mu|_p}{|\zeta_\lambda|_p} &\leq \delta^{\mu-\lambda} \theta_\lambda \dots \theta_{\mu-1} \leq \delta^{\mu-\lambda} \theta_\lambda^{2^\mu - \lambda - 1}, \end{aligned}$$

and therefore

$$\frac{\sum_{\kappa=\lambda+1}^{\mu} |\zeta_\kappa|_p}{|\zeta_\lambda|_p} \leq \frac{1}{\theta_\lambda} \sum_{v=1}^{\infty} \delta^v \theta_\lambda^{2^v} \equiv \tau.$$

Observe that $\tau < 1$ if either $\delta < \frac{1}{2}$, $\rho > 2$, $\theta_\lambda \leq 1$, or $\delta = \frac{1}{2}$, $\rho = 2$, $\theta_\lambda < 1$.

It follows now that

$$\begin{aligned} \frac{1}{|\zeta_\lambda|_p} |\xi_\lambda - \xi_{\mu+1}|_p &= \frac{1}{|\zeta_\lambda|_p} |\zeta_\lambda + \dots + \zeta_\mu|_p \\ &\geq 1 - \frac{|\zeta_{\lambda+1}|_p + \dots + |\zeta_\mu|_p}{|\zeta_\lambda|_p} \geq 1 - \tau, \end{aligned}$$

and therefore, for $\mu \rightarrow \infty$,

$$\frac{|\xi_\lambda - \xi^*|_p}{|\zeta_\lambda|_p} \geq 1 - \tau.$$

We get now, if $\tau < 1$, analogously to (26.7*),

$$|\zeta_\lambda|_p = O(|\xi_\lambda - \xi^*|_p). \quad (26.10)$$

7. It is easy to show that the $\Delta^{(\lambda)}$ are bounded. Indeed, applying for $\lambda > \kappa \geq 0$ the formula (23.26), we have

$$\Delta^{(\lambda)} \leq \frac{\Delta^{(\kappa)}}{1 - \Delta^{(\kappa)} \Delta_q(J^{-1}(\xi_\lambda) - J^{-1}(\xi_\kappa))}$$

and, by repeated application of the formulas (25.3) and (25.11),

$$\Delta_q(J^{-1}(\xi_\lambda) - J^{-1}(\xi_\kappa)) \leq M \sum_{\nu=\kappa}^{\infty} |\zeta_\nu|_p.$$

Further, applying repeatedly (26.6) to two consecutive $\zeta_\mu, \zeta_{\mu+1}$, since $\theta_{\mu+1} \leq \theta_\mu^2 < \theta_\mu$, we have

$$\sum_{\nu=\kappa+1}^{\infty} \frac{|\zeta_\nu|_p}{|\zeta_\kappa|_p} \leq \sum_{\nu=1}^{\infty} \left(\frac{\theta_\kappa}{2\rho - 2\theta_\kappa} \right)^\nu = \frac{\theta_\kappa}{2\rho - 3\theta_\kappa};$$

but this is < 2 as soon as $\theta_\kappa < (4/7)\rho$, which is certainly the case, since $\theta_\kappa < 1$ and $\rho \geq 2$.

We have, therefore,

$$\Delta_q(J^{-1}(\xi_\lambda) - J^{-1}(\xi_\kappa)) \leq 2M |\zeta_\kappa|_p,$$

and hence, since $\rho \geq 2$ and $2M\Delta^{(\kappa)}|\zeta_\kappa|_p \leq \rho M\Delta^{(\kappa)}|\zeta_\kappa|_p = \theta_\kappa$,

$$\Delta^{(\lambda)} \leq \frac{\Delta^{(\kappa)}}{1 - 2M\Delta^{(\kappa)}|\zeta_\kappa|_p} = \frac{\Delta^{(\kappa)}}{1 - \theta_\kappa} \leq \frac{\Delta^{(1)}}{1 - \theta_1} \quad (\lambda > \kappa). \quad (26.11)$$

It follows now from (26.8) that

$$|\zeta_{\lambda+1}|_p = O(|\zeta_\lambda|_p^2), \quad (26.12)$$

and by (26.10)

$$|\zeta_{\lambda+1}|_p = O(|\xi_\lambda - \xi^*|_p^2). \quad (26.13)$$

But then, as $|\xi_{\lambda+1} - \xi^*|_p \leq 2n|\zeta_{\lambda+1}|_p$, (25.14) follows and our Theorems 25.1 and 25.2 are proved.

8. If instead of assuming $\theta_0 < 1$ we only assume $\theta_0 \leq 1$, the inequalities derived in the above proof remain correct, if we replace $<$ signs by \leq signs in all of them save (26.11)–(26.13). We have therefore in particular

$$\theta_{\mu+1} \leq \theta_\mu^2.$$

But then it is sufficient to assume that one of the θ_κ becomes < 1 , since we can then start in our proof with any ξ_κ . In particular, if $\rho > 2$, it follows from (26.7) that already $\theta_1 < 1$. On the other hand, in the case $\theta_1 = 1$, $\rho = 2$, we also still have convergence by virtue of (26.6); however, then it need not be *quadratic* as indicated by (25.14), but could become only *linear*. In any case, for $n = 1$ and the quadratic polynomial $f(x)$ this can happen indeed.

9. The special assumption of Theorem 25.2 that σ' is ≥ 1 , was only necessary in order to insure that S_0 was contained in V_0 , since the estimate (25.11) was only used along S_0 . In all other arguments in which σ' entered, only the inequality (25.13°) was used and not $\sigma' \geq 1$. It is therefore easy to show *the assertions of Theorem 25.2 remain valid if σ' only satisfies the conditions (25.13°)*, provided that the second derivatives of the f_κ exist in a neighborhood of S_0 , are continuous on S_0 , and satisfy on S_0 the inequality (25.11).

To prove this, it is sufficient to show that in any case ξ_1 and ξ_2 lie in V_0 , since then also the inequality (25.11) is satisfied along the segment connecting ξ_1 and ξ_2 . But ξ_1 is the center of V_1 and we have only to show that ξ_2 lies in V_0 , that is, $|\xi_1|_\rho \leq \sigma' |\xi_0|_\rho$, and this follows from (26.6) and (25.13°) as we have

$$|\xi_1|_\rho \leq \frac{1}{2(\rho - 1)} |\xi_0|_\rho < \frac{1}{2\rho - 3} |\xi_0|_\rho.$$

1. Both methods of solution of systems of equations discussed in Chapters 22 and 25–26 are characteristically *local* methods 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 Jordan curve S .

Consider a function $f(\xi)$ defined and continuous on $\Omega + S$ which has on S a fixed value C and is in the interior of Ω everywhere $< C$. Assume that $f(\xi)$ has continuous second derivatives on $\Omega + S$.

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 $C \equiv C_0$ to $C_1 < C_0$. Through the point ξ_1 goes a level surface S_1 of $f(\xi)$ on which the value of f is everywhere equal to C_1 . Starting from ξ_1 , go along the normal to S_1 a segment of the length r_1 with the end point ξ_2 , so that $f(\xi_2) = C_2 < C_1$. 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

$$\frac{\partial f}{\partial x_\nu}(\xi^*) = 0 \quad (\nu = 1, \dots, n). \quad (27.1)$$

4. It appears first that the aim of the method described is the solution of a rather special system of equations (27.1).

However, if we have the general system of equations which can be written in the form

$$f_\nu(\xi) = 0 \quad (\nu = 1, \dots, n), \quad (27.2)$$

we can form $f(\xi)$ as

$$f(\xi) = f_1(\xi)^2 + \dots + f_n(\xi)^2. \quad (27.3)$$

This $f(\xi)$ has in the solution point of (27.2) the absolute minimum 0 and we can hope to obtain this point using Cauchy's method.

5. Analytically the normal in ξ_μ to the surface S_μ has the same direction as the vector $\text{grad } f(\xi_\mu)$ with 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 vanish, we can write, denoting generally by $|\xi|$ the Euclidean length of ξ , i.e., $|\xi|_2$,

$$|f'(\xi_\mu)| = \kappa_\mu, \quad f'(\xi_\mu) = \kappa_\mu \Phi_\mu. \quad (27.4)$$

Then Cauchy's rule can be represented by the formula

$$\xi_{\mu+1} = \xi_\mu - r_\mu \Phi_\mu.$$

6. As to the choice of r_μ , Cauchy's recommendation was to choose r_μ in such a way that the function $f(\xi_\mu - t\Phi_\mu)$ 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 one 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, already 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

$$|\varepsilon_\mu| \leq \varepsilon, \quad \varepsilon < 1 \quad (\mu = 0, 1, \dots) \quad (27.5)$$

and put

$$r_\mu = (1 + \varepsilon_\mu)c\kappa_\mu. \quad (27.6)$$

7. On the other hand, it is also not necessary to keep the direction rigidly fixed as that of ϕ_μ , that is, of the normal to S_μ . We can take instead any direction whose angle with the direction ϕ_μ remains essentially under $\pi/2$. We introduce therefore 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) ,

$$\delta_\mu \equiv (\phi_\mu, \psi_\mu) \geqq \delta, \quad 0 < \delta < 1, \quad (27.7)$$

with a fixed δ .

8. The system of the $\binom{n}{2}$ second derivatives of f can be ordered as an $(n \times n)$ symmetric matrix which we will denote by

$$f''(\xi) = \left(\frac{\partial^2 f(\xi)}{\partial x_i \partial x_j} \right). \quad (27.8)$$

Together with (27.8), we will also have to consider the more general matrix, that is obtained from $f''(\xi)$, if we take in each line of this matrix an argument depending on this line, i.e.,

$$f''(\xi_1, \dots, \xi_n) = \left(\frac{\partial^2 f(\xi_i)}{\partial x_i \partial x_k} \right) \quad (i, k = 1, \dots, n). \quad (27.9)$$

Now we make for our discussion the fundamental assumption that there exists a fixed *positive* Λ^* such that (cf. (23.10))

$$\Lambda(f''(\xi)) \leqq \Lambda^* \quad (\xi \prec \Omega). \quad (27.10)$$

9. In terms of δ and Λ^* the convenient value of c in (27.6) can be now taken as δ/Λ^* so that we have

$$r_\mu = (1 + \varepsilon_\mu) \frac{\delta}{\Lambda^*} \kappa_\mu, \quad (27.11)$$

$$\xi_{\mu+1} = \xi_\mu - r_\mu \psi_\mu. \quad (27.12)$$

10. Denote the set of all points ξ from Ω in which $f'(\xi) = 0$ by Ω^* :

$$\frac{\partial f}{\partial x_i}(\xi) = 0 \quad (i = 1, \dots, n). \quad (27.13)$$

We denote by the symbol $|\xi, \Omega^*|$ the distance of ξ from Ω^* , i.e.,

$$|\xi, \Omega^*| = \inf_{\xi' \prec \Omega^*} |\xi - \xi'|, \quad (27.14)$$

and we say that *a sequence of points ξ_μ tends to Ω^* , $\xi_\mu \rightarrow \Omega^*$, if*

$$|\xi_\mu, \Omega^*| \rightarrow 0 \quad (\mu \rightarrow \infty). \quad (27.15)$$

We will later discuss in what cases it follows from (27.15) 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 the Section 2.

Theorem 27.1. *Take a bounded, open, n -dimensional set Ω with the boundary S and assume that a function $f(\xi)$ is continuous on $\Omega + S$, has a constant 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 $\Omega + \xi_0$, that (27.10) is satisfied and that $f'(\xi_0)$ is $\neq 0$. Then we can, starting from ξ_0 , use the rule (27.12) indefinitely for $\mu = 0, 1, \dots$; all ξ_μ lie in Ω , tend to Ω^* in the sense of (27.15), and we have*

$$\xi_{\mu+1} - \xi_\mu \rightarrow 0 \quad (\mu \rightarrow \infty). \quad (27.16)$$

12. Proof. Putting

$$\xi^{(t)} = \xi_0 - t r_0 \psi_0, \quad 0 \leqq t \leqq 1,$$

develop $f(\xi^{(t)})$ in powers of t with the remainder term of second order. We obtain, using (27.4), for a θ with $0 \leqq \theta \leqq 1$,

$$f(\xi^{(t)}) - f(\xi_0) = -t r_0 \kappa_0(\psi_0, \Phi_0) + \frac{t^2}{2} r_0^2 (\psi_0', f''(\xi_0 - \theta t r_0 \psi_0) \psi_0). \quad (27.17)$$

Since the quotient of the right-side expression by t has a negative limit as $t \downarrow 0$, we see that the right-side 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 (27.7), (27.11), (27.10), and (23.13),

$$\begin{aligned} f(\xi^{(t)}) - f(\xi_0) &\leq -(1 + \varepsilon_0)t \frac{\delta}{A^*} \kappa_0^2 \delta + \frac{t^2}{2} (1 + \varepsilon_0)^2 \frac{\delta^2}{A^{*2}} \kappa_0^2 A^* \\ &= -(1 + \varepsilon_0) \frac{\delta^2}{A^*} \kappa_0^2 \frac{t}{2} (2 - t(1 + \varepsilon_0)) \\ &\leq -\frac{1 - \varepsilon_0^2}{2A^*} \delta^2 \kappa_0^2 t. \end{aligned}$$

Since $\xi^{(t)}$ and the whole segment from ξ_0 to $\xi^{(t)}$, save ξ_0 , lies in Ω , we can go on and up to $t = 1$ and remain always in Ω since $f(\xi^{(t)})$ stays $< C$ and we cannot therefore meet a point of S . We see that ξ_1 lies in Ω and that we have

$$\Delta_0 \equiv f(\xi_0) - f(\xi_1) \geq \frac{1 - \varepsilon_0^2}{2A^*} \delta^2 \kappa_0^2 \geq \frac{1 - \varepsilon^2}{2A^*} \delta^2 \kappa_0^2.$$

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

$$\Delta_\mu \equiv f(\xi_\mu) - f(\xi_{\mu+1}) \geq \frac{1 - \varepsilon^2}{2A^*} \delta^2 \kappa_\mu^2 \quad (\mu = 0, 1, \dots). \quad (27.18)$$

The sequence $f(\xi_\mu)$ is therefore monotonically decreasing and convergent, and we have $\Delta_\mu \rightarrow 0$ and from (27.18)

$$\kappa_\mu \rightarrow 0, \quad f'(\xi_\mu) \rightarrow 0, \quad r_\mu \rightarrow 0, \quad \xi_{\mu+1} - \xi_\mu \rightarrow 0.$$

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 $z = x + iy$:

$$f(z) = u(x, y) + iv(x, y), \quad F(z) = |f(z)|^2 = u^2(x, y) + v^2(x, y), \quad (27.19)$$

where by the Cauchy-Riemann equations

$$u_x' = v_y', \quad u_y' = -v_x'. \quad (27.20)$$

Denoting by $f' = f'(z)$ the derivative with respect to z , we have

$$\frac{\partial f(z)}{\partial x} = f'(z), \quad \frac{\partial f(z)}{\partial y} = if'(z), \quad (27.21)$$

$$\overline{\frac{\partial f(z)}{\partial x}} = u_x' - iv_x' = \overline{f'(z)}, \quad \overline{\frac{\partial f(z)}{\partial y}} = u_y' - iv_y' = -v_x' - iu_x' = -i\overline{f'(z)}.$$

It follows now that

$$\begin{aligned} F_x' &= (\bar{f}\bar{f})_x' = f_x'\bar{f} + \bar{f}f_x' = f'\bar{f} + \bar{f}f' = 2R(\bar{f}\bar{f}'), \\ F_y' &= f_y'\bar{f} + \bar{f}f_y' = if'\bar{f} - if\bar{f}' = \frac{1}{i}(f\bar{f}' - f'\bar{f}) = 2I(\bar{f}\bar{f}'), \\ Z(z) &= Z(x, y) = F_x' + iF_y' = 2f(z)\bar{f}'(z). \end{aligned} \quad (27.22)$$

15. From (27.20) we have further

$$\frac{\partial(u, v)}{\partial(x, y)} = u_x'^2 + v_x'^2 = |f'(z)|^2. \quad (27.23)$$

On the other hand, from (27.19) it follows, using $u_{xx}'' + u_{yy}'' = v_{xx}'' + v_{yy}'' = 0$ and (27.20), that

$$\frac{1}{2}F_{xx}'' = u_x'^2 + v_x'^2 + uu_{xx}'' + vv_{xx}'' = |f'(z)|^2 + R(\bar{f}\bar{f}''),$$

$$\begin{aligned} \frac{1}{2}F_{xy}'' &= u_y'u_x' + v_y'v_x' + uu_{xy}'' + vv_{xy}'' \\ &= (-v_x')u_x' + u_x'v_x' + u(-v_x')_x + v(u_x')_x = I(\bar{f}\bar{f}''), \end{aligned}$$

$$\begin{aligned} \frac{1}{2}F_{yy}'' &= (-uv_x' + vu_x')_y' = -u_y'v_x' + v_y'u_x' - uv_{yx}'' + vu_{yx}'' \\ &= v_x'^2 + u_x'^2 - (uu_{xx}'' + vv_{xx}'') = |f'(z)|^2 - R(\bar{f}\bar{f}''); \end{aligned}$$

collecting these results we have

$$\begin{aligned} F''_{xx} &= 2|f'|^2 + 2R(f\bar{f}''), \\ F''_{xy} &= 2I(f\bar{f}''), \\ F''_{yy} &= 2|f'|^2 - 2R(f\bar{f}''). \end{aligned} \quad (27.24)$$

16. The Hessian matrix of F ,

$$H = \begin{pmatrix} F''_{xx} & F''_{xy} \\ F''_{xy} & F''_{yy} \end{pmatrix},$$

has therefore as its determinant

$$\begin{vmatrix} F''_{xx} & F''_{xy} \\ F''_{xy} & F''_{yy} \end{vmatrix} = 4|f'|^4 - 4((R(f\bar{f}''))^2 + (I(f\bar{f}''))^2) = 4|f'|^4 - 4|f|^2|f''|^2. \quad (27.25)$$

As the trace of this matrix is $F''_{xx} + F''_{yy} = 4|f'|^2$, we obtain as the fundamental equation of H

$$\lambda^2 - 4|f'|^2\lambda + 4|f'|^4 - 4|f|^2|f''|^2 = 0, \quad (27.26)$$

and since the roots of this equation are $2|f'|^2 \pm 2|ff''|$, we have

$$\lambda(H) = 2|f'|^2 - 2|ff''|, \quad \Lambda(H) = 2|f'|^2 + 2|ff''|. \quad (27.27)$$

17. In order to apply the above theory to $F(z)$ given by (27.19), 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 A^* , it is given by the formula

$$A^* = 2 \max_{z \in \Omega + S} (|f'(z)|^2 + |f(z)f''(z)|). * \quad (27.28)$$

In this particular case it is simpler to use $\psi = \phi$ so that the iteration formula is of the type

$$z_1 = z_0 - 2tf(z_0)\bar{f}'(z_0). \quad (27.29)$$

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 (27.29).

* As a matter of fact we have even

$$A^* = 2 \max_{z \in S} (|f'(z)|^2 + \gamma |f''(z)|). \quad (27.28')$$

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 .

THE DERIVED SET AT THE ξ_μ

1. If the set Ω^* in Theorem 27.1 consists of only one point, ζ , then from Theorem 27.1 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 28.1. *Assume a bounded sequence S of points ξ_ν in R_n for which $\xi_{\nu+1} - \xi_\nu \rightarrow 0$. Then the derived set S' of S is a continuum, if ξ_ν 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 closed sets of points without common points.

3. **Proof of Theorem 28.1.** S' is obviously closed. Suppose we have $S' = C_1 + C_2$ where C_1 and C_2 are both closed and have no points in common. Then there exists a positive p such that the distance of any point of C_1 from every point of C_2 does not exceed p . By hypothesis we have for a certain n_0

$$|\xi_{\nu+1} - \xi_\nu| \leq p/3 \quad (\nu \geq n_0). \quad (28.1)$$

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_ν with $\nu > m$ 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 (28.1) $|\xi_k, C_2| > p/3$. We see that

$$p/3 < |\xi_k, C_2| \leq 2p/3, \quad k > m. \quad (28.2)$$

4. There exists therefore an infinite sequence of indices k_1, k_2, \dots for which (28.2) holds. A cluster point ζ of this sequence belongs to S and satisfies the relation

$$p/3 \leq |\zeta, C_2| \leq 2p/3. \quad (28.3)$$

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 27.1 we have

$$\xi_\mu \rightarrow \zeta. \quad (28.4)$$

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

$$\limsup_{\mu \rightarrow \infty} \frac{|\xi_{\mu+N} - \zeta|}{|\xi_\mu - \zeta|} < 1. * \quad (28.5)$$

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 $f(\xi)$ has 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 28.2. *Assume under the conditions of Theorem 27.1 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

* 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.

$$\lambda = \frac{1}{2\sqrt{n}} \lambda(f_0''), \quad \Lambda = 2\Lambda(f_0''). \quad (28.6)$$

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

$$\lambda(f''(\xi)) > \lambda, \quad \Lambda(f''(\xi)) < \Lambda \quad (\xi \prec U). \quad (28.7)$$

We can therefore also assume that Λ^* introduced in (27.10) and used in (27.11) is equal to Λ .

9. We have, by what has been said at the end of Chapter 19,

$$\Lambda(f_0''^{-1}) = \frac{1}{\lambda(f_0'')}. \quad (28.8)$$

further, obviously,

$$\Lambda(f''(\eta_1, \dots, \eta_n)^{-1}) \rightarrow \Lambda(f_0''^{-1}) = \frac{1}{\lambda(f_0'')}, \quad (28.9)$$

if η_1, \dots, η_n tend independently to the origin. Taking the neighborhood U of the origin sufficiently small we have therefore also

$$\Lambda(f''(\eta_1, \dots, \eta_n)^{-1}) < \frac{2}{\lambda(f_0'')} = \frac{1}{\lambda\sqrt{n}} \quad (\eta_1, \dots, \eta_n \prec U). \quad (28.10)$$

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 (28.7),

$$\lambda|\xi_\mu|_2^2 \leq 2f(\xi_\mu) \leq \Lambda|\xi_\mu|_2^2. \quad (28.11)$$

11. Apply to each component of $f'(\xi_\mu)$ the mean value theorem; we obtain

$$\frac{\partial f(\xi_\mu)}{\partial x_\kappa} = \sum_{v=1}^n \frac{\partial^2 f(\eta_v)}{\partial x_\kappa \partial x_v} x_v^{(\mu)}, \quad \xi_\mu = (x_1^{(\mu)}, \dots, x_n^{(\mu)}),$$

where η_κ lies on the segment joining ξ_μ with the origin, that is, in U . We can therefore write, in notation (27.9),

$$f'(\xi_\mu) = f''(\eta_1, \dots, \eta_n) \xi_\mu$$

and, solving this with respect to ξ_μ and using (27.4),

$$\xi_\mu = f''(\eta_1, \dots, \eta_n)^{-1} f'(\xi_\mu) = \kappa_\mu f''(\eta_1, \dots, \eta_n)^{-1} \Phi_\mu.$$

Therefore by (23.5)

$$|\xi_\mu|_2 \leq \kappa_\mu \Delta_2(f''(\eta_1, \dots, \eta_n)^{-1}).$$

Using now (23.12) and (28.10) we get $|\xi_\mu|_2 \leq \kappa_\mu / \lambda$, and combining this with (28.11), we obtain

$$\kappa_\mu^2 \geq \lambda^2 |\xi_\mu|_2^2 \geq 2 \frac{\lambda^2}{\Lambda} f(\xi_\mu). \quad (28.12)$$

12. Using (28.11) we have for any positive integer N

$$\frac{|\xi_{\mu+N}|_2^2}{|\xi_\mu|_2^2} \leq \frac{\Lambda}{\lambda} \frac{f(\xi_{\mu+N})}{f(\xi_\mu)}. \quad (28.13)$$

On the other hand, we have from (27.18) and (28.12)

$$\Delta_\mu \geq \frac{1 - \varepsilon^2}{2} \frac{\delta^2}{\Lambda} \left(\frac{2\lambda^2}{\Lambda} f(\xi_\mu) \right) = (1 - \varepsilon^2) \left(\frac{\lambda \delta}{\Lambda} \right)^2 f(\xi_\mu)$$

and therefore, using (27.18),

$$f(\xi_{\mu+1}) = f(\xi_\mu) - \Delta_\mu \leq f(\xi_\mu) \left(1 - \frac{1 - \varepsilon^2}{\Lambda^2} \lambda^2 \delta^2 \right).$$

This can be written, putting

$$\theta = \sqrt{1 - (1 - \varepsilon^2)(\lambda \delta / \Lambda)^2}, \quad (28.14)$$

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 (28.13) we obtain

$$\frac{|\xi_{\mu+N}|_2^2}{|\xi_\mu|_2^2} \leq \theta^{2N} \frac{A}{\lambda}, \quad (28.15)$$

If we now choose N so large that we have

$$\theta^{2N} \frac{A}{\lambda} < 1,$$

the relation (28.5) holds and Theorem 28.2 is proved.

CONDITION FOR THE REGULAR MINIMUM OF THE FUNCTION (27.3)

13. If Theorems 27.1 and 28.2 are applied to the solution of the system (27.2) using (27.3), the following theorem is useful:

Theorem 28.3. Assume that the equations (27.2) 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 (27.3) 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 (27.3)

$$\frac{1}{2}f'_{x_\kappa} = \sum_{v=1}^n f_v f'_{vx_\kappa}, \quad \frac{1}{2}f''_{vx_\kappa x_\lambda} = \sum_{v=1}^n f'_{vx_\lambda} f'_{vx_\kappa} + \sum_{v=1}^n f_v f''_{vx_\kappa x_\lambda}.$$

Then from (27.2) 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

$$\sum_{\kappa=1}^n f'_{\nu x_\kappa}(\zeta) x_\kappa = 0 \quad (\nu = 1, \dots, n).$$

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_ν vanishes at ζ . This proves our theorem.

Observe that Theorem 28.3 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 the Theorem 28.1 that the sequence z_μ constructed according to the rule of Theorem 27.1 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 $\Omega + S$, $f(z)f'(z)$ 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 (27.29) cannot lead to a complex zero of $f(z)$, if we start with a *real* z_0 . Therefore the corresponding sequence z_ν must converge to a real zero of $f'(z)$.

15. In the case of polynomials our method becomes a regular procedure for solving algebraic equations in the real and complex domain. Indeed, suppose that the sequence z_ν started from any z_0 tends to a zero ζ_1 of $f'(z)$ with $f'(\zeta_1) \neq 0$ —this is recognized by forming the values of $f(z)$ and $f'(z)$ in the course of iteration. Then, developing $f(z)$ for $z = \zeta_1 + u$ in powers of u we have, denoting by n the degree of $f(z)$,

$$f(z) = f(\zeta_1) + \sum_{\nu=1}^n \frac{u^\nu}{\nu!} f^{(\nu)}(\zeta_1).$$

Here we can, by the famous argument used by Cauchy in his proof of the fundamental theorem of algebra, find a value of $u = u_0$ such that $|f(\zeta_1 + u_0)| < |f(\zeta_1)|$. Starting with the value $\zeta_1 + u_0$ we obtain a new sequence of z , such that all $|f(z_v)|$ are $< |f(\zeta_1)|$ so that this sequence certainly does not converge to ζ_1 . If it does not converge to a zero of $f(z)$ it must converge to a zero ζ_2 of $f'(z)$, distinct from ζ_1 .

Treating ζ_2 in the same way we obtain again a sequence z , such that all $|f(z_v)|$ are $< |f(\zeta_2)| < |f(\zeta_1)|$ and this sequence cannot converge to ζ_1 or ζ_2 . Repeating the same procedure, if necessary, we find, after at the most $n - 1$ zeros of $f'(z)$, finally a sequence z , convergent to a zero of $f(z)$.

In practice the convergence to a zero of $f'(z)$ is too slow for computational purposes. One possibility to overcome this difficulty consists in getting the next zero of $f'(z)$ by the same method applied directly to $f'(z)$. However, the great number of choices then makes the programming hardly feasible. In some cases it may be worth while to compute all zeros of $f'(z)$ at the beginning. This can be used, for instance, to prepare a routine for equations of a given small degree.*

* The difficulties pointed out in this section have been overcome while the proofs of this book were corrected, too late, however, to include this result in the book.

1. By narrowing down the conditions of Theorem 28.2 we can even ensure that the convergence of the ξ_μ to ζ becomes *strictly linear*.

Theorem 29.1. *Assume about Ω , S , $f(\xi)$, and ξ_0 the conditions of Theorem 27.1. Form the sequence ξ_μ in (27.12) taking there*

$$r_\mu = \alpha_\mu \kappa_\mu \psi_\mu, \quad (29.1)$$

where ψ_μ is restricted by the conditions $|\psi_\mu| = 1$ and (27.7), and κ_μ, Φ_μ are given by (27.4), while the α_μ stay in the interval $(0, 2/\Lambda(f''(\xi_0)))$. Assume that the ξ_μ tend to a point ζ inside Ω in which $f(\xi)$ has a regular minimum. Put

$$f''(\zeta) = f_0'', \quad \Lambda(f_0'') = \Lambda_0, \quad \lambda(f_0'') = \lambda_0. \quad (29.2)$$

Take two positive numbers p, p' satisfying the conditions

$$p' < \frac{2}{\Lambda_0} - \frac{2}{\lambda_0 + \Lambda_0}, \quad p < \frac{2}{\lambda_0 + \Lambda_0}, \quad p < \frac{\lambda_0^2}{2\Lambda_0^2}, \quad (29.3)$$

$$p' < \frac{2\lambda_0}{\lambda_0 + \Lambda_0} \leq 1, \quad (29.4)$$

and put

$$\delta' = 1 - \frac{1}{2} \left(\frac{\lambda_0}{\Lambda_0} \right)^2 + p, \quad \delta'' = 1 - \frac{1}{8} \frac{p'^2}{(2 - p')^2}. \quad (29.5)$$

Assume further that the α_μ run through the interval $\langle p, (2 - p')/\Lambda_0 \rangle$ and that for the α_μ situated in the interval

$$p \leq \alpha_\mu \leq \frac{2}{\lambda_0 + \Lambda_0} \quad (29.6)$$

we have $\delta_\mu \geq \delta'$ and for α_μ in the interval

$$\frac{2}{\lambda_0 + A_0} < \alpha_\mu \leq \frac{2 - p'}{A_0} \quad (29.7)$$

we have $\delta_\mu \geq \delta''$. Then the ξ_μ converge to ζ linearly and we have more precisely

$$\lim_{\mu \rightarrow \infty} \frac{|\xi_{\mu+1} - \zeta|}{|\xi_\mu - \zeta|} \leq \begin{cases} 1 - \frac{p^2 A_0^2}{\lambda_0} & \left(p \leq \alpha_\mu \leq \frac{2}{\lambda_0 + A_0} \right), \\ 1 - \frac{p'}{2} & \left(\frac{2}{\lambda_0 + A_0} \leq \alpha_\mu \leq \frac{2 - p'}{A_0} \right). \end{cases} \quad (29.8)$$

2. Proof. Without loss of generality we can assume $\zeta = 0$. Observe that from (29.4) follows $2/(\lambda_0 + A_0) < (2 - p')/A_0$.

The formula (27.12) can be written, using (29.1), as

$$\xi_{\mu+1} - \xi_\mu = -\alpha_\mu \kappa_\mu \psi_\mu = -\alpha_\mu \kappa_\mu \Phi_\mu + \alpha_\mu \kappa_\mu (\Phi_\mu - \psi_\mu). \quad (29.9)$$

Here we have by (27.4) $\kappa_\mu \Phi_\mu = f'(\xi_\mu)$. 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 $f''(0) \equiv f_0''$. We can therefore write

$$f'(\xi_\mu) = f_0'' \xi_\mu + o(|\xi_\mu|) \quad (29.10)$$

and, introducing this into (29.9),

$$\xi_{\mu+1} = (I - \alpha_\mu f_0'') \xi_\mu + \alpha_\mu \kappa_\mu (\Phi_\mu - \psi_\mu) + o(|\xi_\mu|). \quad (29.11)$$

3. Since we have $|\Phi_\mu|_2 = |\psi_\mu|_2 = 1$, 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

$$|\Phi_\mu - \psi_\mu|_2^2 = \sum_1^n (h_i - k_i)^2 = 2 - 2 \sum_1^n h_i k_i = 2 - 2\delta_\mu,$$

$$|\Phi_\mu - \psi_\mu|_2 = \sqrt{2 - 2\delta_\mu}. \quad (29.12)$$

Further, using (29.10) and (23.13),

$$\kappa_\mu = |f'(\xi_\mu)| = |f_0''\xi_\mu| + o(|\xi_\mu|) \leq A_0|\xi_\mu| + o(|\xi_\mu|).$$

As to the first right-hand term of (29.11), we obtain

$$A(I - \alpha_\mu f_0'') = 1 - \alpha_\mu \lambda_0, \quad \lambda(I - \alpha_\mu f_0'') = 1 - \alpha_\mu A_0$$

and therefore, by virtue of (23.13),

$$\begin{aligned} |(I - \alpha_\mu f_0'')\xi_\mu| &\leq M_\mu |\xi_\mu|, \\ M_\mu &= \max(|1 - \lambda_0 \alpha_\mu|, |1 - A_0 \alpha_\mu|). \end{aligned} \quad (29.13)$$

Introducing these estimates into (29.11) we get

$$|\xi_{\mu+1}|/|\xi_\mu| \leq M_\mu + \alpha_\mu A_0 \sqrt{2 - 2\delta_\mu} + o(1). \quad (29.14)$$

4. If now α_μ satisfies (29.6), we have

$$-(1 - \alpha_\mu A_0) \leq 1 - \alpha_\mu \lambda_0$$

and therefore by (29.13)

$$M = 1 - \alpha_\mu \lambda_0.$$

Inequality (29.14) becomes, since here $\delta_\mu \geq \delta'$,

$$\begin{aligned} |\xi_{\mu+1}|/|\xi_\mu| &\leq 1 - \alpha_\mu (\lambda_0 - A_0 \sqrt{2 - 2\delta_\mu}) + o(1) \\ &\leq 1 - \alpha_\mu (\lambda_0 - A_0 \sqrt{2 - 2\delta'}) + o(1). \end{aligned}$$

Introduce here the expression for δ' from (29.5). We get

$$\lambda_0 - A_0 \sqrt{2 - 2\delta'} = \lambda_0 - \sqrt{\lambda_0^2 - 2\hat{p}A_0^2} = \frac{2\hat{p}A_0^2}{\lambda_0 + \sqrt{\lambda_0^2 - 2\hat{p}A_0^2}} > \frac{\hat{p}A_0^2}{\lambda_0}$$

and therefore, by (29.6),

$$|\xi_{\mu+1}|/|\xi_\mu| \leq 1 - \hat{p}\alpha_\mu \frac{A_0^2}{\lambda_0} + o(1) \leq 1 - \frac{\hat{p}^2 A_0^2}{\lambda_0} + o(1).$$

We obtain in the case (29.6) the first relation (29.8).

5. If, on the other hand, α_μ satisfies (29.7) the inequality

$$\alpha_\mu A_0 - 1 > 1 - \alpha_\mu \lambda_0 > 1 - \alpha_\mu A_0$$

is verified immediately and in this case it follows that

$$M_\mu = \alpha_\mu A_0 - 1.$$

Inequality (29.14) now becomes, by virtue of (29.7),

$$\begin{aligned} |\xi_{\mu+1}|/|\xi_\mu| &\leq \alpha_\mu A_0(1 + \sqrt{2 - \mu} 2\delta) - 1 + o(1) \\ &\leq (2 - p')(1 + \sqrt{2 - 2\delta''}) - 1 + o(1), \end{aligned}$$

since in our case $\delta_\mu \geq \delta''$.

On the other hand, it follows from (29.5) that

$$\begin{aligned} 2 - 2\delta'' &= \frac{p'^2}{4(2 - p')^2}, \quad 1 + \sqrt{2 - 2\delta''} = \frac{4 - p'}{2(2 - p')}, \\ (2 - p')(1 + \sqrt{2 - 2\delta''}) &= \frac{4 - p'}{2} = 2 - \frac{p'}{2}, \end{aligned}$$

and the second assertion of (29.8) follows immediately.

EXAMPLE

6. We show now by an example that if we choose t in Section 6 of Chapter 27 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 :

$$f(\xi) \equiv \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1, \quad a > b > 0. \quad (29.15)$$

Choose ξ_0 and ξ'_0 on E_0 (see Fig. 7) in the first and in the fourth quadrant, 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 $0\xi'_0$, so that we have

$$\xi_1 = \sigma \xi'_0, \quad 0 < \sigma < 1. \quad (29.16)$$

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

$$\xi_1' = \sigma \xi_0. \quad (29.17)$$

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 $0\xi_0$, $0\xi_0'$, so that

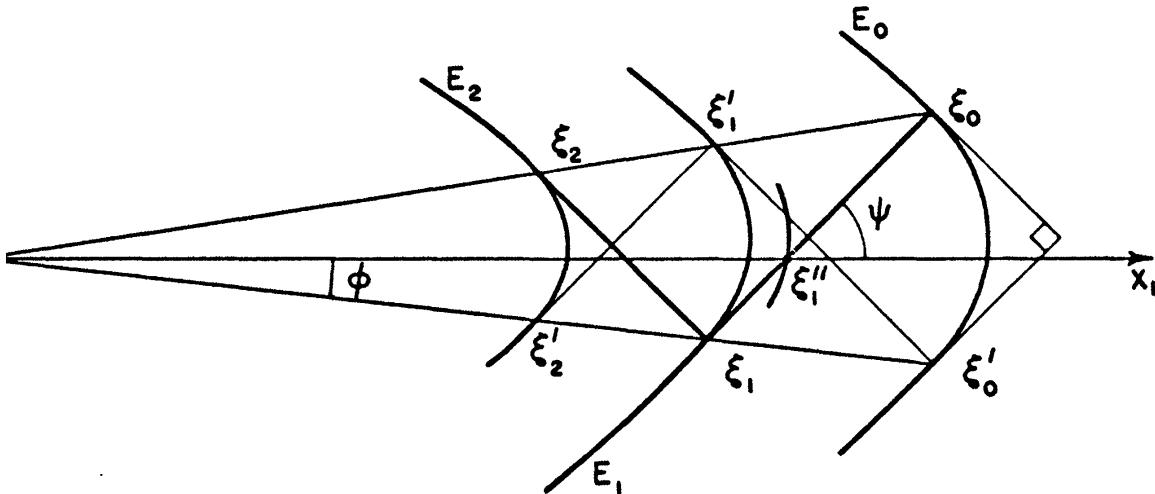


FIG. 7.

$$\xi_2 = \sigma \xi_1', \quad \xi_2' = \sigma \xi_1.$$

Continuing in the same way we have generally

$$\xi_{\nu+1} = \sigma \xi_{\nu}', \quad \xi_{\nu+1}' = \sigma \xi_{\nu}. \quad (29.18)$$

Now put

$$q_{\nu} = \frac{|\xi_{\nu+1}|}{|\xi_{\nu}|}. \quad (29.19)$$

Then it follows from (29.18), eliminating the ξ' , that

$$q_{\nu+1} = q_{\nu-1}, \quad q_{\nu} q_{\nu+1} = \sigma^2. \quad (29.20)$$

Further, we easily obtain

$$\xi_{2\nu+i} = \sigma^{2\nu} \xi_i \quad (i = 0, 1; \quad \nu = 1, 2, \dots). \quad (29.21)$$

We see that the convergence of the ξ_i to zero is not better than linear.

8. We are going to prove now that the convergence of the ξ_ν to zero is indeed *linear*, proving that

$$\frac{|\xi_{\nu+1}|}{|\xi_\nu|} \leq \frac{c^2}{b^2}; \quad c^2 = a^2 - b^2 \quad (\nu = 0, 1, \dots). \quad (29.22)$$

It is sufficient to prove (29.22) for $\nu = 0$.

Let the line $\xi_0\xi_1$ cut the real axis at ξ_1'' . Denote the angles $\xi_0\xi_1''x_1$ by ψ and $\xi_1'0x_1$ by ϕ . The equation of the normal to the ellipse E_0 at ξ_0 is

$$\frac{ux_2}{b^2} - \frac{vx_1}{a^2} = \frac{c^2}{a^2b^2} x_1 x_2,$$

where we have for the coordinates (x_1, x_2) of ξ_0 in the usual parametrical representation

$$x_1 = a \cos t, \quad x_2 = b \sin t.$$

Putting $v = 0$, we obtain $u = (c^2/a^2)x_1$,

$$|\xi_1''| = \frac{c^2}{a^2} x_1. \quad (29.23)$$

On the other hand, denoting the coordinates of ξ_0' by (x_1', x_2') , we have

$$x_1' = a \cos t', \quad x_2' = b \sin t',$$

and since the tangents at ξ_0 and ξ_0' are orthogonal, as the orthogonality condition

$$\tan t \tan |t'| = \frac{b^2}{a^2}. \quad (29.24)$$

9. We have now for the angle ψ from Fig. 7

$$\begin{aligned} \tan \psi &= \frac{x_2}{x_1 - |\xi_1''|} = \frac{x_2}{x_1} \frac{1}{1 - c^2/a^2}, \\ \tan \psi &= \frac{a^2}{b^2} \frac{b}{a} \tan t = \frac{a}{b} \tan t. \end{aligned} \quad (29.25)$$

As to the angle ϕ , we have

$$\tan \phi = \frac{|x_2'|}{x_1'} = \frac{b}{a} \tan t'$$

and by (29.24)

$$\tan \phi = \frac{b^3}{a^3} \cot t. \quad (29.26)$$

It now follows that

$$\tan(\phi + \psi) = \frac{(a/b) \tan t + (b^3/a^3) \cot t}{1 - (b^2/a^2)} = \frac{1}{c^2} \left(\frac{a^3}{b} \tan t + \frac{b^3}{a} \cot t \right).$$

Since the right-hand expression remains finite for $0 < t < \pi/2$, it follows that $\phi + \psi$ 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. 7 that both angles ϕ and ψ are acute. Therefore $\phi + \psi$ lies always between 0 and $\pi/2$ so that, as indicated in Fig. 7, 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 (29.23) 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 , is linear and also that

$$\sigma \leq \frac{c^2}{a^2}. \quad (29.27)$$

10. Figure 7 illustrates also 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.

11. Although the example (29.15) is a very special one, in the general case of a function $f(x, y)$ continuous with its derivatives of the second order in a neighborhood of ζ , we can assume, after a suitable change of coordinates, $f(x, y)$ in the form

$$f(x, y) = C \left(\frac{x^2}{a^2} + \frac{y^2}{b^2} - c + o(x^2 + y^2) \right), \quad a \geq b > 0.$$

If we have here $a > b$ it is easy to prove that (29.18) still remains true in the limit. The case $a = b$ is of course an exception.

12. This exception presents itself, however, always in the very important case, considered in the last sections of Chapters 27 and 28, of a function $f(z)$ of a complex variable $z = x + iy$. In this case we have to consider the iteration by the formula

$$z' = z - tZ(z), \quad (29.28)$$

where

$$Z(z) = 2f(z)f'(z), \quad (29.29)$$

where the real positive t is chosen in such a way as to make $|f'(z')|^2$ a minimum.

If we assume that the sequence z_n obtained in this manner converges to a zero of $f(z)$ it can be shown that this convergence is at least quadratic. The discussion of this case is, however, too complicated to be presented here.

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:

$$\begin{aligned} f(z) &= a_0 z^n + \dots + a_n, & a_0 &= 1, \\ g(z) &= b_0 z^n + \dots + b_n, & b_0 &= 1. \end{aligned} \tag{A.1}$$

Let the n zeros of $f(z)$ be x_1, \dots, x_n , those of $g(z)$, 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

$$\gamma = \max_v (|x_v|, |y_v|), \quad \Gamma = \max_{v>0} (|a_v|^{1/v}, |b_v|^{1/v}). \tag{A.2}$$

It is well known that $\gamma \leq 2\Gamma$.*

* This estimate can be improved by a result due to Lagrange but now almost forgotten (J. L. Lagrange, "De la résolution des équations numériques," Paris, Duprat, an VI (1797/8), p. 16; reprinted in *Oeuvres complètes*, Vol. VIII, p. 32): If the numbers $|a_v|^{1/v}$ are ordered increasingly, then let L_f be the sum of the n th and $(n-1)$ st of these numbers. Then all $|x_v|$ are $\leq L_f$. Using this, we obtain for our γ the bound $\gamma \leq \max[L_f, L_g]$.

Introduce now the expression

$$\varepsilon = \sqrt[n]{\sum_{v=1}^n |b_v - a_v| \gamma^{n-v}}. \quad (\text{A.3})$$

Then we have from

$$f(z) - g(z) = \sum_{v=1}^n (a_v - b_v) z^{n-v},$$

denoting by x_0 any one of the zeros x_1, \dots, x_n ,

$$|g(x_0)| = \prod_{v=1}^n |x_0 - y_v| \leq \varepsilon^n. \quad (\text{A.4})$$

3. We see that to our x_0 there exists one of the y_v , y_0 , such that

$$|x_0 - y_0| \leq \varepsilon. \quad (\text{A.5})$$

Since $f(z)$ and $g(z)$ in the above argument can be interchanged, we see that *in the ε neighborhood of any zero of one of the polynomials $f(z)$ and $g(z)$ there is always one zero of the other polynomial*. This does not, however, signify that we can always order x_v and y_v in such a way that we have $|x_v - y_v| \leq \varepsilon$ ($v = 1, \dots, n$), and this last result is indeed no longer true for $n > 2$. (See the example at the end of this appendix in Section 10.)

4. However, we are going to prove the following

Theorem. *The zeros x_v and y_v can be ordered in such a way that we have*

$$|x_v - y_v| < 2n\varepsilon \quad (v = 1, \dots, n). \quad (\text{A.6})$$

In proving (A.6), we can assume that the above ε neighborhoods of x_v and y_v are *open, i.e., circles without their circumference*. Indeed, if for a root x_0 of $f(z)$ we have for no y_v : $|x_0 - y_v| < \varepsilon$, then it follows from (A.4) that we have $|x_0 - y_v| = \varepsilon$ ($v = 1, \dots, n$). But then, since in the ε neighborhood of any x_v there is one of the y_v , we see that in this case the distance of x_v from every y_v is $\leq 3\varepsilon < 4\varepsilon$ and this implies (A.6) for $n > 1$ in this special case.

5. 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 z plane, the boundary of which consists of a finite number of regular arcs; let the functions $f(z)$, $h(z)$ 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 $f(z) + th(z)$ becomes $= 0$ on the boundary of B . Then the number $N(t)$ of the zeros of $f(z) + th(z)$ 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 $u(z) = f(z) + t_0 h(z)$ then has a positive lower limit p on the boundary of B . Then if $|\delta|$ is sufficiently small, the function $v(z) = \delta h(z)$ has its modulus less than p everywhere on the boundary of B . Since, therefore, everywhere on the boundary of B we have $|u| > |v|$, by the theorem of Rouché, $u + v$ has the same numbers of zeros inside B as the function $u(z)$.

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.

6. Proof of the Theorem. We call two zeros x_ν and x_μ of $f(z)$ *neighbors* if their distance is $\leq 2\epsilon$. Two zeros x_ν , x_μ of $f(z)$ will be called *connected*, if there is a sequence of zeros of $f(z)$ beginning with x_ν and ending with x_μ : $x_\nu, x_{\lambda_1}, x_{\lambda_2}, \dots, x_{\lambda_k}, x_\mu$, such that each zero of the sequence except x_ν is a neighbor of the immediately preceding one.

Thus all n zeros of $f(z)$ 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 ϵ neighborhoods of all zeros contained in a group 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_ν in common if $\kappa \neq \lambda$.

7. We consider now the group g_1 and assume that it contains ρ zeros of $f(z)$, counted, of course, according to their multiplicity. Consider now the function

$$g_t(z) = f(z) + t[g(z) - f(z)] \quad (0 \leq t \leq 1). \quad (\text{A.7})$$

Observe that for t of the interval $\langle 0, 1 \rangle$ $g_t(z)$ has no zero on B_1 . Indeed, such a zero would have the distance $\geq \epsilon$ from all zeros of

$f(z)$, while from our above result and the assumptions in Section 4 it follows that its distance from at least one x_ν is $\leq \varepsilon t^{1/n}$ for $t < 1$ and $< \varepsilon$ for $t = 1$.

It follows now by our lemma that the function (A.7) has a constant number of zeros inside G_1 for $0 \leq t \leq 1$. But for $t = 0$ this number is ρ , and therefore we see that $g(z)$ also has inside G_1 exactly ρ zeros y_ν . But then, since the distance of any of these y_ν from any of the x_ν in g_1 is $\leq (2\rho - 1)\varepsilon < 2n\varepsilon$, the relation (A.6) holds for x_ν and y_ν contained in G_1 . The theorem now follows immediately by applying this argument to each group g_κ .

8. In order to use (A.6), we must of course find a convenient estimate of ε . Putting

$$\delta = \max_\nu |b_\nu - a_\nu|, \quad (\text{A.8})$$

we have from (A.3), using $\gamma \leq 2\Gamma$,

$$\varepsilon^n \leq \delta \sum_{\nu=0}^{n-1} (2\Gamma)^\nu.$$

On the other hand, we have for any $u \geq 0$ the relation*

$$\sum_{\nu=0}^{n-1} u^\nu \leq \max(1, u^n) \min\left(n, \frac{1}{|1-u|}\right). \quad (\text{A.9})$$

We obtain therefore

$$\varepsilon \leq \delta^{1/n} \max(1, 2\Gamma) \sqrt[n]{\min\left(n, \frac{1}{|1-2\Gamma|}\right)}. \quad (\text{A.10})$$

Inequality (A.10) may be inconvenient because all differences $a_\nu - b_\nu$ are used with the same weight. Another estimate is obtained by assuming

$$|b_\nu - a_\nu| \leq \sigma \Gamma^\nu \quad (\nu = 1, \dots, n); \quad (\text{A.11})$$

* Inequality (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 $(1-u^n)/(1-u)$.

then we have from (A.3)

$$\begin{aligned} \varepsilon^n &\leq \sigma \sum_{\nu=1}^n \Gamma^\nu (2\Gamma)^{n-\nu} = \sigma \Gamma^n (1 + 2 + \dots + 2^{n-1}), \\ \varepsilon &\leq 2\Gamma \sigma^{1/n}. \end{aligned} \quad (\text{A.12})$$

9. 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

$$y_1 = 1.02681, \quad y_2 = 0.97389, \quad y_{3,4} = 0.99965 \pm i0.026455.$$

Take the polynomial (A.13) as $g(z)$, and $(z - 1)^4$ as $f(z)$. Then all x_ν are 1 and $\max_\nu |y_\nu - x_\nu| = 0.02681 = y_1 - 1$, while $\varepsilon^4 = 49 \cdot 10^{-8} y_1^2 = (y_1 - 1)^4$, $\varepsilon = y_1 - 1$. We have therefore in this case the equality sign in (A.5).

10. We show finally in an example that it is impossible to deduce for $n = 3$ the relation (A.6) with the coefficient 1 instead of $2n$. 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

$$\begin{aligned} x_1 &= -\sqrt[3]{2}, \quad x_2 = -\sqrt[3]{2}, \quad x_3 = \frac{1}{\sqrt[3]{4}} = 0.62996, \\ y_1 &= 0, \quad y_2 = 0, \quad y_3 = -\frac{3}{2}\sqrt[3]{2} = -1.89 \end{aligned}$$

and it is obviously impossible to reorder y_1, y_2, y_3 in such a way that we get $|y_\nu - x_\nu| \leqq \varepsilon = 1$.

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. *Consider two polynomials*

$$f(z) = a_0 z^n + a_1 z^{n-1} + \dots + a_n, \quad g(z) = b_0 z^n + b_1 z^{n-1} + \dots + b_n \quad (\text{B.1})$$

and assume that $a_0 a_n \neq 0$, i.e., that $f(z)$ has n finite zeros x_ν ($\nu = 1, \dots, n$), all $\neq 0$. Assume further that for a certain positive τ with

$$4n\tau^{1/n} \leq 1 \quad (\text{B.2})$$

we have

$$|b_\nu - a_\nu| \leq \tau |a_\nu| \quad (\nu = 0, 1, \dots, n). \quad (\text{B.3})$$

Then the n zeros y_1, \dots, y_n of $g(z)$ can be ordered in such a way that we have

$$\left| \frac{y_\nu}{x_\nu} - 1 \right| < 8n\tau^{1/n} \quad (\nu = 1, \dots, n). \quad (\text{B.4})$$

3. Proof. Put $|x_\nu| = r_\nu$ ($\nu = 1, \dots, n$). We can assume that

$$r_1 \geq r_2 \geq \dots \geq r_n, \quad |y_1| \geq |y_2| \geq \dots \geq |y_n|.$$

Consider the polynomial

$$F(z) = |a_0| \prod_{\nu=1}^n (z + r_\nu) = S_0 z^n + \dots + S_n. \quad (\text{B.5})$$

We have obviously

$$S_0 = |a_0|, \quad S_\nu \geq |a_\nu| \quad (\nu = 1, \dots, n). \quad (\text{B.6})$$

It follows now from (B.3) and (B.6) that

$$|b_\nu - a_\nu| \leq \tau S_\nu \quad (\nu = 0, 1, \dots, n). \quad (\text{B.7})$$

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_ν , for which we have

$$\min_\nu \left| 1 - \frac{y_0}{x_\nu} \right| = \left| 1 - \frac{y_0}{x_0} \right|. \quad (\text{B.8})$$

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 $S_0 = |a_0|$,

$$\prod_{\nu=1}^n |y_0 - x_\nu| \leq \tau \prod_{\nu=1}^n (r + r_\nu). \quad (\text{B.9})$$

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 $r_\nu = |x_\nu|$ 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_σ 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 , $\sigma = l + 1, \dots, k$,

$$\prod_{\sigma=l+1}^k \left| 1 - \frac{y_0}{x_\sigma} \right| \leq \tau \prod_{\sigma=l+1}^k \left(1 + \frac{r}{r_\sigma} \right) \left(\frac{1 + q}{1 - q} \right)^{n-s} \quad (\text{B.14})$$

But here we have for any r_σ from the third class

$$1 + \frac{r}{r_\sigma} \leq 1 + \frac{1}{q} = \frac{1 + q}{q}.$$

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

$$\left| 1 - \frac{y_0}{x_0} \right|^s \leq \left(\frac{1-q}{\cdot q} \right)^s, \quad \left| 1 - \frac{y_0}{x_0} \right| \leq \frac{1-q}{q}.$$

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_ν be the inside of the circle around x_ν with the radius ε_ν ($\nu = 1, \dots, n$). Then it follows from (B.16) that each zero y_μ of $g(z)$ is contained in one of the *open* neighborhoods U_ν of the x_ν .

8. We proceed now as in Appendix A. We call two zeros x_ν, x_μ *neighbors* if we have

$$|x_\nu - x_\mu| \leq \varepsilon_\nu + \varepsilon_\mu. \quad (\text{B.19})$$

From (B.19) and (B.18) we have

$$x_\mu = \frac{1 + \theta \varepsilon}{1 + \theta' \varepsilon} x_\nu, \quad |\theta|, |\theta'| \leq 1. \quad (\text{B.20})$$

Two zeros x_{ν_1}, x_{ν_l} will be called *connected* if there exists a sequence $x_{\nu_1}, x_{\nu_2}, \dots, x_{\nu_l}$ 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

$$x_{\nu_l} = x_{\nu_1} \prod_{\lambda=1}^{l-1} \frac{1 + \theta_\lambda \varepsilon}{1 + \theta'_{\lambda'} \varepsilon}, \quad |\theta_\lambda|, |\theta'_{\lambda'}| \leq 1, \quad (\text{B.21})$$

and therefore *a fortiori*

$$x_{\nu_l} = x_{\nu_1} \prod_{\lambda=1}^{n-1} \frac{1 + \theta_\lambda \varepsilon}{1 + \theta'_{\lambda'} \varepsilon}, \quad |\theta_\lambda|, |\theta'_{\lambda'}| \leq 1, \quad (\text{B.22})$$

since the $\theta_\lambda, \theta'_{\lambda'}$ with $\lambda \geq l$ can be taken as 0.

9. We decompose all x_ν into groups g_1, \dots, g_k , such that two x_ν from the same group are always *connected*, while two x_ν of different groups are never connected. For each group we form the sum set of the corresponding neighborhoods U_ν , $G_\kappa = \sum_{x_\nu \in g_\kappa} U_\nu$. 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_ν 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 we replace in our result (B.16) 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 the 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 from 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_ν contained in G_1 is the same as the number of the x_ν inside G_1 , and the analogous result holds of course for every G_κ .

11. We reorder the y_ν in such a way that all y_ν contained in a set G_κ have the same indices as the x_ν from the same G_κ . This obviously can be done in many ways if a G_κ contains more than one x_ν .

Consider now a zero y_ν 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 $|y_\nu - x_\mu| \leq r_\mu \varepsilon$, $y_\nu = x_\mu(1 + \theta\varepsilon)$, $|\theta| \leq 1$. On the other hand x_μ is connected with x_ν . It follows therefore from (B.22) that

$$y_\nu = x_\nu(1 + \theta\varepsilon) \prod_{\mu=1}^{n-1} \frac{1 + \theta_{\mu\varepsilon}}{1 + \theta_{\mu'}\varepsilon},$$

$$\frac{y_\nu}{x_\nu} - 1 = (1 + \theta\varepsilon) \prod_{\mu=1}^{n-1} \frac{1 + \theta_{\mu\varepsilon}}{1 + \theta_{\mu'}\varepsilon} - 1.$$

The right-hand expression is majorized by $[(1 + \varepsilon)^n / (1 - \varepsilon)^{n-1}] - 1$ and we have

$$\left| \frac{y_\nu}{x_\nu} - 1 \right| \leq \phi(\varepsilon), \quad \phi(\varepsilon) = \frac{(1 + \varepsilon)^n}{(1 - \varepsilon)^{n-1}} - 1. \quad (\text{B.24})$$

12. We now put

$$\delta = \frac{2\tau'}{1 - \tau'}, \quad \tau' = \tau^{1/n}, \quad (\text{B.25})$$

and discuss first $\phi(\delta)$. Using (B.2), $\tau' \leq 1/4n$, we have by (B.25)

$$\phi(\delta) = \frac{(1 + \tau')^n}{(1 - \tau')(1 - 3\tau')^{n-1}} - 1 = \tau' \psi(\tau'),$$

where $\psi(\tau')$ is a power series with *positive* coefficients. We have therefore

$$\psi(\tau') \leq \psi\left(\frac{1}{4n}\right), \quad \phi(\delta) \leq \psi\left(\frac{1}{4n}\right)\tau'.$$

On the other hand

$$\psi\left(\frac{1}{4n}\right) = 4n \left[\frac{(1 + 1/4n)^n}{(1 - 3/4n)^{n-1}(1 - 1/4n)} - 1 \right].$$

The first term in the square brackets is here

$$\begin{aligned} \frac{4n}{4n-1} \left(1 + \frac{1}{4n}\right)^n \left(1 + \frac{3}{4n-3}\right)^{n-1} &< \left(1 + \frac{1}{4n-1}\right) \\ &\times \left(1 + \frac{3/4}{n-1}\right)^{n-1} \left(1 + \frac{1/4}{n}\right)^n \\ &< \left(1 + \frac{1}{4n-1}\right) e^{3/4} e^{1/4}, \end{aligned}$$

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 $\psi(1/4n) < 7.9n$. We have therefore

$$\phi(\delta) < 7.9n\tau'. \quad (\text{B.26})$$

Observe now that by (B.17) ε can be chosen arbitrarily between δ and 1. Since $\phi(\varepsilon)$ is continuous for $\varepsilon < 1$, we can therefore from the beginning choose ε so near δ that we have $\phi(\varepsilon) < 8n\tau'$ and the inequality (B.4) is proved.

13. It may be added that if in the relation (B.4) the x_ν, y_ν are replaced by $|x_\nu|, |y_\nu|$, 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 ν with $a_\nu = 0$: $b_\nu = a_\nu$; and if a_ν is very small, we have from (B.3) only a very small range of values for b_ν . 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), $|b_0 - 1| \leq \tau, |b_n + 1| \leq \tau, b_\nu = 0$ ($0 < \nu < n$). On the other hand, in this case we have $S_\nu = \binom{n}{\nu}$ and using (B.7) our conditions for b_ν become

$$|b_0 - 1| \leq \tau, \quad |b_n + 1| \leq \tau, \quad |b_\nu| \leq \binom{n}{\nu} \tau \quad (0 < \nu < n - 1).$$

However, in order to apply (B.7) in this way, we would have to know all $|x_\nu|$, which is usually not the case.

15. We will now treat the problem of how *positive* constants T_ν can be formed directly from the $|a_\nu|$ in such a way that we have

$$|a_\nu| \leq T_\nu \leq S_\nu \quad (\nu = 1, \dots, n - 1), \quad (\text{B.27})$$

and that therefore the conditions

$$|b_\nu - a_\nu| \leq \tau T_\nu, \quad (\text{B.28})$$

can be used instead of (B.3).

* Compare A. Ostrowski, *Mathematische Miszellen* XXIV, "Zur relativen Stetigkeit von Wurzeln algebraischer Gleichungen," *Jahresber. Deut. Math.-Ver.* **58**, 98–102, 1956.

For this purpose we need an inequality which goes back to Newton.
If the polynomial

$$\Phi(z) = q_0 z^n + \binom{n}{1} q_1 z^{n-1} + \dots + \binom{n}{\nu} q_\nu z^{n-\nu} + \dots + q_n, \quad q_0 \neq 0, \quad (\text{B.29})$$

has real coefficients and n real roots, then we have

$$q_\nu^2 \geq q_{\nu-1} q_{\nu+1} \quad (\nu = 1, \dots, n-1). \quad (\text{B.30})$$

This is obvious for $n = 2$ and is proved in the general case by induction using Rolle's theorem.*

16. We assume from now on that

$$a_0 = 1, \quad (\text{B.31})$$

write $F(z)$ from (B.5) in the form

$$F(z) = \sum_{\nu=0}^n \binom{n}{\nu} s_\nu z^{n-\nu}, \quad \binom{n}{\nu} s_\nu = S_\nu, \quad s_0 = S_0 = 1, \quad (\text{B.32})$$

and put

$$|a_\nu| = \binom{n}{\nu} k_\nu \quad (\nu = 0, \dots, n); \quad (\text{B.33})$$

then we have from (B.30)

$$s_\nu^2 \geq s_{\nu-1} s_{\nu+1} \quad (\nu = 1, \dots, n-1). \quad (\text{B.34})$$

Suppose now that for a certain index ν we have $a_\nu = 0$, $a_{\nu-1} a_{\nu+1} \neq 0$. Then it follows from (B.34) and (B.6) that $\sqrt[k_{\nu-1} k_{\nu+1}]{|a_{\nu-1} a_{\nu+1}|} \leq s_\nu$ and therefore, if we use (B.33),

$$\sqrt{\left(1 + \frac{1}{\nu}\right)\left(1 + \frac{1}{n-\nu}\right)|a_{\nu-1} a_{\nu+1}|} \leq S_\nu.$$

* Compare, for instance, G. H. Hardy, J. E. Littlewood, G. Polya, *Inequalities*, p. 53, 54, Cambridge Univ. Press, 1934.

We can therefore replace in the corresponding relation (B.3) $|a_\nu|$ in the right-hand expression by

$$\sqrt{\left(1 + \frac{1}{\nu}\right)\left(1 + \frac{1}{n-\nu}\right)|a_{\nu-1}a_{\nu+1}|}.$$

If a sequence of a_ν vanishes, e.g.,

$$a_{\nu_1} \neq 0, \quad a_{\nu_1+1} = a_{\nu_1+2} = \dots = a_{\nu_2-1} = 0, \quad a_{\nu_2} \neq 0, \quad (\text{B.35})$$

then it is easy to show that the corresponding $|a_\nu|$ in the inequality (B.3) can be replaced by the expressions

$$\binom{n}{\nu} \binom{n}{\nu_2}^{(\nu_1-\nu)/(\nu_2-\nu_1)} \binom{n}{\nu_1}^{(\nu-\nu_2)/(\nu_2-\nu_1)} a_{\nu_1}^{(\nu-\nu_1)/(\nu_2-\nu_1)} a_{\nu_1}^{(\nu_2-\nu)/(\nu_2-\nu_1)}. \quad (\text{B.36})$$

However, in this case it will be preferable to use a systematic geometrical approach to our problem of building up the expressions for the T_ν .

17. Put

$$\sigma_\nu = \ln s_\nu, \quad \kappa_\nu = \ln k_\nu \quad (\nu = 0, 1, \dots, n), \quad (\text{B.37})$$

where, if $k_\nu = 0$, the corresponding κ_ν is $-\infty$. Then we have from (B.34) and (B.6)

$$\sigma_\nu \geqq \frac{\sigma_{\nu-1} + \sigma_{\nu+1}}{2}, \quad (\text{B.38})$$

$$\kappa_\nu \leqq \sigma_\nu. \quad (\text{B.39})$$

The inequality (B.38) has a very simple geometric interpretation. If we mark in the (ν, σ) plane the points with the coordinates (ν, σ_ν) , $\nu = 0, 1, \dots, n$, and connect them by rectilinear segments (see Fig. 8), we obtain a polygonal line P_S , which is *convex from above*. On the other hand, if we mark the points (ν, κ_ν) ($\nu = 0, 1, \dots, n$), these points lie by (B.39) *below* P_S or *on* this polygonal line.

We draw now 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 (ν, κ_ν) , while all other points (ν, κ_ν) lie either on P_T or below. It follows from (B.3) that no point of P_T is situated above

P_S . Denote for each ν , $\nu = 0, 1, \dots, n$, by τ_ν the ordinate of the point of P_T corresponding to the abscissa ν ; then we have obviously

$$\kappa_\nu \leqq \tau_\nu \leqq \sigma_\nu, \quad (\text{B.40})$$

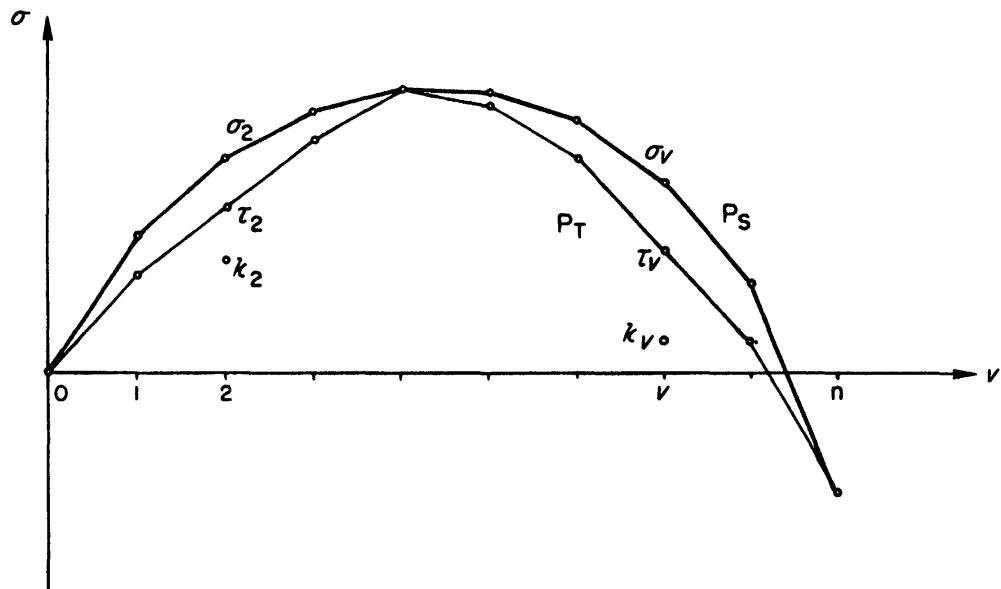


FIG. 8.

and therefore, if we put

$$\binom{n}{\nu} e^{\tau_\nu} = T_\nu, \quad (\text{B.41})$$

(B.27) is certainly satisfied. In this way we can replace the conditions (B.3) by the conditions (B.28), where the constants T_ν 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 $x = \phi(y)$ be the inverse function of $y = f(x)$. Then we have seen in Chapter 2 that we can write

$$\phi^{(n)}(y) = y_1^{-(2n-1)} X(y_1, \dots, y_n), \quad (\text{C.1})$$

where X_n is a polynomial in y_1, \dots, y_n :

$$X_n = \sum a_{\alpha_1 \dots \alpha_n} y_1^{\alpha_1} \dots y_n^{\alpha_n}, \quad (\text{C.2})$$

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

$$a_{\alpha_1 \dots \alpha_n} = (-1)^{n-\alpha_1-1} \frac{(2n-\alpha_1-2)!}{\alpha_2!(2!)^{\alpha_2}\alpha_3!(3!)^{\alpha_3} \dots \alpha_n!(n!)^{\alpha_n}}. \quad (\text{C.4})$$

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 $\phi^{(n)}(y)$. 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

$$\Phi^{(n)}(y) = \sum (-1)^{n+\beta_1} \frac{(\beta_1 - 1)!}{\beta_2!(2!)^{\beta_2} \dots \beta_n!(n!)^{\beta_n}} y_1^{-\beta_1} y_2^{\beta_2} \dots y_n^{\beta_n}, \quad (\text{C.5})$$

where the β_ν are subject to the conditions

$$\beta_1 \leq 2n - 1, \quad \beta_\nu \geq 0 \quad (\nu > 1), \quad \sum_{\nu > 1} \beta_\nu = \beta_1 - n, \quad (\text{C.6})$$

$$\sum_{\nu > 1} \nu \beta_\nu = \beta_1 - 1. \quad (\text{C.7})$$

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) $\Phi'(y) = 1/y_1$.

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) $\Phi^{(n)}(y)$ 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

$$T_{\gamma_1 \dots \gamma_{n+1}} = T \equiv (-1)^{n+1+\gamma_1} \frac{(\gamma_1 - 1)!}{\gamma_2!(2!)^{\gamma_2} \dots \gamma_n!(n!)^{\gamma_n} \gamma_{n+1}![(n+1)!]^{\gamma_{n+1}}}, \quad (\text{C.8})$$

where we put

$$\Phi^{(n+1)}(y) = \sum T_{\gamma_1 \dots \gamma_{n+1}} y_1^{-\gamma_1} y_2^{\gamma_2} \dots y_{n+1}^{\gamma_{n+1}}. \quad (\text{C.9})$$

In this proof we can assume without loss of generality that $\gamma_{n+1} = 0$.

Our γ_ν satisfy the conditions which are analogous to (C.6) and (C.7), and the condition corresponding to (C.7) is

$$\sum_{\nu \geq 2} \nu \gamma_\nu = \gamma_1 - 1. \quad (\text{C.10})$$

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

$$\Phi^{(n+1)}(y) = \frac{1}{y_1} \left(y_2 \frac{\partial}{\partial y_1} + y_3 \frac{\partial}{\partial y_2} + \dots + y_n \frac{\partial}{\partial y_{n-1}} + y_{n+1} \frac{\partial}{\partial y_n} \right) \Phi^{(n)}(y), \quad (\text{C.11})$$

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

$$y_1^{-\beta_1} y_2^{\beta_2} \dots y_n^{\beta_n}, \quad (\text{C.12})$$

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

$$-(\gamma_1 - 2)(-1)^{n+\gamma_1-2} \frac{(\gamma_1 - 3)!}{(\gamma_2 - 1)! (2!)^{\gamma_2-1} \gamma_3! (3!)^{\gamma_3} \dots \gamma_n! (n!)^{\gamma_n}},$$

and this is $[2\gamma_2/(\gamma_1 - 1)]T$.

5. Consider now for a $\nu > 1$ the contribution arising from $(y_{\nu+1}/y_1)\partial/\partial y_\nu$. This operation must be applied to the term of (C.5) corresponding to the monomial (C.12) with

$$\begin{aligned} \beta_1 &= \gamma_1 - 1, & \beta_\nu &= \gamma_\nu + 1, & \beta_{\nu+1} &= \gamma_{\nu+1} - 1, \\ \beta_\mu &= \gamma_\mu, & (\mu &\neq 1, \nu, \nu + 1). \end{aligned}$$

The corresponding contribution to $T_{\gamma_1 \dots \gamma_{n+1}}$ is then

$$\frac{(\gamma_\nu + 1)(-1)^{n+\gamma_1-1}(\gamma_1 - 2)!}{\dots (\gamma_\nu + 1)! (\nu!)^{\nu+1} (\gamma_{\nu+1} - 1)! [(\nu + 1)!]^{\nu+1-1} \dots},$$

where the factors in the denominator corresponding to $\gamma_2, \dots, \gamma_{\nu-1}$, $\gamma_{\nu+2}, \dots, \gamma_n$ are the same as in (C.8) and are not explicitly written down; this expression is obviously

$$\frac{(\nu + 1)! \gamma_{\nu+1}}{\nu! (\gamma_1 - 1)} T = \frac{(\nu + 1) \gamma_{\nu+1}}{\gamma_1 - 1} T.$$

Forming now the sum of all contributions, we get

$$\frac{T}{\gamma_1 - 1} \left(\sum_{\nu=2}^n (\nu + 1) \gamma_{\nu+1} + 2\gamma_2 \right). \quad (\text{C.13})$$

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 $y_3 = y_4 = \dots = 0$ and in (C.5) $\beta_2 = n - 1$, $\beta_1 = 2n - 1$,

$$\Phi^{(n)}(y) = (-1)^{n-1} 3 \cdot 5 \dots (2n-3) \frac{y_2^{n-1}}{y_1^{2n-1}}. \quad (\text{C.14})$$

If $f(x)$ is a *cubic polynomial*, we have $y_4 = y_5 = \dots = 0$ and $\Phi^{(n)}(y)$ becomes

$$\Phi^{(n)}(y) = \sum (-1)^{n+\beta_1} \frac{(\beta_1 - 1)!}{\beta_2! 2^{\beta_2} \beta_3! 6^{\beta_3}} y_1^{-\beta_1} y_2^{\beta_2} y_3^{\beta_3}$$

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 $\beta_3 = \nu$ and obtain

$$\beta_2 = n - 2\nu - 1, \quad \beta_1 = 2n - \nu - 1.$$

ν is ≥ 0 but must remain $\leq (n-1)/2$, since β_2 is ≥ 0 . Our expression for $\Phi^{(n)}(y)$ now becomes

$$\Phi^{(n)}(y) = y_2^{n-1} y_1^{1-2\nu} \sum_{\nu=0}^{\frac{1}{2}(n-1)} (-1)^{n-\nu-1} \frac{(2n-\nu-2)!}{(n-2\nu-1)! \nu! 2^{n-\nu-1} 3^\nu} \left(\frac{y_1 y_3}{y_2^2} \right)^\nu. \quad (\text{C.15})$$

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:

$$F(P_\nu), \quad G(P_\nu) \quad (\nu = 1, 2, 3).$$

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

$$L_1(P_\nu) = F(P_\nu), \quad L_2(P_\nu) = G(P_\nu) \quad (\nu = 1, 2, 3) \quad (\text{D.1})$$

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 ν from any of the equations (D.2) the corresponding equation (D.1); thus we obtain

$$\begin{aligned} a(x - x_\nu) + b(y - y_\nu) &= -F(P_\nu) \\ e(x - x_\nu) + f(y - y_\nu) &= -G(P_\nu) \end{aligned} \quad (\nu = 1, 2, 3). \quad (\text{D.3})$$

2. The result of the elimination of a, b, e, f from (D.3) amounts to the statement that the (3×4) matrix

$$(x - x_\nu \quad y - y_\nu \quad F(P_\nu) \quad G(P_\nu)) \quad (\nu = 1, 2, 3)$$

* *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, Nr. 120, pp. 186–188, 1809; C. F. Gauss, *Werke*, Vol. VII, pp. 150–152.

has rank ≤ 2 . We obtain the two equations

$$\begin{vmatrix} x - x_\nu \\ F(P_\nu) \\ G(P_\nu) \end{vmatrix} = 0, \quad \begin{vmatrix} y - y_\nu \\ F(P_\nu) \\ G(P_\nu) \end{vmatrix} = 0, \quad (\text{D.4})$$

which were given by Gauss in order to obtain from the three given approximations P_ν to the solution in question a fourth improved one.

If we assume that the determinant

$$\Delta = \begin{vmatrix} 1 \\ F(P_\nu) \\ G(P_\nu) \end{vmatrix} \quad (\nu = 1, 2, 3) \quad (\text{D.5})$$

is $\neq 0$, the solution of (D.4) is given by

$$x = \frac{1}{\Delta} \begin{vmatrix} x_\nu \\ F(P_\nu) \\ G(P_\nu) \end{vmatrix}, \quad y = \frac{1}{\Delta} \begin{vmatrix} y_\nu \\ F(P_\nu) \\ G(P_\nu) \end{vmatrix} \quad (\nu = 1, 2, 3). \quad (\text{D.6})$$

3. The condition $\Delta \neq 0$ is equivalent to the condition that the three points $[F(P_\nu), G(P_\nu)]$ ($\nu = 1, 2, 3$) are not collinear. On the other hand, we obtain from (D.5), using (D.3), the identical relation

$$\Delta = (af - be) \begin{vmatrix} 1 \\ x_\nu \\ y_\nu \end{vmatrix} \quad (\nu = 1, 2, 3). \quad (\text{D.7})$$

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.

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

$$\Psi(y) = \frac{y\psi[\psi(y)] - \psi(y)^2}{y - 2\psi(y) + \psi[\psi(y)]}. \quad (\text{E.1})^\dagger$$

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*[‡]

* 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 "Aitkin's method." In this method the use of the expression (E.1) serves to derive from a given complete iteration sequence a better convergent one. However, by this "Aitkin's transformation" a linearly convergent sequence remains, as a rule, linearly convergent after transformation.

‡ F. A. Willers, *Z. angew. Math. Mech.* **28**, 125–126 (1948).

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 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 $z_n = x_n - \zeta$ 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 $\alpha \neq 0, \alpha \neq 1$, 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 $x \rightarrow 0$

$$\begin{aligned} \psi(x) &= \alpha x + o(x), \\ \psi[\psi(x)] &= \alpha[\alpha x + o(x)] + o[\alpha x + o(x)] = \alpha^2 x + o(x), \\ \psi[\psi(x)] - 2\psi(x) + x &= (\alpha - 1)^2 x + o(x), \\ \psi(x)^2 &= \alpha^2 x^2 + o(x^2), \\ x\psi[\psi(x)] - \psi(x)^2 &= o(x^2), \end{aligned}$$

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 $\psi(x) - \alpha(x - \zeta)$ as $x \rightarrow \zeta$.

* 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.

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.

6. Proof. Without loss of generality assume $\zeta = 0$. Then we have from (E.2)

$$\psi(x) = \alpha x + E(x)|x|^\lambda,$$

$$\begin{aligned} \psi[\psi(x)] &= \alpha[\alpha x + E(x)|x|^\lambda] + E[\psi(x)]|\alpha x + E(x)|x|^\lambda|^\lambda \\ &= \alpha^2 x + \{\alpha E(x) + |\alpha|^\lambda E[\psi(x)]\}|x|^\lambda + o(x^\lambda). \end{aligned}$$

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*

$$\Psi(x) - \zeta = O(|x - \zeta|^{2\lambda - 1}), \quad (\text{E.8a})$$

$$\Psi(x) - \zeta = o(|x - \zeta|^{2\lambda - 1}). \quad (\text{E.8b})$$

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,

$$x\psi[\psi(x)] - \psi(x)^2 = -E(x)^2|x|^{2\lambda} + o(|x|^{2\lambda}).$$

* 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)," *Archive for Rational Mechanics and Analysis*, Vol. 4 (1959), pp. 152–165.

From (E.1) we obtain finally

$$\Psi(x) = -E(x)^2 x |x|^{2\lambda-2} + o(|x|^{2\lambda-1}), \quad (\text{E.9})$$

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 $x \geq 0$ 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) $g[\psi(x)] - g(x)$, 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)

$$g'(\xi) = \psi'(\xi) - 1 = T(\xi)|\xi|^{\lambda-1} = \gamma|x|^{\lambda-1} + o(|x|^{\lambda-1}).$$

and therefore 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, the 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

$$f(x) \equiv (x - \xi)(x - \eta) = 0, \quad (\text{F.1})$$

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

$$x_{v+1} = x_v - \frac{(x_v - \xi)(x_v - \eta)}{2x_v - \xi - \eta} = x_v + \frac{a_v b_v}{a_v + b_v},$$

$$a_{v+1} = a_v - \frac{a_v b_v}{a_v + b_v} = \frac{a_v^2}{a_v + b_v},$$

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 $\xi \neq \eta$,

$$a_\nu = \frac{a^{2^\nu}(a-b)}{a^{2^\nu} - b^{2^\nu}}, \quad b_\nu = \frac{b^{2^\nu}(a-b)}{a^{2^\nu} - b^{2^\nu}} \quad (\xi \neq \eta) \quad (\text{F.6})$$

and, if $\xi = \eta$, $a = b$,

$$a_\nu = b_\nu = \frac{a}{2^\nu} \quad (\xi = \eta). \quad (\text{F.6a})$$

3. Indeed, (F.6) is obvious for $\xi \neq \eta$ and $\nu = 0$. Suppose that (F.6) is true for a ν ; we have then from (F.5)

$$a_{\nu+1} = \frac{a^{2^\nu+1}(a-b)^2/(a^{2^\nu} - b^{2^\nu})^2}{(a^{2^\nu} + b^{2^\nu})(a-b)/(a^{2^\nu} - b^{2^\nu})} = \frac{a^{2^\nu+1}(a-b)}{a^{2^{\nu+1}} - b^{2^{\nu+1}}}.$$

Since the value of $b_{\nu+1}$ is obtained by symmetry, (F.6) is proved by induction.

In the case $\xi = \eta$, $a_\nu = b_\nu$, (F.5) becomes

$$a_{\nu+1} = b_{\nu+1} = \frac{a_\nu}{2},$$

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

$$\alpha_\nu = \xi - x_\nu \rightarrow 0 \quad (\nu \rightarrow \infty).$$

For $\xi \neq \eta$ we have under the hypothesis (F.2) $|a| < |b|$,

$$a_\nu \sim (b-a) \left(\frac{a}{b} \right)^{2^\nu} \rightarrow 0 \quad (\nu \rightarrow \infty).$$

If, on the other hand, $|a| = |b|$, $a \neq b$, we see from (F.6) that $|a_\nu| = |b_\nu|$, and in the case of convergence both a_ν and b_ν must tend to 0. But, on the other hand, it follows from (F.6) that $a_\nu - b_\nu = a - b$, and if a_ν and b_ν 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 7.1 are satisfied.

Keeping the notation (F.4), put $\phi = b/a$; then we have either $\phi = 1$ or $|\phi| \neq 1$. From Section 2 we obtain

$$h_v = x_{v+1} - x_v = \frac{a_v b_v}{a_v + b_v} = a^{2^v} b^{2^v} \frac{(a - b)}{a^{2^v+1} - b^{2^v+1}} \quad (|\phi| \neq 1),$$

$$h_v = \frac{a}{2^{v+1}} \quad (\phi = 1).$$

Therefore, if we use $\phi = b/a$,

$$h_v = \begin{cases} \frac{a}{2^{v+1}} & (\phi = 1) \\ a\phi^{2^v} \frac{\phi - 1}{\phi^{2^v+1} - 1} & (|\phi| \neq 1). \end{cases} \quad (\text{F.7})$$

On the other hand, by (F.4)

$$f'(x_v) = 2x_v - \xi - \eta = -(a_v + b_v) = -(a^{2^v} + b^{2^v}) \frac{a - b}{a^{2^v} - b^{2^v}}$$

$$= -a(\phi - 1) \frac{\phi^{2^v} + 1}{\phi^{2^v} - 1} \quad (|\phi| \neq 1),$$

and for $\phi = 1$,

$$f'(x_v) = -\frac{a}{2^{v+1}} \quad (\phi = 1).$$

Therefore, since in our case the number M of Theorem 7.1 is 2,

$$\frac{2Mh_v}{f'(x_v)} = -4 \frac{\phi^{2^v}(\phi^{2^v} - 1)}{(\phi^{2^v+1} - 1)(\phi^{2^v} + 1)} = -4 \frac{\phi^{2^v}}{(\phi^{2^v} + 1)^2}$$

$$= -\frac{4}{(\phi^{2^v-1} + \phi^{-2^v-1})^2} \quad (|\phi| \neq 1), \quad (\text{F.8})$$

while for $\phi = 1$ we have

$$\frac{2Mh_v}{f'(x_v)} = -1 \quad (\phi = 1). \quad (\text{F.8a})$$

6. We see that in the case $\rho = 1$ we have for each ν the limiting case $|2Mh_\nu/f'(x_\nu)| = 1$. What happens in the case $\rho \neq 1$?

From (F.8) it follows that the modulus of the left-hand expression tends to 0 as $\nu \rightarrow \infty$. Therefore, the conditions of Theorem 7.1 are satisfied from a certain ν on. On the other hand, choosing ρ conveniently, we can insure that the conditions of Theorem 7.1 do not hold for $\nu = 0, 1, \dots, N$, where N can be chosen as large as we like. Indeed, if we take $\rho = \rho e^{i\alpha}$, we have

$$|\rho^{2^\nu-1} + \rho^{-2^\nu-1}|^2 = \rho^{2^\nu} + \rho^{-2^\nu} + 2 \cos 2^\nu \alpha.$$

Take here $\alpha = \pi/2^N$, then we obtain for $\nu = 0, 1, \dots, N$

$$|\rho^{2^\nu-1} + \rho^{-2^\nu-1}|^2 = \rho^{2^\nu} + \rho^{-2^\nu} + 2 \cos \frac{\pi}{2^N}.$$

But if we take $\rho > 1$ sufficiently near to 1, we can insure that the inequality $\rho^{2^\nu} + \rho^{-2^\nu} < 4 - 2 \cos(\pi/2^N)$ holds for $\nu = 0, 1, \dots, N$, and the modulus of the right-hand expression in (F.8) remains > 1 for all these values of ν .

7. We further ask whether the modulus of the expression (F.8) can be 1 for one ν or two consecutive values of ν . 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) $|\rho| \neq 1$, the modulus of (F.8) certainly cannot be 1 for two consecutive values of ν .

We see in particular that if for a value of ν the expression $|2Mh_\nu/f'(x_\nu)|$ is = 1, this expression becomes < 1 for all greater ν unless our quadratic polynomial has a double zero.

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 $f'(x_\nu)$ 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

$$x_{\nu+1} = x_\nu - cf(x_\nu)$$

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

$$\begin{aligned} x_1 &= 2.1 \\ x_2 &= 2.094\ 568\ 1 \\ x_3 &= 2.094\ 551\ 481\ 72 \end{aligned}$$

II

$$\begin{aligned} x_1 &= 2.1 \\ x_2 &= 2.093\ 9 \\ x_3 &= 2.094\ 627 \\ x_4 &= 2.094\ 542\ 7 \\ x_5 &= 2.094\ 552\ 5 \\ x_6 &= 2.094\ 551\ 363 \end{aligned}$$

III

$$\begin{aligned} y_0 &= 2.1 \\ x_1 &= 2.093\ 9 \\ y_1 &= 2.094\ 551\ 72 \\ x_2 &= 2.094\ 551\ 481\ 367\ 28 \end{aligned}$$

IV

$$\begin{aligned} y_0 &= 2.1 \\ x_1 &= 2.094\ 563\ 28 \\ y_1 &= 2.094\ 551\ 481\ 162\ 069\ 454\ 26 \\ x_2 &= 2.094\ 551\ 481\ 542\ 326\ 591\ 9 \end{aligned}$$

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_n obtained by using the simplified formula and by replacing $f'(x_n)$ by $f'(\zeta)$. 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 $f'(x_n)$ not at every step but *only at every second step*. This rule can be interpreted in the following way: from the n th approximation x_n of ζ we obtain the next approximation x_{n+1} by taking

$$y_n = x_n - \frac{f(x_n)}{f'(x_n)}, \quad x_{n+1} = y_n - \frac{f(y_n)}{f'(x_n)}. \quad (\text{G.1})$$

To discuss the rapidity of the convergence of the sequence x_n we assume that x_n (and therefore y_n) tend to ζ and $f'(\zeta) \neq 0$. We have then from (6.9), replacing there x_0 by x_n and x_1 by y_n ,

$$\frac{y_n - \zeta}{(\zeta - x_n)^2} \rightarrow \frac{1}{2} \frac{f''(\zeta)}{f'(\zeta)}. \quad (\text{G.2})$$

On the other hand, since for $y_n \rightarrow \zeta$,

$$f(y_n) = (y_n - \zeta)f'(\zeta) + O[(y_n - \zeta)^2],$$

we have from (G.1) and (G.2)

$$f'(x_n)(x_{n+1} - \zeta) = f'(x_n)(y_n - \zeta) - (y_n - \zeta)f'(\zeta) + O[(y_n - \zeta)^2],$$

$$\begin{aligned} f'(x_n) \frac{x_{n+1} - \zeta}{y_n - \zeta} &= f'(x_n) - f'(\zeta) + O(y_n - \zeta) \\ &= f''(\xi)(x_n - \zeta) + O[(x_n - \zeta)^2], \quad \xi \prec (x_n, \zeta), \end{aligned}$$

$$f'(x_n) \frac{x_{n+1} - \zeta}{(y_n - \zeta)(x_n - \zeta)} \rightarrow f''(\zeta);$$

therefore, since $f'(\zeta) \neq 0$,

$$\frac{x_{n+1} - \zeta}{(y_n - \zeta)(x_n - \zeta)} \rightarrow \frac{f''(\zeta)}{f'(\zeta)}, \quad (\text{G.3})$$

and finally from (G.2)

$$\frac{x_{\nu+1} - \zeta}{(x_{\nu} - \zeta)^3} \rightarrow \frac{1}{2} \left(\frac{f''(\zeta)}{f'(\zeta)} \right)^2. \quad (\text{G.4})$$

3. Now observe that the computation by (G.1) requires *three horners*. Therefore, "at the price" of six horners, we move from x_{ν} to $x_{\nu+2}$, where

$$\frac{x_{\nu+2} - \zeta}{(x_{\nu} - \zeta)^9} \rightarrow \frac{1}{16} \left(\frac{f''(\zeta)}{f'(\zeta)} \right)^8. \quad (\text{G.5})$$

On the other hand, if we use the formula (6.3) three times, we obtain "at the price" of six horners $x_{\nu+3}$, where

$$\frac{(x_{\nu+3} - \zeta)}{(x_{\nu} - \zeta)^8} \rightarrow \left(\frac{f''(\zeta)}{2f'(\zeta)} \right)^7. \quad (\text{G.6})$$

We see that our new rule (G.1) is indeed an improvement of the rule (6.3), although of course the necessity of computing $f'(x_{\nu})$ with more decimals than in the classical Newton-Raphson method is a drawback.

In column III of the table on p. 251, 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 $x_{\nu} - \zeta$ missing in (G.6), as long as $x_{\nu} - \zeta$ 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 $f'(x_{\nu})$ by a convenient combination of $f(x_{\nu})$ and $f(x_{\nu-1})$.

We shall now discuss the following rule, which can be used in this direction.

Starting with x_0 , put

$$y_{\nu} = x_{\nu} - \frac{f(x_{\nu})}{f'(x_{\nu})}, \quad x_{\nu+1} = y_{\nu} - \frac{f(y_{\nu})(y_{\nu} - x_{\nu})}{2f(y_{\nu}) - f(x_{\nu})}. \quad (\text{G.7})$$

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

$$\frac{x_{v+1} - \zeta}{(x_v - \zeta)^4} \rightarrow \frac{1}{24} \frac{f''(\zeta)}{f'(\zeta)^3} [3f''(\zeta)^2 - 2f'(\zeta)f'''(\zeta)], \quad (G.8)$$

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. The 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

$$\frac{-1}{f'_0} \frac{f_0 - f_1}{\left(-\frac{f_0}{f'_0}\right)} = \frac{f_0 - f_1}{f_0}$$

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 the formula (11.30) we have

$$\frac{\zeta - x_{v+1}}{(x_v - \zeta)^2(y_v - \zeta)} \rightarrow \frac{1}{6} f^{(3)}(\zeta) f'(\zeta)^{-1} - \frac{1}{4} f''(\zeta)^2 f'(\zeta)^{-2}. \quad (G.9)$$

But here we have by (6.9) $y_v - \zeta \sim (x_v - \zeta)^2 f''(\zeta)/2f'(\zeta)$ and from (G.9) follows immediately the relation (G.8).

7. In column IV of the table on p. 251, we give the values of the y_0, y_1 and x_1, x_2 , computed by the 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 computing by the 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_{v+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

$$X = (-1)^{n-1} y_1 \dots y_n \sum_{v=1}^n \frac{x_v}{y_v} \frac{1}{F'(y_v)} \quad (\text{H.2})$$

and for $1 \leq k \leq n$

$$D_k = \frac{1}{y_1 \dots y_n} \frac{\partial X}{\partial y_k} \quad (1 \leq k \leq n). \quad (\text{H.3})$$

Then we have

$$D_n \sim \frac{-1}{y_1 \dots y_n}, \quad (\text{H.4})$$

$$D_{n-1} \sim \frac{1}{y_1 \dots y_{n-2} y_{n-1}^2},$$

$$D_{n-2} \sim \frac{-1}{y_1 \dots y_{n-3} y_{n-2}^3} \quad (n \geq 3), \quad (\text{H.5})$$

$$D_k = o\left(\frac{1}{y_1 \dots y_{n-2} y_k^2}\right) \quad (1 \leq k < n-2, \quad n \geq 4). \quad (\text{H.6})$$

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)

$$G(y_k) \sim (-1)^{k-1} y_1 \dots y_{k-1} y_k^{n-k}. \quad (\text{H.8})$$

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 $k = n$

$$\frac{G'(y_n)}{G(y_n)} = - \sum_{\nu=1}^{n-1} \frac{1}{y_\nu} \frac{1}{1 - y_n/y_\nu} \sim \frac{-1}{y_{n-1}};$$

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)

$$\frac{G'(y_k)}{G(y_k)^2} \sim \begin{cases} (-1)^{k-1} \frac{n-k}{y_1 \dots y_{k-1} y_k^{n-k+1}} & (k < n) \\ \frac{(-1)^n}{y_1 y_2 \dots y_{n-2} y_{n-1}^2} & (k = n). \end{cases} \quad (\text{H.9})$$

5. On the other hand, for a $\nu \neq k$ we have

$$G'(y_\nu) = \prod_{\substack{\sigma=1 \\ \sigma \neq k, \nu}}^n (y_\nu - y_\sigma),$$

and therefore

$$G'(y_\mu) \sim (-1)^{\mu-1} y_1 \dots y_{\mu-1} y_\mu^{n-\mu-1} \quad (\mu < k), \quad (\text{H.10})$$

$$G'(y_\lambda) \sim (-1)^\lambda \frac{y_1 \dots y_{\lambda-1} y_\lambda^{n-\lambda}}{y_k} \quad (\lambda > k). \quad (\text{H.11})$$

We have further from (H.2)

$$X = \sum_{\nu=1}^n U_\nu, \quad U_\nu = (-1)^{n-1} y_1 \dots y_n \frac{x_\nu}{y_\nu} \frac{1}{G'(y_\nu)},$$

where in particular, by virtue of (H.7),

$$U_k = (-1)^{n-1} y_1 \dots y_n \frac{x_k}{y_k} \frac{1}{G(y_k)},$$

$$U_\nu = (-1)^{n-1} y_1 \dots y_n \frac{x_\nu}{y_\nu} \frac{1}{G'(y_\nu)(y_\nu - y_k)} \quad (\nu \neq k).$$

We have, therefore, differentiating with respect to y_k ,

$$\frac{\partial U_k}{\partial y_k} = y_1 \dots y_n \frac{x_k}{y_k} \frac{(-1)^n G'(y_k)}{G(y_k)^2}$$

and for $\nu \neq k$

$$\frac{\partial U_\nu}{\partial y_k} = \frac{y_1 \dots y_n}{y_k} \frac{x_\nu}{y_\nu} \frac{(-1)^{n-1}}{G'(y_\nu)} \frac{\partial}{\partial y_k} \frac{y_k}{y_\nu - y_k} = \frac{y_1 \dots y_n}{y_k} \frac{x_\nu}{y_\nu} \frac{(-1)^{n-1}}{G'(y_\nu)} \frac{y_\nu}{(y_\nu - y_k)^2},$$

therefore

$$D_k = \frac{1}{y_1 \dots y_n} \frac{\partial X}{\partial y_k} = \sum_{\nu=1}^n T_\nu, \quad (H.12)$$

$$T_k = \frac{(-1)^n x_k G'(y_k)}{y_k G(y_k)^2}, \quad T_\nu = \frac{(-1)^{n-1} x_\nu}{y_k G'(y_\nu) (y_\nu - y_k)^2} \quad (\nu \neq k).$$

6. But now it follows from (H.9) that

$$T_k \sim \frac{(-1)^{n-k+1} (n-k)}{y_1 \dots y_{k-1} y_k^{n-k+1}} \quad (k < n), \quad (H.13)$$

$$T_k \sim \frac{1}{y_1 y_2 \dots y_{n-2} y_{n-1}^2} \quad (k = n),$$

and from (H.10) and (H.11) for $\mu < k$ and $\lambda > k$:

$$T_\mu \sim \frac{(-1)^{n-1} y_\mu}{y_k (-1)^{\mu-1} y_1 \dots y_{\mu-1} y_\mu^{n-\mu-1} y_\mu^2}$$

$$= \frac{(-1)^{n-\mu}}{y_1 \dots y_{\mu-1} y_\mu^{n-\mu} y_k} \quad (\mu < k), \quad (H.14)$$

$$T_\lambda \sim \frac{(-1)^{n-1} y_\lambda y_k}{y_k (-1)^\lambda y_1 \dots y_{\lambda-1} y_\lambda^{n-\lambda} y_k^2}$$

$$= \frac{(-1)^{n-\lambda-1}}{y_1 \dots y_{\lambda-1} y_\lambda^{n-\lambda-1} y_k^2} \quad (\lambda > k). \quad (H.15)$$

From (H.14) and (H.13) we have for $k < n$

$$\frac{T_\mu}{T_k} \sim \frac{(-1)^{k-\mu-1}}{n-k} \left(\frac{y_\mu}{y_k} \right) \left(\frac{y_{\mu+1}}{y_\mu} \right) \dots \left(\frac{y_{k-1}}{y_\mu} \right) \left(\frac{y_k}{y_\mu} \right)^{n-k},$$

and this tends to 0; therefore,

$$\frac{T_\mu}{T_k} \rightarrow 0 \quad (\mu < k < n). \quad (\text{H.16})$$

Further, we have from (H.13), (H.14) and (H.15), whether $k = n$, or $k = n - 1$, or $k \leq n - 2$,

$$\frac{T_n}{T_{n-1}} \sim -\frac{y_n}{y_{n-1}} \rightarrow 0. \quad (\text{H.17})$$

7. We have on the other hand from (H.14) and (H.13), for $k = n$ and $\mu < n - 1$,

$$\frac{T_\mu}{T_{n-1}} \sim (-1)^{n-\mu-1} \left(\frac{y_\mu}{y_\mu} \right) \left(\frac{y_{\mu+1}}{y_\mu} \right) \dots \left(\frac{y_{n-1}}{y_\mu} \right) \rightarrow 0 \quad (\mu < n - 1 < k).$$

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)

$$D_n \sim T_{n-1} \sim \frac{-1}{y_1 \dots y_n};$$

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 $k = n - 1$

$$D_{n-1} \sim T_{n-1} \sim \frac{1}{y_1 \dots y_{n-2} y_{n-1}^2};$$

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 $k = n - 2$

$$\frac{T_{n-2}}{T_{n-1}} \sim \frac{(-1)^3 2 y_1 \dots y_{n-2} y_{n-2}^2}{y_1 \dots y_{n-3} y_{n-2}^3} = -2,$$

$T_k/T_{n-1} \rightarrow -2$, and it follows now from (H.17) and (H.16) that

$$D_{n-2} \sim -T_{n-1} \sim \frac{-1}{y_1 \dots y_{n-3} y_{n-2}^3},$$

which proves (H.5).

10. We assume now $k < n - 2$. Then we have from (H.13) and (H.15) with $\lambda = n - 1$

$$\frac{T_k}{T_{n-1}} \sim (-1)^{n-k+1} (n-k) \frac{y_k y_{k+1}}{y_k^2} \dots \frac{y_{n-2}}{y_k} \rightarrow 0, \quad (\text{H.18})$$

and from (H.15) for $\lambda < n - 1$

$$\begin{aligned} \frac{T_\lambda}{T_{n-1}} &\sim \frac{(-1)^{n-\lambda+1} y_1 \dots y_{n-2} y_k^2}{y_1 \dots y_{\lambda-1} y_{\lambda}^{n-\lambda-1} y_k^2} \\ &= (-1)^{n-\lambda+1} \frac{y_\lambda y_{\lambda+1}}{y_\lambda^2} \dots \frac{y_{n-2}}{y_\lambda} \quad (\lambda < n - 1). \end{aligned}$$

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

$$T_\nu = o(T_{n-1}) \quad (\nu < n - 2, \quad \nu = n)$$

and

$$T_{n-2} + T_{n-1} = o(T_{n-1}),$$

and by (H.12)

$$D_k = o(T_{n-1}) = o\left(\frac{1}{y_1 \dots y_{n-2} y_k^2}\right) \quad (k < n - 2);$$

our lemma is proved.

11. Now, using the formulas (H.3)–(H.6), we see that in order to make

$$\sum_{\nu=1}^n \delta_\nu \frac{\partial X}{\partial y_\nu} = O(y_1 \dots y_n)$$

it is sufficient to make (a)

$$\delta_n = O(y_1 \dots y_n),$$

$$\delta_{n-1} = O(y_1 \dots y_{n-2} y_{n-1}^2),$$

$$\delta_{n-2} = O(y_1 \dots y_{n-3} y_{n-2}^3),$$

and (b)

$$\delta_k = O(y_1 \dots y_{n-2} y_k^2) \quad (k = 1, \dots, n-3).$$

Here, the results contained in (a) are very satisfactory. Indeed, the y_ν used in these estimates can be considered as already known at the moment the decision about the corresponding δ_ν 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_ν with $\nu > k$.

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 $\varphi'(0) = 1$,

$$\varphi(w) = w + \sum_{\nu=2}^{n-3} a_\nu w^\nu + O(w^{n-2}),$$

and therefore for $w = y_\lambda$

$$x_\lambda = y_\lambda + \sum_{\nu=2}^{n-3} a_\nu y_\lambda^\nu + O(y_\lambda^{n-2}) \quad (1 < \lambda \leq n-3).$$

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 $k < \lambda \leq n$*

$$x_\lambda = y_\lambda + \sum_{\nu=2}^{n-k-2} a_\nu y_\lambda^\nu + O(y_\lambda^{n-k-1}) \quad (k < \lambda \leq n) \quad (\text{H.19})$$

with certain constants a_ν independent of λ . Then we have instead of (H.6)

$$D_k \sim \frac{(-1)^{n-k+1}}{y_1 \dots y_{k-1} y_k^{n-k+1}} \quad (1 \leq k \leq n-3). \quad (\text{H.20})$$

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 ν

$$\frac{y_\lambda^\nu}{\eta(\eta - y_\lambda)^2 G'(\eta)} \sim \frac{(-1)^\lambda y_\lambda^{\nu-n+\lambda}}{\eta^2 y_1 \dots y_k y_{k+1} \dots y_{\lambda-1}}.$$

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

$$\frac{y_\lambda^\nu}{\eta(\eta - y_\lambda)^2 G'(y_\lambda)} = O\left(\frac{y_\lambda^{\nu+1-(n-k)}}{y_1 \dots y_k \eta^2}\right) = O\left(\frac{\eta^{\nu-1-(n-k)}}{y_1 \dots y_k}\right),$$

if we assume $\nu + 1 \geq n - k$. Dividing this by T_k and using the first equivalence (H.13), we obtain

$$\frac{y_\lambda^\nu}{\eta(\eta - y_\lambda)^2 G'(y_\lambda)} = O(\eta^{\nu-1} T_k) \quad (\nu + 1 \geq n - k, \quad k < \lambda \leq n). \quad (\text{H.21})$$

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

$$G_1(y_\lambda) \sim (-1)^{k-1} y_1 \dots y_{k-1}, \quad G_2(\eta) \sim \eta^{n-k}, \quad (\text{H.22})$$

$$\eta \frac{G_2'(\eta)}{G_2(\eta)} = \sum_{\lambda=k+1}^n \frac{\eta}{\eta - y_\lambda} \rightarrow n - k.$$

Introducing the equivalence for $G_1(y_\lambda)$ into (H.21), we obtain further

$$\frac{y_\lambda^\nu}{(\eta - y_\lambda)^2 G_2'(y_\lambda)} = O(y_1 \dots y_k \eta^{\nu-1} T_k) (\nu + 1 \geq n - k, \quad k < \lambda \leq n). \quad (\text{H.23})$$

15. Consider now the expression

$$K_\nu \equiv \sum_{\lambda=k+1}^n \frac{y_\lambda^\nu}{(\eta - y_\lambda)^2 G_2'(y_\lambda)} = -\frac{\partial}{\partial \eta} \frac{1}{G_2(\eta)} \sum_{\lambda=k+1}^n \frac{y_\lambda^\nu G_2(\eta)}{(\eta - y_\lambda) G_2'(y_\lambda)}. \quad (\text{H.24})$$

For $1 \leq \nu < n - k$ the last sum to the right is by Lagrange's interpolation formula identical to η^ν , and we have therefore by (H.22)

$$K_\nu = -\frac{\partial}{\partial \eta} \frac{\eta^\nu}{G_2(\eta)} = \frac{\eta^{\nu-1}}{G_2(\eta)} \left[\eta \frac{G_2'(\eta)}{G_2(\eta)} - \nu \right] \sim \frac{n-k-\nu}{\eta^{(n-k+1)-\nu}},$$

and in particular

$$K_1 \sim (n-k-1)\eta^{k-n}, \quad K_\nu = O(\eta^{\nu-(n-k)-1}) \quad (\nu > 1).$$

Dividing by T_k and using (H.13), we obtain finally

$$\frac{K_1}{T_k} \sim (-1)^{n-k-1} \frac{n-k-1}{n-k} y_1 \dots y_k, \quad (H.25)$$

$$\frac{K_\nu}{T_k} = O(y_1 \dots y_k \eta^{\nu-1}) \quad (\nu > 1). \quad (H.26)$$

The last estimate has been deduced for $\nu < n - k$. It also remains valid, however, for $\nu \geq n - k$ by virtue of (H.23) and (H.24).

16. We now introduce the expressions

$$S_\nu \equiv (-1)^{n-1} \sum_{\lambda=k+1}^n \frac{y_\lambda^\nu}{\eta(\eta-y_\lambda)^2 G'(\eta)} \quad (\nu = 1, \dots, n-k-2). \quad (H.27)$$

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

$$D_k = T_k + S_1 + \sum_{\nu=2}^{n-k-2} a_\nu S_\nu + \sum_{\lambda=k+1}^n O\left(\frac{y_\lambda^{n-k-1}}{\eta G'(y_\lambda)(\eta-y_\lambda)^2}\right) + o(T_k).$$

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 $\nu = n - k - 1$, since $n - k - 1 > 1$. We have therefore

$$D_k = T_k + S_1 + \sum_{\nu=2}^{n-k-2} a_\nu S_\nu + o(T_k). \quad (H.28)$$

17. We consider first the case $k = 1$. In this case we have from (H.27) and (H.24), since $G_2(y) = G(y)$,

$$\frac{S_\nu}{T_k} = (-1)^{n-1} \frac{K_\nu}{\eta T_k},$$

and we have therefore from (H.25) and (H.26) for $k = 1$ the relations

$$S_\nu = o(T_k) \quad (\nu > 1), \quad (\text{H.29})$$

$$S_1 \sim -\left(1 - \frac{1}{n-k}\right) T_k. \quad (\text{H.30})$$

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

$$(-1)^{n-k} y_1 \dots y_k S_\nu = \sum_{\lambda=k+1}^n \prod_{\mu=1}^{k-1} \frac{1}{1 - y_\lambda/y_\mu} \frac{y_\lambda^\nu}{G_2'(y_\lambda)(\eta - y_\lambda)^2}. \quad (\text{H.31})$$

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

$$(-1)^{n-k} y_1 \dots y_k S_\nu = K_\nu + \sum_{\sigma=1}^{n+1} F_\sigma K_{\nu+\sigma}. \quad (\text{H.34})$$

But by virtue of (H.26), each term of the right-hand sum is equal to

$$o(\eta^{-\sigma} y_1 \dots y_k \eta^{\nu+\sigma-1} T_k) = o(y_1 \dots y_k \eta^{\nu-1} T_k).$$

19. It follows now from (H.34) and (H.26) that

$$S_\nu = o(T_k) \quad (\nu > 1),$$

$$S_1 = (-1)^{n-k} \frac{K_1}{y_1 \dots y_k} + o(T_k),$$

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 $\delta_k (1 \leq k \leq n)$ the conditions (c)

$$\delta_k = O(y_1 \dots y_{k-1} y_k^{n-k+1}) \quad (1 \leq k \leq n),$$

which can be considered as completely satisfactory in the same sense as the conditions (a) in Section 11.

Accelerating Iterations with Superlinear Convergence

1. We consider in what follows a sequence z_ν ($\nu = 1, 2, \dots$) convergent to ζ and assume that for a constant $s > 1$ we have

$$\frac{|z_{\nu+1} - \zeta|}{|z_\nu - \zeta|^s} \rightarrow \alpha \quad (\nu \rightarrow \infty, \quad \alpha \neq 0, \quad \alpha \neq \infty). \quad (\text{I.1})$$

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_ν is given by an *iteration formula of the first order*,

$$z_{\nu+1} = \Phi(z_\nu),$$

or more generally if $z_{\nu+1}$ is obtain by an *iteration of the order k*,

$$z_{\nu+1} = \Phi(z_\nu, z_{\nu-1}, \dots, z_{\nu-k+1}).$$

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

$$\Delta_n = \max(|\varepsilon_n|, |\varepsilon_{n-1}|, |z_{n-1} - \zeta|^{s-1}). \quad (\text{I.4})$$

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,

$$s_{n+1} \frac{|z_n - z_{n+1}|^{s+1}}{|z_{n-1} - z_n|^s} = s_{n+1} \alpha^{s+1} \delta^{s^2} [1 + O(\Delta_n)],$$

$$z_{n+1} - \zeta = s_{n+1} \alpha^{s+1} \delta^{s^2} [1 + O(\Delta_n)].$$

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

$$\min(p, s - 1) = d, \quad (\text{I.12})$$

we have obviously $\Delta_n = O(\delta^d)$ and we can replace (I.6) by

$$Z - \zeta = O(\delta^{s+d}). \quad (\text{I.13})$$

In general 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 $\operatorname{sgn}(z_{n+1} - \zeta)$ must be known. In many cases it can be considered as known, namely, when in the case of real z_v 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 also 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 $O(t_2^r) = O(1/\ln \delta_v)$, 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 the approximation (I.2) can be used in the case of an iteration $z_{v+1} = \phi(z_v)$ 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 ,

$$z_{v+k} = \phi(z_v, z_{v+1}, \dots, z_{v+k-1}).$$

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 that 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

$$\begin{aligned} \frac{z_v - \zeta}{z_v - z_{v+1}} - 1 &= \frac{z_{v+1} - \zeta}{z_v - \zeta} \left(1 - \frac{z_{v+1} - \zeta}{z_v - \zeta} \right)^{-1} = O(|z_v - \zeta|^{s-1}): \\ |z_{v+1} - \zeta| &= \alpha |z_v - z_{v+1}|^s (1 + O(|z_v - \zeta|)) (1 + O(|z_v - \zeta|^{s-1})), \\ z_{v+1} - \zeta &= \alpha |z_v - z_{v+1}|^s \text{sgn}(z_{v+1} - \zeta) + O(|z_v - \zeta|^{\min(s+1, 2s-1)}), \\ \zeta &= z_{v+1} - \alpha |z_v - z_{v+1}|^s \text{sgn}(z_{v+1} - \zeta) \\ &\quad + O(|z_v - \zeta|^{\min(s+1, 2s-1)}). \end{aligned} \quad (\text{I.15})$$

We see that the expression

$$Z^* = z_{v+1} - \alpha |z_v - z_{v+1}|^s \text{sgn}(z_{v+1} - \zeta). \quad (\text{I.16})$$

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

$$f(z) \equiv a_0 + a_1 z + \dots = 0 \quad (a_0 = 1), \quad (\text{J.1})$$

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_ν with $\nu > n$ are 0.

2. We shall use the development

$$\Phi(z) \equiv \frac{1}{f(z)} = \sum_{\nu=0}^{\infty} P_\nu z^\nu, \quad (\text{J.2})$$

where $\Phi(z)$ has a radius of convergence $\rho_0 > 0$. The coefficients P_ν 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

$$1 \equiv (a_0 + a_1 z + a_2 z^2 + \dots)(P_0 + P_1 z + P_2 z^2 + \dots),$$

and therefore

$$\left. \begin{array}{l} a_0 P_0 = 1 \\ a_1 P_0 + a_0 P_1 = 0 \\ \vdots \\ a_\nu P_0 + a_{\nu-1} P_1 + \dots + a_0 P_\nu = 0 \end{array} \right\}; \quad (\text{J.3})$$

from these formulas the P_ν are easily obtained one after another.

If in particular $f(z)$ is a polynomial of degree n , the formulas (J.3) with $\nu \geq n$ show that the P_ν 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

$$D_{1,v} = \begin{vmatrix} a_1 & a_2 & \dots & a_v \\ a_0 & a_1 & \dots & a_{v-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{2-v} & a_{3-v} & \dots & a_1 \end{vmatrix} \quad (v = 1, 2, \dots), \quad D_{1,0} = 1, \quad (\text{J.4})$$

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 ν th term of this infinite series can be easily represented (using a Sylvester's determinantal formula) in the form

$$\frac{P_{\nu-1}}{P_\nu} - \frac{P_{\nu-2}}{P_{\nu-1}} = - \frac{D_{2,\nu-1}}{D_{1,\nu-1} D_{1,\nu}}, \quad *$$

if we put

$$D_{2,\nu} = \left| \begin{array}{cccc} a_2 & a_3 & \dots & a_{\nu+1} \\ a_1 & a_2 & \dots & a_\nu \\ \vdots & \vdots & \ddots & \vdots \\ a_{3-\nu} & a_{4-\nu} & \dots & a_2 \end{array} \right| \quad (\nu = 1, 2, \dots), \quad D_{2,0} = 1.$$

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:

$$\xi_1 = - \sum_{\nu=1}^{\infty} \frac{D_{2,\nu-1}}{D_{1,\nu-1} D_{1,\nu}}.$$

* Indeed, let us denote generally by

$$\Delta \begin{pmatrix} \alpha_1, & \alpha_2, \dots \\ \beta_1, & \beta_2, \dots \end{pmatrix}$$

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's formula we have to use is then

$$\Delta \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Delta \begin{pmatrix} \nu \\ \nu \end{pmatrix} - \Delta \begin{pmatrix} \nu \\ 1 \end{pmatrix} \Delta \begin{pmatrix} 1 \\ \nu \end{pmatrix} = \Delta \Delta \begin{pmatrix} 1 & \nu \\ 1 & \nu \end{pmatrix}.$$

If we take now the determinant $D_{1,\nu}$ in (J.4) as Δ , we obtain easily

$$\begin{aligned} \Delta \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= D_{1,\nu-1}, & \Delta \begin{pmatrix} \nu \\ \nu \end{pmatrix} &= D_{1,\nu-1}, & \Delta \begin{pmatrix} \nu \\ 1 \end{pmatrix} &= D_{2,\nu-1}, \\ \Delta \begin{pmatrix} 1 \\ \nu \end{pmatrix} &= a_0^{\nu-1} = 1, & \Delta \begin{pmatrix} 1 & \nu \\ 1 & \nu \end{pmatrix} &= D_{1,\nu-2} \end{aligned}$$

and therefore from Sylvester's formula

$$D_{1,\nu-1}^2 - D_{2,\nu-1} = D_{1,\nu} D_{1,\nu-2};$$

this gives, together with (J.6), the assertion.

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

$$D_{2,1}, \quad D_{2,2}, \quad \dots, \quad D_{2,v-1},$$

$$D_{1,1}, \quad D_{1,2}, \quad \dots, \quad D_{1,v},$$

while by (J.6) the knowledge of $D_{1,v}$ and $D_{1,v-1}$ is completely sufficient.

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 $\xi_1 \xi_2 \dots \xi_k$ 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 the 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

$$\frac{S(z)}{N(z)} = \sum_{\nu=0}^{\infty} y_{\nu} z^{\nu}. \quad (\text{J.18})$$

Multiplying both sides of (J.18) by $N(z)$ and comparing the coefficients of equal powers of z on both sides, we get

$$y_{\nu} A_k + \dots + y_{\nu-k} A_0 = 0 \quad (\nu \geq k). \quad (\text{J.19})$$

Consider now the determinants

$$D_{\nu} = \begin{vmatrix} y_{\nu} & y_{\nu+1} & \dots & y_{\nu+k-1} \\ y_{\nu-1} & y_{\nu} & \dots & y_{\nu+k-2} \\ \vdots & \vdots & & \vdots \\ y_{\nu-k+1} & y_{\nu-k+2} & \dots & y_{\nu} \end{vmatrix}, \quad \nu \geq k-1. \quad (\text{J.20})$$

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

$$-\frac{y_{\nu-k}}{A_k}, \quad -\frac{y_{\nu-k+1}}{A_k}, \dots, \quad -\frac{y_{\nu-1}}{A_k}.$$

If we bring this row by $k-1$ permutations into the k th row, we have, except for the factor

$$\frac{(-1)^k}{A_k} = \frac{1}{\xi_1 \dots \xi_k},$$

the determinant $D_{\nu-1}$. Thus we obtain

$$D_{\nu-1} = \xi_1 \dots \xi_k D_{\nu}. \quad (\text{J.21})$$

8. We are now going to prove that

$$D_{k-1} = \begin{vmatrix} y_{k-1} & y_k & \dots & y_{2k-2} \\ y_{k-2} & y_{k-1} & \dots & y_{2k-3} \\ \vdots & \vdots & & \vdots \\ y_0 & y_1 & \dots & y_{k-1} \end{vmatrix}$$

does not vanish. Indeed, otherwise we would have the n relations

$$\begin{aligned} \alpha_{k-1}y_0 + \alpha_{k-2}y_1 + \dots + \alpha_0y_{k-1} &= 0, \\ \alpha_{k-1}y_1 + \alpha_{k-2}y_2 + \dots + \alpha_0y_k &= 0, \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \\ \alpha_{k-1}y_{k-1} + \alpha_{k-2}y_k + \dots + \alpha_0y_{2k-2} &= 0, \end{aligned} \tag{J.22}$$

where the constants $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ do not all vanish.

Set

$$H(z) = \alpha_{k-1}z^{k-1} + \alpha_{k-2}z^{k-2} + \dots + \alpha_0$$

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

$$P_v = y_v + c_v \quad (v = 0, 1, \dots). \tag{J.23}$$

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) .†

* This result apparently goes back to L. Kronecker (1881).

† For any determinant D we denote the corresponding matrix by the symbol (D) .

We consider the matrix

$$X = \begin{pmatrix} -A_1 & 1 & 0 & \dots & 0 \\ -A_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ -A_{k-1} & 0 & \dots & & 1 \\ -A_k & 0 & \dots & & 0 \end{pmatrix}, \quad (\text{J.24})$$

A_k being the coefficients of $N(z)$ in (J.13), and form the product $(D_\nu)X$. To obtain the first row in this product, we take the first row of (D_ν) :

$$y_\nu, \quad y_{\nu+1}, \dots, \quad y_{\nu+k-1}. \quad (\text{J.25})$$

Multiplying it by the first column of X , we obtain

$$-A_1 y_\nu - A_2 y_{\nu+1} - \dots - A_k y_{\nu+k-1},$$

and this is, by (J.19) and (J.13), equal to $y_{\nu-1}$. The following elements of the first row of $(D_\nu)X$ are obtained by multiplying (J.25) by the 2nd, 3rd, ..., k th columns of X and are easily seen to be $y_\nu, y_{\nu+1}, \dots, y_{\nu+k-2}$. Therefore the complete first row of $(D_\nu)X$ is

$$y_{\nu-1}, \quad y_\nu, \quad y_{\nu+1}, \dots, \quad y_{\nu+k-2}. \quad (\text{J.26})$$

The following rows of $(D_\nu)X$ are obtained from (J.26) by diminishing consecutively the indices of the elements of (J.26) by 1, since this is true in (D_ν) . We see that our product is just the matrix $(D_{\nu-1})$,

$$(D_\nu)X = (D_{\nu-1}),$$

and it follows that

$$\begin{aligned} (D_\nu)X^{\nu-k+1} &= (D_{\nu-1}), & (D_\nu)^{-1} &= X^{\nu-k+1}(D_{\nu-1})^{-1}, \\ (D_\nu)^{-1} &= X^\nu T, & T &= X^{-k+1}(D_{\nu-1})^{-1}. \end{aligned} \quad (\text{J.27})$$

10. Set

$$C_\nu = \begin{pmatrix} c_\nu & c_{\nu+1} & \dots & c_{\nu+k-1} \\ c_{\nu-1} & c_\nu & \dots & c_{\nu+k-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{\nu-k+1} & c_{\nu-k+2} & \dots & c_\nu \end{pmatrix}. \quad (\text{J.28})$$

We obtain from (J.17), (J.20), (J.23) and (J.27)

$$\begin{aligned} (\Delta_\nu) &= (D_\nu) + C_\nu, \\ (\Delta_\nu)(D_\nu)^{-1} &= I + C_\nu X^\nu T. \end{aligned}$$

Taking the determinants, we get

$$\frac{\Delta_\nu}{D_\nu} = |I + C_\nu X^\nu T|. \quad (\text{J.29})$$

We now have to obtain an estimate for X^ν as $\nu \rightarrow \infty$, and in order to do so, we must obtain the fundamental roots of X .

Now we have

$$|zI - X| = \begin{vmatrix} A_1 + z & -1 & 0 & \dots & 0 & 0 & 0 \\ A_2 & z & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{k-2} & 0 & 0 & \dots & z & -1 & 0 \\ A_{k-1} & 0 & 0 & \dots & 0 & z & -1 \\ A_k & 0 & 0 & \dots & 0 & 0 & z \end{vmatrix} = N(z). \quad (\text{J.30})$$

Indeed, putting for $\nu \leq k$,

$$N_\nu(z) = A_0 z^\nu + \dots + A_\nu, \quad N_k(z) = N(z),$$

we have

$$N_k(z) = z N_{k-1}(z) + A_k.$$

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

$$|zI - X| = (-1)^{k-1} A_k (-1)^{k-1} + z N_{k-1}(z) = N(z).$$

We see that the fundamental roots of X are just given by*

$$\xi_1, \xi_2, \dots, \xi_k.$$

We have in particular (Chapter 19, Section 6) $\lambda_X = |\xi_k|$.

* This goes back to S. Günther (1876).

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

$$|X^\nu|_\infty < \sigma(\lambda_X + \varepsilon)^\nu = \sigma(|\xi_k| + \varepsilon)^\nu \quad (\nu = 1, 2, \dots).$$

On the other hand, from (J.16) it follows that for the matrix (J.28)

$$|C_\nu|_\infty = o(q^\nu) \quad (\nu \rightarrow \infty),$$

and therefore by (19.19)

$$\begin{aligned} |C_\nu X^\nu|_\infty &= o[(q|\xi_k| + q\varepsilon)^\nu], \\ |C_\nu X^\nu T|_\infty &= o[(q|\xi_k| + q\varepsilon)^\nu] \quad (\nu \rightarrow \infty). \end{aligned}$$

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

$$\varepsilon = \frac{\delta}{2q},$$

we have

$$q(|\xi_k| + \varepsilon) = \frac{|\xi_k|}{\rho} + \frac{\delta}{2} + \frac{\delta}{2} = \theta,$$

and therefore by (J.29)

$$\frac{A_\nu}{D_\nu} = 1 + o(\theta^\nu) \quad \left(1 > \theta > \frac{|\xi_k|}{\rho}, \quad \nu \rightarrow \infty\right).$$

12. From this we have finally by (J.21), for all θ with $1 > \theta > |\xi_k|/\rho$,

$$\frac{A_{\nu-1}}{A_\nu} \rightarrow \xi_1 \dots \xi_k, \quad \frac{A_{\nu-1}}{A_\nu} = \xi_1 \dots \xi_k + o(\theta^\nu). \quad (\text{J.31})$$

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 $> \max(|\xi_1|, \dots, |\xi_k|)$. 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.

The 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 et appl.* [4] **8**, 101–186, 1892; partly reprinted in J. Hadamard's *Selecta*, Gauthier-Villars, Paris, pp. 19–37, 1935).

13. If we have in the notation of Section 6

$$|\xi_1| < |\xi_2| < |\xi_3| < \dots < |\xi_k|, \quad (\text{J.32})$$

the determination of the ξ_1, \dots, ξ_k can be carried out by a recursive process without computing the determinants Δ_ν . Indeed, multiplying (J.2) by $1 - z/\xi_1$, we obtain

$$\frac{1 - z/\xi_1}{f(z)} = \sum_{\nu=0}^{\infty} Q_\nu z^\nu, \quad Q_0 = 1, \quad Q_\nu = P_\nu - \frac{1}{\xi_1} P_{\nu-1} \quad (\nu > 0). \quad (\text{J.33})$$

From this and from (J.10) we obtain

$$\frac{Q_{\nu-1}}{Q_\nu} \rightarrow \xi_2. \quad (\text{J.34})$$

Multiplying the series in (J.33) again by $1 - z/\xi_2$, we get the series

$$\sum_{\nu=0}^{\infty} R_\nu z^\nu, \quad R_0 = 1, \quad R_\nu = Q_\nu - \frac{1}{\xi_2} Q_{\nu-1} \quad (\nu > 0), \quad (\text{J.35})$$

and from this again

$$\frac{R_{\nu-1}}{R_\nu} \rightarrow \xi_3, \quad (\text{J.36})$$

and so on. If we have computed from the beginning the coefficients P_ν 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	2.467291
E_v	0.4	0.067	0.0084	0.0010	0.00011

Further, we obtain the values of Δ_{v-1}/Δ_v for $k = 2, 3, 4$ 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 $1/9 = \pi^2/4 \cdot 1/(3 \cdot \pi/2)^2$, 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

$$\phi(\lambda) \equiv |A - \lambda I| = 0, \quad \psi(\lambda) \equiv |B - \lambda I| = 0 \quad (\text{K.1})$$

the corresponding characteristic polynomials and equations. Denote the zeros of $\phi(\lambda)$ by λ_ν , and those of $\psi(\lambda)$ by λ'_ν . Put

$$M = \max(|a_{\mu\nu}|, |b_{\mu\nu}|) \quad (\mu, \nu = 1, \dots, n), \quad (\text{K.2})$$

$$\frac{1}{nM} \sum_{\lambda, \nu} |a_{\mu\nu} - b_{\mu\nu}| = \delta. \quad (\text{K.3})$$

Then to every root λ'_ν of $\psi(\lambda)$ belongs a certain root λ_ν of $\phi(\lambda)$ such that we have

$$|\lambda'_\nu - \lambda_\nu| \leq (n + 2)M\delta^{1/n}. \quad (\text{K.4})$$

Further, for a suitable ordering of λ_ν and λ'_ν we have

$$|\lambda_\nu - \lambda'_\nu| \leq 2(n + 1)^2 M \delta^{1/n} \quad (\nu = 1, \dots, n). \quad (\text{K.5})$$

2. We prove first the

Lemma. Under the hypotheses of the theorem, we have for any λ with $|\lambda| \leq nM$

$$|\phi(\lambda) - \psi(\lambda)| \leq (n + 2)^n M^n \delta. \quad (\text{K.6})$$

Proof of the Lemma. Denote the $a_{\mu\nu}$ in an arbitrarily chosen order by $\alpha_1, \dots, \alpha_n$, and the $b_{\mu\nu}$ in the corresponding order by β_1, \dots, β_n . We can then write

$$\phi(\lambda) = P(\alpha_1, \dots, \alpha_n), \quad \psi(\lambda) = P(\beta_1, \dots, \beta_n),$$

$$\phi(\lambda) - \psi(\lambda) = \sum_{k=1}^n A_k \quad (\text{K.7})$$

with

$$\Delta_\kappa = P(\alpha_1, \dots, \alpha_\kappa, \beta_{\kappa+1}, \dots, \beta_n) - P(\alpha_1, \dots, \alpha_{\kappa-1}, \beta_\kappa, \dots, \beta_n).$$

Since $P(\alpha_1, \dots, \alpha_n)$ 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)

$$|\Phi(\lambda) - \Psi(\lambda)| \leq (n+2)^{n-1}M^{n-1} \sum_{\kappa=1}^{n^2} |\alpha_\kappa - \beta_\kappa| \leq (n+2)^n M^n \delta,$$

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)

$$|\Phi(\lambda')| = |\Phi(\lambda') - \Psi(\lambda')| \leq (n+2)^n M^n \delta,$$

$$\prod_{v=1}^n |\lambda' - \lambda_v| \leq [(n+2)M \delta^{1/n}]^n; \quad (\text{K.9})$$

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 6 and 7 of Appendix A can be applied without any change. We see that by ordering the λ_v and λ'_v conveniently, we have

$$|\lambda'_v - \lambda_v| \leq 2n(n+2)M \delta^{1/n} \leq 2(n+1)^2 M \delta^{1/n} \quad (v = 1, \dots, n);$$

(K.5) is proved.

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 only. To obtain the explicit expressions of the U_v put, *assuming that all x_v are distinct*, $f(x_m) = a$, $f(x_v) = 0$ ($v = 1, \dots, m-1$). Then, referring to the definition of divided differences we see that we obtain in this case (L.1) dividing a successively by

$$x_m - x_1, x_m - x_2, \dots, x_m - x_{m-1};$$

it follows that

$$U_m = \frac{1}{(x_m - x_1)(x_m - x_2) \dots (x_m - x_{m-1})}.$$

Since (L.1) is symmetric in x_1, \dots, x_m , we have

$$[x_1, \dots, x_m]f = \sum_{v=1}^m \frac{f(x_v)}{(x_v - x_1) \dots (x_v - x_{v-1})(x_v - x_{v+1}) \dots (x_v - x_m)}. \quad (\text{L.2})$$

2. To write this 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 will prove 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,

$$\frac{\prod_{m-1 \geq \mu > \nu \geq 1} (x_\mu - x_\nu)}{\prod_{m \geq \mu > \nu \geq 1} (x_\mu - x_\nu)} = \frac{1}{\prod_{\nu=1}^{m-1} (x_m - x_\nu)},$$

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

$$[x_1, \dots, x_{m_1}; y_1, \dots, y_{m_2}; \dots; z_1, \dots, z_{m_k}]f$$

$$= \frac{|Z(x_1)Z(x_2)\dots Z(x_{m_1})Z(y_1)\dots Z(z_{m_k})|}{|N(x_1)N(x_2)\dots N(x_{m_1})N(y_1)\dots N(z_{m_k})|}. \quad (\text{L.5})$$

Subtracting both in the numerator and in 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_ν column ($1 < \nu \leq m_1$) we can replace each x_ν column by

$$[x_\nu, x_1]Z(x_1), \quad [x_\nu, x_1]N(x_1).$$

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

$$Z(x_1), \quad [x_2, x_1]Z(x_1), \dots, \quad [x_{m_1}, \dots, x_2, x_1]Z(x_1);$$

$$N(x_1), \quad [x_2, x_1]N(x_1), \dots, \quad [x_{m_1}, \dots, x_2, x_1]N(x_1).$$

The same procedure can be applied to the columns depending on the y_v, \dots, z_v and we obtain finally

$$[x_1, \dots, x_{m_1}; y_1, \dots, y_{m_2}; \dots; z_1, \dots, z_{m_k}]f \quad (\text{L.6})$$

$$= \frac{|Z(x_1)[x_1, x_2]Z(x_1) \dots [x_1, \dots, x_{m_1}]Z(x_1)Z(y_1) \dots [z_1, \dots, z_{m_k}]Z(z_1)|}{|N(x_1)[x_1, x_2]N(x_1) \dots [x_1, \dots, x_{m_1}]N(x_1) \dots [z_1, \dots, z_{m_k}]N(z_1)|}.$$

4. Assume now k distinct variables t_1, \dots, t_k and let in (L.6) 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

$$[t_1^{m_1}, \dots, t_k^{m_k}]f = \frac{Z^*}{N^*},$$

$$Z^* = |Z(t_1)Z'(t_1) \dots Z^{(m_1-1)}(t_1)Z(t_2) \dots Z^{(m_k-1)}(t_k)|,$$

$$N^* = |N(t_1)N'(t_1) \dots N^{(m_1-1)}(t_1) \dots N^{(m_k-1)}(t_k)|, \quad (\text{L.7})$$

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

$$\Delta(x) = \begin{vmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_{m_1} \\ \vdots & & \vdots \\ x_1^{m_1-1} & \dots & x_{m_1}^{m_1-1} \end{vmatrix} = \prod_{\mu > v} (x_\mu - x_v)$$

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

$$N_0 = \frac{N}{\Delta(x) \Delta(y) \dots \Delta(z)}. \quad (\text{L.9})$$

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

$$N_0 = R(x, y) \dots R(x, z) \dots R(y, z) \dots$$

and this tends indeed to the right-side product in (L.8).

M

Remainder Terms in Interpolation Formulas

1. The formulas (1.23a) (and (1.23c)) 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 (1.23a) only the *existence* of the derivative $f^{(n)}$ throughout the corresponding interval and in the case of (1.23c) 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 (1.15) has been proved in Chapter 1, Section 10 only under the assumption of the *continuity* of $f^{(m)}(x)$. This raises the question whether (1.15) still exists if only the *existence* of $f^{(m)}(x)$ is assumed.

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 $F(x_v) = 0$ ($v = 1, \dots, n + 1$), $x_v \prec J_x$, and that $F^{(n)}(x)$ exists in J_x . Then there exists a $\xi \prec J_x$ 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

$$F(x) = x^{n+1}, \quad J_x: 0 \leq x \leq 1;$$

then $\xi = 0$ and does not lie in (J_x) .

3. Proof. Let $g(x) \equiv F'(x)$. 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 ν , then $g(x)$ will have x_0 as a zero of multiplicity $\nu - 1$. 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 $\xi \prec (J_x)$ 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 $x_v \prec J_x$ ($v = 1, \dots, n$) 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 $\xi \prec (J_x)$ such that

$$\frac{f(x_0)}{g(x_0)} = \frac{f^{(n)}(\xi)}{g^{(n)}(\xi)}, \quad \xi \prec (J_x) \quad (x_0 \neq x_v; \quad v = 1, \dots, n). \quad (\text{M.1})$$

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 $\xi \prec J_x$ 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 discuss the existence of (1.15) under the assumption that the derivatives of $f(t)$ up to the order $\max_k(m_k - 1)$ are continuous in $\langle t_1, \dots, t_k \rangle$. We can use the formula (L.6) where the value of the denominator is given by (L.8), and all x_v, y_v, \dots, z_v are chosen from $\langle t_1, \dots, t_k \rangle$. Then in the numerator in (L.6) the columns

$$Z(x_1), [x_1, x_2]Z(x_1), \dots, [x_1, x_2, \dots, x_m]Z(x_1)$$

have the corresponding expressions $Z(t_1), Z'(t_1), \dots, 1/(m-1)!Z^{(m-1)}(t_1)$ as their limits, since only the derivatives of $f(x_1)$ up to the order $m_1 - 1$ are implied and these are assumed to be continuous. Applying this argument to all sets of variables x_v, y_v, \dots, z_v we obtain then indeed the representation (L.7).

Now, putting $m = m_1 + \dots + m_k - 1$ we see that if the m th derivative of f exists in $\langle t_1, \dots, t_{m_k} \rangle$ then (1.15) certainly exists for $k > 1$.

8. The situation is different if the derivatives of the order $m_k - 1$ are assumed to exist but are not necessarily continuous. Already for $m = 1$, $[x_2, x_1]f$ 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 , unless t lies between x_1, x_2 . Consider, for instance, the function

$$f(x) = x^2 \sin \frac{1}{x} \quad (x \neq 0), \quad f(0) = 0,$$

which has for $x \neq 0$ the derivative $2x \sin(1/x) - \cos(1/x)$ and at $x = 0$ the derivative 0. Putting

$$\frac{1}{u_\nu} = \frac{\pi}{4} + 2\nu\pi \quad (\nu = 0, 1, \dots),$$

we have

$$f'(u_\nu) = \frac{\sqrt{2}}{\pi/4 + \nu\pi} - \sqrt{1/2},$$

and we see that the values of $f'(u_\nu)$ tend to $-\sqrt{\frac{1}{2}}$. Choose now $x_1^{(\nu)} \neq x_2^{(\nu)}$ and both so close to u_ν that $[x_2^{(\nu)}, x_1^{(\nu)}]f(x) = -\sqrt{\frac{1}{2}} + O(1/\nu)$. Then we have indeed

$$x_1^{(\nu)} \rightarrow 0, \quad x_2^{(\nu)} \rightarrow 0, \quad [x_2^{(\nu)}, x_1^{(\nu)}]f(x) \rightarrow -\sqrt{\frac{1}{2}} \neq f'(0).$$

9. We return now to the general interpolation formula (1.20)–(1.22) 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 the formulas (1.20), (1.21) 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 (1.24) for the remainder exists and remains valid. Indeed, in this case we have $\max m_\kappa - 1 \leq n - 1$.

10. We write now $L_{n-1}(x) \equiv L(x)$ 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

$$g(x) = \prod_{\nu=1}^n (x - x_\nu) = \prod_{\kappa=1}^k (x - t_k)^{m_\kappa}, \quad g^{(n)} = n!,$$

(M.3) becomes

$$f(x) = L_{n-1}(x) + \frac{f^{(n)}(\xi)}{n!} \prod_{\nu=1}^n (x - x_\nu), \quad \xi \prec \langle x, t_1, \dots, t_k \rangle. \quad (\text{M.4})$$

Comparing this with (1.24), we see that we have

$$[x, t_1^{m_1}, \dots, t_k^{m_k}] f = \frac{1}{n!} f^{(n)}(\xi), \quad \xi \prec \langle x, t_1, \dots, t_k \rangle \quad (\text{M.5})$$

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 $J(\zeta)$:

$$f(\zeta) = f'(\zeta) = \dots = f^{(n)}(\zeta) = 0 \quad (\text{N.3})$$

so that

$$f(x) = (x - \zeta)^n \phi(x), \quad (\text{N.4})$$

where $\varphi(x)$ is continuous in $J(\zeta)$ with $\varphi(\zeta) = 0$.* Then we have for $k = 0, 1, \dots, n - 1$:

$$(x - \zeta)^k \varphi^{(k)}(x) \rightarrow 0 \quad (x \rightarrow \zeta, \quad x \prec J(\zeta), \quad k = 0, 1, \dots, n - 1). \quad (\text{N.5})$$

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 $p \leq n$,

$$F(x) = \frac{f(x)}{(x - \zeta)^p}. \quad (\text{N.6})$$

* 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

$$\lim \varphi(x) = \lim \frac{f(x)}{(x - \zeta)^n} = \frac{1}{n!} \lim \frac{f^{(n-1)}(x)}{x - \zeta} = \frac{1}{n!} f^{(n)}(\zeta) = 0.$$

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) $F(x) = (x - \zeta)^{n-p}\Phi(x)$. Differentiating this k times we get

$$\begin{aligned} F^{(k)}(x) &= \sum_{\kappa=0}^k \binom{k}{\kappa} \Phi^{(\kappa)}(x) (k - \kappa)! \binom{n-p}{k-\kappa} (x - \zeta)^{n-p-k+\kappa}, \\ \frac{F^{(k)}(x)}{(x - \zeta)^{n-p-k}} &= \sum_{\kappa=0}^k \binom{k}{\kappa} (k - \kappa)! \binom{n-p}{k-\kappa} [(x - \zeta)^\kappa \Phi^{(\kappa)}(x)], \end{aligned}$$

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 $k \leq n - p$

$$F^{(k)}(x) \rightarrow 0 \quad (x \rightarrow \zeta; \quad x \prec J(\zeta); \quad k = 0, 1, \dots, n - p). \quad (\text{N.8})$$

In order to derive from this result some information about $F^{(k)}(\zeta)$, we need a further

Lemma 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 :

$$f^{(n)}(x) \rightarrow \alpha_n \quad (x \rightarrow \zeta, \quad x \prec J). \quad (\text{N.9})$$

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 J .

5. Proof. Define in J the function $\psi(x)$ as having the values $f^{(n)}(x)$ in J and α_n in ζ . The function $\psi(x)$ is continuous in 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

$$\lim_{x \rightarrow \zeta} f^{(n-1)}(x) = f^{(n-1)}(\zeta) + \int_b^\zeta \psi(u) du \equiv \alpha_{n-1} \quad (x \prec J).$$

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

$$f^{(v)}(x) \rightarrow \alpha_v \quad (x \rightarrow \zeta, \quad x \prec J, \quad v = 0, 1, \dots, n).$$

We define now $f(x)$ in ζ as α_0 . We have then for $x \prec J$ 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 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

$$g^{(\lambda)}(x) = \sum_{v=0}^{\lambda} \binom{\lambda}{v} (\lambda - v)! \binom{p}{\lambda - v} (x - \zeta)^{p - \lambda + v} G^{(v)}(x),$$

$$g^{(\lambda)}(x) = \sum_{\kappa=0}^{n-p} \binom{\lambda}{\kappa} (\lambda - \kappa)! \binom{p}{\lambda - \kappa} (x - \zeta)^{p-\lambda+\kappa} G^{(\kappa)}(x)$$

$$+ (x - \zeta)^{n-\lambda} \sum_{\kappa=n-p+1}^{\lambda} \binom{\lambda}{\kappa} (\lambda - \kappa)! \binom{p}{\lambda - \kappa} (x - \zeta)^{p-n+\kappa} G^{(\kappa)}(x).$$

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 $\binom{p}{\lambda - \kappa} = 0$ for $\lambda - \kappa > p$, $\kappa < \lambda$.

We obtain

$$g^{(\lambda)}(x) = \sum_{\nu=\max(0, \lambda-p)}^{\min(\lambda, n-p)} C_{\lambda\nu} (x - \zeta)^{\nu-\lambda+p} G^{(\nu)}(x)$$

$$+ o((x - \zeta)^{n-\lambda}) \quad (x \prec J, x \rightarrow \zeta, 0 < \lambda < n) \quad (\text{N.17})$$

For $\lambda < p$ it follows now that $g^{(\lambda)}(x) \rightarrow 0$. For $\lambda \geq p$ we can write, taking the term with $\nu = \lambda - p$ out,

$$g^{(\lambda)}(x) = \sum_{\nu=\lambda-p+1}^{\min(\lambda, n-p)} C_{\lambda\nu} (x - \zeta)^{\nu-\lambda+p} G^{(\nu)}(x)$$

$$+ o((x - \zeta)^{n-\lambda})$$

$$+ p! \binom{\lambda}{\lambda-p} G^{(\lambda-p)}(x),$$

where $C_{\lambda\nu}$ are convenient numerical constants and $x \prec J$.

In this formula we let x go to ζ . Again, in the first right-hand sum the derivatives $G^{(\nu)}(x)$ 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 existance 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 $p = n$,

$$\frac{g^{(n-1)}(x) - g^{(n-1)}(\zeta)}{x - \zeta} = n! G(\zeta) + o(1) \tag{N.19}$$

But here, for $x \rightarrow \xi$, 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 J , a limit. This proves the existance 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 $J(\zeta)$, assume that $z(\omega)$ is continuous for $\omega \prec U$ together with $z', \dots, z^{(n-1)}$. Put

$$G^*(x) = z(G(x)). \quad (\text{N.20})$$

Then $G^*(x)$ satisfies also 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

$$G^{*(m)}(x) = \sum c_{\alpha_1, \dots, \alpha_m, \beta} G'^{\alpha_1}(x) \dots G^{(m)\alpha_m}(x) z^{(\beta)}(G), \quad (\text{N.21})$$

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 $m \leq n-p$, $G^{*(m)}(x)$ has obviously 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 ,

$$z(\omega) = (1 + \omega)^{1/p} = \sum_{v=0}^{\infty} \binom{1/p}{v} \omega^v,$$

* 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 .

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.

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)

$$k = -\frac{u}{u'} = -p \frac{y}{y'},$$

$$X_1 = 1, \quad \frac{1}{2!} X_2 = -\frac{1}{2} \frac{u''}{u'} = \frac{(\rho - 1)y'^2 - \rho yy''}{2\rho yy'},$$

$$\begin{aligned} \frac{1}{3!} X_3 &= \frac{3u''^2 - u'u'''}{6u'^2} \\ &= \frac{(\rho - 2)(\rho - 1)y'^4 - 3\rho(\rho - 1)yy'^2y'' - \rho^2y^2y'y''' + 3\rho^2y^2y''^2}{6\rho^2y^2y'^2}. \end{aligned}$$

Instead of (2.20) we obtain in our case

$$\begin{aligned} \zeta - x &= k + \frac{(\rho - 1)y'^2 - \rho yy''}{2\rho yy'} k^2 & (N.30) \\ &+ \frac{(\rho - 2)(\rho - 1)y'^4 - 3\rho(\rho - 1)yy'^2y'' - \rho^2y^2y'y''' + 3\rho^2y^2y''^2}{6\rho^2y^2y'^2} k^3 \\ &+ \frac{1}{4!} X_4 k^4 + O((x - \zeta)^5). \end{aligned}$$

17. For certain purposes we need the values of $X_p(u'/u', u''/u', \dots, u^{(p)}/u')$ 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_p , for $x = \zeta$ 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 $x - \zeta = \xi$,

$$f(x) = \xi^\rho \alpha(1 + \beta\xi + \gamma\xi^2 + \delta\xi^3 + O(\xi^4)),$$

where

$$\begin{aligned} \alpha &= \frac{f^{(\rho)}(\zeta)}{\rho!}, & \beta &= \frac{f^{(\rho+1)}(\zeta)}{f^{(\rho)}(\zeta)} \frac{1}{\rho+1}, \\ \gamma &= \frac{f^{(\rho+2)}(\zeta)}{(\rho+1)(\rho+2)f^{(\rho)}(\zeta)}, & \delta &= \frac{f^{(\rho+3)}(\zeta)}{(\rho+1)(\rho+2)(\rho+3)f^{(\rho)}(\zeta)}. \end{aligned} \quad (N.31)$$

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

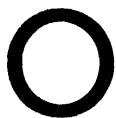
$$u' = \alpha^{1/p}, \quad \frac{u''}{u'} = \frac{2\beta}{p} = \frac{2}{p(p+1)} \frac{f^{(p+1)}(\zeta)}{f^{(p)}(\zeta)}, \quad (\text{N.32})$$

$$\begin{aligned}
\frac{u'''}{u'} &= 3 \frac{2p\gamma - (p-1)\beta^2}{p^2} \\
&= 3 \frac{2p(p+1)f^{(p)}(\zeta)f^{(p+2)}(\zeta) - (p-1)(p+2)(f^{(p+1)}(\zeta))^2}{(p+1)^2(p+2)p^2f^{(p)}(\zeta)^2}, \quad (\text{N.33})
\end{aligned}$$

$$\begin{aligned}
\frac{u^{(4)}}{u'} &= 4 \frac{(p-1)(2p-1)\beta^3 - 6p(p-1)\beta\gamma + 6p^2\delta}{p^3} \\
&= 4 \left(\frac{(p-1)(2p-1)(p+2)(p+3)f^{(p+1)}(\zeta)^3}{p^3(p+3)(p+1)^3(p+2)f^{(p)}(\zeta)^3} \right. \\
&\quad - \frac{6p(p^2-1)(p+3)f^{(p)}(\zeta)f^{(p+1)}(\zeta)f^{(p+2)}(\zeta)}{p^3(p+3)(p+1)^3(p+2)f^{(p)}(\zeta)^3} \\
&\quad \left. + \frac{6p^2(p+1)^2f^{(p)}(\zeta)^2f^{(p+3)}(\zeta)}{p^3(p+3)(p+1)^3(p+2)f^{(p)}(\zeta)^3} \right). \quad (\text{N.34})
\end{aligned}$$

For the corresponding values of X_1, \dots, X_4 we have now

$$\begin{aligned}
X_1 &= 1, \quad X_2 = -\frac{2}{p(p+1)} \frac{f^{(p+1)}(\zeta)}{f^{(p)}(\zeta)}, \\
\frac{1}{3}p^2(p+1)^2(p+2)f^{(p)}(\zeta)^2X_3 &= (p+2)(p+3)f^{(p+1)}(\zeta)^2 \\
&\quad - 2p(p+1)f^{(p)}(\zeta)f^{(p+2)}(\zeta), \\
\frac{1}{8}p^3(p+1)^3(p+2)(p+3)f^{(p)}(\zeta)^3X_4 &= 3p(p+1)(p+3)(p+4)f^{(p)}(\zeta)f^{(p+1)}(\zeta)f^{(p+2)}(\zeta) \\
&\quad - (p+2)^2(p+3)(p+4)f^{(p+1)}(\zeta)^3 - 3p^2(p+1)^2f^{(p)}(\zeta)^2f^{(p+3)}(\zeta).
\end{aligned}$$



Laguerre Iterations

1. Consider n real quantities a_1, \dots, a_n and form with them

$$a_1 + \dots + a_n = a, \quad a_1^2 + \dots + a_n^2 = b. \quad (O.1)$$

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

$$\Phi_N(u) \equiv Nu^2 - 2au + a^2 - (N-1)b = 0 \quad (O.2)$$

has two real roots

$$u_{1,2} = \frac{1}{N}(a \pm \sqrt{(N-1)(Nb - a^2)}) \quad (O.3)$$

and that, if $u_1 \leq u_2$, we have

$$u_1 \leq a_1 \leq u_2. \quad (O.4)$$

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 $u = u_2$,

$$\Phi_N(u) = N(u^2 - b) - (2au - a^2 - b) = 0. \quad (O.5)$$

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 $u^2 - b \leq 0$.

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

$$N^2(u^2 - b) = 2aNu - (\alpha + \beta) - 2A^2. \quad (\text{O.7})$$

If 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 $\Phi_{\bar{N}}(u) = 0$. Then we have

$$\bar{N}(\bar{u}^2 - b) = 2a\bar{u} - a^2 - b$$

and it follows therefore for $\Phi_N(\bar{u})$ that

$$\Phi_N(\bar{u}) = N(\bar{u}^2 - b) - (2a\bar{u} - a^2 - b) = (N - \bar{N})(\bar{u}^2 - b) \geq 0,$$

$$\Phi_{\bar{N}}(u) = (\bar{N} - N)(u^2 - b) \leq 0.$$

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 ,

$$u_1 \leqq \frac{1}{x - \zeta_v} \leqq u_2 \quad (\text{O.8})$$

where u_1 is the smaller and u_2 the greater of the roots given by

$$u_{1,2} = \frac{1}{Nf(x)} [f'(x) \pm \sqrt{(N-1)((N-1)f'(x)^2 - Nf(x)f''(x))}],$$

$$u_1(x) \leqq u_2(x). \quad (\text{O.9})$$

Since $b > 0$, none of the u_1 , u_2 can vanish.

5. We put in what follows

$$v_1(x) = \frac{1}{Nf(x)} [f'(x) - \sqrt{(N-1)((N-1)f'(x)^2 - Nf(x)f''(x))}],$$

$$v_2(x) = \frac{1}{Nf(x)} [f'(x) + \sqrt{(N-1)((N-1)f'(x)^2 - Nf(x)f''(x))}]. \quad (\text{O.10})$$

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 $(\zeta_\nu, \zeta_{\nu+1})$ 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:

$$y_0' = x, \quad y_{\kappa+1}' = y_\kappa' - \frac{1}{u_2(y_\kappa')} \quad (\kappa = 0, 1, \dots), \quad (O.14)$$

$$z_0' = x, \quad z_{\kappa+1}' = z_\kappa' - \frac{1}{u_1(z_\kappa')} \quad (\kappa = 0, 1, \dots). \quad (O.15)$$

Then it follows from our discussion that the y_κ' monotonically decrease and remain $> \zeta_\nu$, while the z_κ' monotonically increase and remain $< \zeta_{\nu+1}$. Both sequences are therefore convergent and, since by (O.9) in the limit points $f(x)$ must vanish, we see that

$$y_\kappa' \downarrow \zeta_\nu, \quad z_\kappa' \uparrow \zeta_{\nu+1}.$$

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 $(\zeta_\nu, \zeta_{\nu+1})$, 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 $\max(\zeta, \eta)$ and the z_κ tend to $\min(\zeta, \eta)$, 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 $\nu = n$,

$$x - \frac{1}{u_2(x)} > \zeta_n.$$

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 $u_1(x)$ is > 0 , then it follows from (O.8), applied to $\nu = 1$, 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 $u_1(x)$ is < 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

$$u_1(x)u_2(x) = \frac{(N-1)f(x)f''(x) - (N-2)f'(x)^2}{Nf(x)^2} \quad (O.17)$$

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 ζ 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 , or z , by x_ν . We have then obviously, as follows from (O.14) and (O.15), $|u(x_\nu)| \rightarrow \infty$ and even, since the convergence is monotonic, $u(x_\nu) \rightarrow \pm \infty$.

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_ν to ζ it follows also by (O.16) that, from a ν onward, the $u(x_\nu)$ have the same sign as $x_\nu - \zeta$, so that we have finally from a certain ν onward

$$\operatorname{sgn}(x_\nu - \zeta) = \operatorname{sgn} \frac{f(x_\nu)}{f'(x_\nu)} = \operatorname{sgn} u(x_\nu).$$

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 $x = x_\nu \rightarrow \zeta$ 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 $\sigma(x_\nu) \rightarrow 0$, we can develop, from a ν onward, writing x for x_ν until formula (O.25),

$$\begin{aligned} \sqrt{1 - \frac{N}{N - 1} \sigma(x)} &= 1 - \frac{1}{2} \frac{N}{N - 1} \sigma(x) - \frac{1}{8} \left(\frac{N}{N - 1} \right)^2 \sigma(x)^2 + O(\sigma(x)^3), \\ u(x) &= \frac{f'(x)}{f(x)} \left[1 - \frac{1}{2} \sigma(x) - \frac{1}{8} \frac{N}{N - 1} \sigma(x)^2 + O(\sigma(x)^3) \right], \\ \frac{1}{u(x)} &= \frac{f(x)}{f'(x)} \left[1 + \frac{1}{2} \sigma(x) \right. \\ &\quad \left. + \frac{1}{8} \frac{N}{N - 1} \sigma(x)^2 + \frac{1}{4} \sigma(x)^2 + O(\sigma(x)^3) \right], \\ \frac{1}{u(x)} &= \frac{f(x)}{f'(x)} \left[1 + \frac{\sigma(x)}{2} \right. \\ &\quad \left. + \frac{3}{8} \left(1 + \frac{1}{3(N - 1)} \right) \sigma(x)^2 + O(\sigma(x)^3) \right]. \end{aligned} \quad (\text{O.21})$$

14. We use now the Schröder series (2.20), replacing there 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

$$x - \zeta = \frac{f(x)}{f'(x)} + \frac{1}{2} \frac{f''(x)f(x)^2}{f'(x)^3} + \frac{3f''(x)^2 - f'(x)f'''(x)}{6f'(x)^5} f(x)^3 + O(f(x)^4).$$

Using again the notation (O.19) and observing that $f(x) = O(x - \zeta)$ we see that this becomes

$$x - \zeta = \frac{f(x)}{f'(x)} \left[1 + \frac{1}{2}\sigma(x) + \frac{3f''(x)^2 - f'(x)f'''(x)}{6f'(x)^2} \cdot \left(\frac{f(x)}{f'(x)} \right)^2 \right] + O((x - \zeta)^4). \quad (\text{O.22})$$

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

$$f_0' \equiv f'(\zeta) \neq 0, \quad f_0'' \equiv f''(\zeta), \quad f_0''' \equiv f'''(\zeta), \quad (\text{O.23})$$

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 here x by x_ν , and observing that the left-side numerator becomes $x_{\nu+1} - \zeta$, we have

$$\frac{x_{\nu+1} - \zeta}{(x_\nu - \zeta)^3} \rightarrow \frac{1}{24f_0'^2} \left[3f_0''^2 - 4f_0'f_0''' - \frac{3}{N-1} f_0''^2 \right]. \quad (\text{O.25})$$

15. In the 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

$$g(y) = a + \binom{n-1}{1}by + \binom{n-1}{2}cy^2 + \dots,$$

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 the 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

$$\frac{f(x)}{f'(x)} \sim \frac{1}{p}(x - \zeta), \quad \frac{f(x)f''(x)}{f'(x)^2} \rightarrow 1 - \frac{1}{p}. \quad (O.28)$$

We have now from (O.18)

$$N(x - \zeta)u(x) \rightarrow p[1 \pm \sqrt{(N-1)(N-1-N(1-1/p))}],$$

$$\frac{1}{u(x)} / (x - \zeta) \rightarrow \frac{N}{p \pm w}, \quad w = \sqrt{(N-1)(N-p)p}. \quad (O.29)$$

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

$$\frac{N}{p \pm w} \leq 0, \quad p \pm w > 0.$$

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,

$$p(N-1)(N-p) \leq p^2 - pN + \frac{N^2}{4}.$$

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's 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 the Chapter 18. We use the notations introduced in the 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

$$1 > \frac{\delta}{2^{n+1}T} > \eta_1 \geq \dots \geq \eta_{n-1}, \quad \eta_n < \delta^{p+1}\eta_{n-1}, \quad \eta < \frac{\delta}{2^{n+1}T} \quad (\text{P.2})$$

$$\eta^{p-1} > \frac{T}{4\delta^{p+1}} \prod_{\kappa=1}^{n-1} \eta_\kappa, \quad (\text{P.3})$$

$$\eta^{p-1} \leq \frac{4T}{\delta} \prod_{\kappa=1}^{n-1} (\eta + \eta_\kappa). \quad (\text{P.4})$$

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 $\eta + \eta_\kappa$ either by $2\eta_\kappa$ or by 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 from (P.9) would follow

$$\frac{\delta}{2^{n+1}T} \leq \eta^{n-p} \leq \eta,$$

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,

$$|z_\nu - \zeta| \equiv \xi_\nu. \quad (\text{P.12})$$

Lemma 2. Assume under the hypotheses of Theorem 18.1 that $p > 1$. Put $|\gamma| = T$ and assume δ satisfies (P.1), and N is such that we have

$$\xi_\nu < \frac{\delta}{2^{n+1}T}, \quad \xi_\nu \max(C_1, C_2) \leq \frac{1}{3} \quad (\nu \geq N). \quad (\text{P.13})$$

Then, if we have for a $\nu \geq N$,

$$\xi_{\nu+n+1} < \delta^{p+1} \min(\xi_{\nu+1}, \dots, \xi_{\nu+n}), \quad \nu \geq N, \quad (\text{P.14})$$

we have also

$$\xi_{\nu+n+2} < \delta \xi_{\nu+n+1}. \quad (\text{P.15})$$

8. Proof. Without loss of generality we can and will assume $\zeta = 0$. Assume that (P.15) is false so that we have

$$\xi_{\nu+n+2} \geq \delta \xi_{\nu+n+1}. \quad (\text{P.16})$$

From (17.22), we have, replacing ε_0 by $m \equiv \max(\xi_{\nu+1}, \dots, \xi_{\nu+n})$ and since $|\gamma| = T$,

$$\begin{aligned} \xi_{\nu+n+1}^{\phi} &= U_{\nu} T \prod_{\kappa=1}^n |z_{\nu+n+1} - z_{\nu+\kappa}|, \\ U_{\nu} &= \frac{1 + \theta_2 m C_2}{1 + \theta_1 m C_1}, \quad |\theta_2| \leq 1, \quad |\theta_1| \leq 1. \end{aligned} \tag{P.17}$$

It follows then from the second inequality (P.13) that

$$\frac{1}{2} \leq U_{\nu} \leq 2 \quad (\nu \geq N). \tag{P.18}$$

We have therefore from (P.17) and (P.14)

$$\xi_{\nu+n+1}^{\phi} > \frac{T}{2} \prod_{\kappa=1}^n \xi_{\nu+\kappa} (1 - \delta)^n$$

and, using (P.11) and again (P.14),

$$\begin{aligned} \xi_{\nu+n+1}^{\phi-1} &> \frac{T}{4} \frac{\xi_{\nu+1}}{\xi_{\nu+n+1}} \prod_{\kappa=2}^n \xi_{\nu+\kappa}, \\ \xi_{\nu+n+1}^{\phi-1} &> \frac{T}{4 \delta^{\phi+1}} \prod_{\kappa=2}^n \xi_{\nu+\kappa}. \end{aligned} \tag{P.19}$$

9. On the other hand, replacing ν in (P.17) by $\nu + 1$ and using (P.18), we have

$$\xi_{\nu+n+2}^{\phi} \leq 2T |z_{\nu+n+2} - z_{\nu+n+1}| \prod_{\kappa=2}^n (\xi_{\nu+n+2} + \xi_{\nu+\kappa}).$$

By (P.16) we have here, however,

$$|z_{\nu+n+2} - z_{\nu+n+1}| \leq \left(1 + \frac{1}{\delta}\right) \xi_{\nu+n+2} \leq \frac{2}{\delta} \xi_{\nu+n+2}$$

and get therefore

$$\xi_{\nu+n+2}^{\phi-1} \leq \frac{4T}{\delta} \prod_{\kappa=2}^n (\xi_{\nu+n+2} + \xi_{\nu+\kappa}). \tag{P.20}$$

Consider now the $n - 1$ numbers $\xi_{\nu+2}, \dots, \xi_{\nu+n}$. Arrange them in decreasing order and denote them in this order by $\eta_1, \eta_2, \dots, \eta_{n-1}$. Further, put

$$\xi_{\nu+n+1} = \eta_n, \quad \xi_{\nu+n+2} = \eta.$$

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*

$$\xi_\nu < \frac{\delta(\nu+1)^{2n}}{2^{n+1}|\gamma|} \quad (\nu \geq N) \quad (\text{P.21})$$

and that, beyond (P.14),

$$\xi_{\nu+n+1} \leq \delta(\nu+1)^{2n} \cdot \min(\xi_{\nu+1}, \dots, \xi_{\nu+n}). \quad (\text{P.22})$$

Then we have

$$\xi_{\nu+n+\kappa+1} \leq \delta(\nu+1)^{2n-\kappa} \xi_{\nu+n+\kappa} \quad (\kappa = 1, 2, \dots, 2n). \quad (\text{P.23})$$

Proof. To prove Lemma 3 it is sufficient to apply Lemma 2 $2n$ times, replacing there δ by

$$\delta(\nu+1)^{2n-1}, \delta(\nu+1)^{2n-2}, \dots, \delta^{\nu+1}, \delta.$$

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 $\nu + n = \mu$,

$$\xi_{\mu+\kappa+1} \leq \delta \xi_{\mu+\kappa} \quad (\kappa = 0, 1, \dots, 2n). \quad (\text{P.24})$$

It is easy to see that (P.22) is satisfied for infinitely many indices ν . Indeed, this will follow from

$$\liminf_{\nu \rightarrow \infty} \frac{\xi_{\nu+n+1}}{\min(\xi_{\nu+1}, \dots, \xi_{\nu+n})} = 0. \quad (\text{P.25})$$

If (P.25) were not true, there would exist a positive $P < 1$, such that for all $\nu \geq 0$

$$\xi_{\nu+n+1} \geq P \min(\xi_{\nu+1}, \dots, \xi_{\nu+n}).$$

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

$$\xi_{\mu_\tau} \geq P\xi_{\mu_{\tau+1}} \quad (\tau = 0, 1, \dots, t-1),$$

$$\mu_\tau > \mu_{\tau+1} \geq \mu_\tau - n \quad (\tau = 0, 1, \dots, t-1), \quad \mu_t \leq n + 1.$$

Here we have obviously $t < \mu_0 = \mu$ and

$$\xi_\mu \geq P^t \min(\xi_1, \dots, \xi_n),$$

$$\xi_\mu \geq P^\mu \min(\xi_1, \dots, \xi_{n+1}),$$

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:

$$\prod_{\kappa=1}^n = \prod_{\kappa=1}^{n-\pi} \cdot \prod_{\kappa=n-\pi+1}^n.$$

The first right-hand product is here

$$\xi_{\mu+\pi+2}\xi_{\mu+\pi+3}\dots\xi_{\mu+n+1}.$$

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

$$\prod_{\kappa=n-\pi+1}^n \xi_{\mu+\pi+1+\kappa} = \prod_{\sigma=1}^{\pi} \xi_{\mu+n+1+\sigma},$$

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

$$\xi_{\mu+n+\pi+2}^\phi > P_\pi \xi_{\mu+n+1}^{n-\pi} \quad (0 < \pi < n). \quad (\text{P.30})$$

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 , occurring in the 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 ν_0 onward

$$\xi_{\nu+1} \leq \delta \xi_\nu \quad (\nu \geq \nu_0),$$

and since $\delta > 0$ can be assumed arbitrarily small, finally, under the hypotheses of Theorem 18.1,

$$\frac{\xi_{\nu+1}}{\xi_\nu} \rightarrow 0 \quad (\nu \rightarrow \infty). \quad (\text{P.34})$$

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)

$$\xi_{\nu+n+1}^p = \Gamma^{n-p} (1 + O(\xi_{\nu+1})) \prod_{\kappa=1}^n \xi_{\nu+\kappa} \prod_{\kappa=1}^n \left(1 + O\left(\frac{\xi_{\nu+n+1}}{\xi_{\nu+\kappa}}\right) \right).$$

We can write this as

$$\xi_{\nu+n+1}^p = \Gamma^{n-p} \prod_{\kappa=1}^n \xi_{\nu+\kappa} (1 + \varepsilon_\nu) \quad (\text{P.36})$$

where

$$\varepsilon_\nu = O\left(\xi_{\nu+1} + \sum_{\kappa=1}^n \frac{\xi_{\nu+\kappa+1}}{\xi_{\nu+\kappa}}\right) \rightarrow 0. \quad (\text{P.37})$$

Now put for $\nu = 1, 2, \dots$

$$t_\nu = -\log(\Gamma\xi_\nu) \rightarrow \infty, \quad w_\nu = t_{\nu+1} - t_\nu \rightarrow \infty, \quad v_\nu = w_\nu - 1. \quad (\text{P.38})$$

Then we obtain from (P.36)

$$pt_{\nu+n+1} - \sum_{\kappa=1}^n t_{\nu+\kappa} = O(\varepsilon_\nu). \quad (\text{P.39})$$

Replacing in this formula ν by $\nu - 1$ and subtracting we get

$$pw_{\nu+n} - \sum_{\kappa=1}^n w_{\nu+\kappa-1} = o(1).$$

Introducing here the v_ν by (P.38) and $\kappa - 1$ instead of κ as the summation variable, we have finally

$$pv_{\nu+n} - \sum_{\kappa=0}^{n-1} v_{\nu+\kappa} = o(1) + n - p.$$

18. Since $v_\nu \rightarrow \infty$ it follows now that from a certain ν onward, $\nu \geq \nu_0$, all v_ν are positive and we have

$$pv_{\nu+n} - \sum_{\kappa=0}^{n-1} v_{\nu+\kappa} > 0 \quad (\nu \geq \nu_0).$$

Putting then $v_{\nu_0+\nu} = u_\nu$, the conditions of Theorem 12.3 are satisfied for the sequence u_ν , and we have therefore, for an $\alpha > 0$ and a σ which is, by (13.17), > 1 ,

$$u_\nu \geq \alpha \sigma^\nu \quad (\nu = 1, 2, \dots),$$

that is,

$$v_\nu \geq \frac{\alpha}{\sigma^{\nu_0}} \sigma^\nu \quad (\nu \geq \nu_0).$$

But then it follows from (P.38) that for a certain positive β we have

$$w_\nu \geq \beta \sigma^\nu \quad (\nu \geq \nu_0),$$

$$\frac{\xi_\nu}{\xi_{\nu+1}} \geq \exp(\beta \sigma^\nu) \quad (\nu \geq \nu_0),$$

and we see that we have

$$\frac{\xi_{\nu+1}}{\xi_\nu} = O(s^\nu)$$

for any arbitrarily small positive s . Since the same holds, by (18.13), for ξ_ν , we have in (P.37) $\varepsilon_\nu = O(s^\nu)$ so that (P.39) becomes

$$t_{\nu+n+1} - \frac{1}{p} \prod_{\kappa=1}^n t_{\nu+\kappa} = O(s^\nu) \quad (\text{P.40})'$$

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

$$t_\nu = \alpha \mu_{n,p}^\nu + O(q_{n,p}^\nu),$$

where α must be > 0 since $t_\nu \rightarrow \infty$. Using (P.38) we get finally

$$|z_\nu - \zeta| = \frac{1}{\Gamma} \exp(-\alpha \mu_{n,p}^\nu) (1 + O(q_{n,p}^\nu)), \quad (\text{P.41})$$

$$\frac{|z_{\nu+1} - \zeta|}{|z_\nu - \zeta|^{\mu_{n,p}}} = \Gamma^{\mu_{n,p}-1} + O(q_{n,p}^\nu), \quad (\text{P.42})$$

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).

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. Lab., Stanford Univ., 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 Publ., 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, 1964.

Chapter 1

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. 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," *Ann. École Norm. Sup. Paris* **4**,

* Numbers in square brackets refer to the bibliography on this page.

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. Cf. 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,” *Jahresber. deut. Math.-Ver.* 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*,” Paris, Buré frères, 1829, reprinted in *Oeuvres complètes*, 2nd series, Vol. IV, Gauthier-Villars, Paris, 1899. Compare in particular in this reprint pp. 576, 578, 593–594. As to Theorem 7.1, Cauchy assumes instead of $|f'(z_0)| \geq 2|h_0|M$ the relation

$$\min_{K_0} |f'(z)| > 4|h_0|M.$$

In Theorem 7.2 Cauchy’s condition is $\min_{J_0} |f'(x)| > 2|h_0|M$ instead of (7.1).

Chapter 8

Formula (8.1) is due to Schröder [5], p. 325. Theorem 8.1 is due to Ostrowski, 1st ed.

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), (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 γ* (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

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," *Tôhoku Math. J.* **2**, 140, 1912. The discussion of Section 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 Schröder [5], p. 329. Theorems 14.1 and 14.2 are due to Ostrowski, 1st ed.

Chapters 15 and 16

Unpublished. For the background of the method cf. 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. Industrial and Appl. Math.* **1**, Series B, pp. 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. u. Physik* **48**, 84–95, 1939. For a more general treatment of norms of matrices cf. 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, 1930. Another special case appears in G. Schulz, "Über die Lösung von Gleichungen durch Iteration," *Z. angew. Math. Mech.* 22, 234–235, 1942. In this 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 content of the first sections belongs to the classical theory of matrices. The inequality (23.20) is special case of a formula derived in the author's paper quoted in the note to Chapter 19. The results of Sections 6–7 are found partly in the author's paper "On some Metrical Properties of Operator Matrices and Matrices Partitioned into Blocks," *J. Math. Anal. and Appl.* 2, 1961, 161–209. I owe the idea of using relation (23.21) to an observation by Prof. F. L. Bauer.

Chapter 24

The results of these sections 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.

Chapters 25 and 26

The first convergence proof of the Newton-Raphson method for $n = 2$ was given by F. A. Willers, but Willers' conditions implied the *third derivatives*. The first proof implying only the second derivatives was given by the author (1936) for $n = 2$. The proof for general n , which presents only formal complications but no difficulties in substance, was given in a doctoral thesis by K. Bussmann and communicated by F. Rehbock in *Z. angew. Math. Mech.* 22, 361–362, 1942. (The complete thesis was submitted to the T. H. Braunschweig in 1940 but apparently never published.) The norm used by Bussmann was $|\xi|_\infty$. L. V. Kantorovich succeeded then in adapting the proof to very general problems of functional analysis. Compare his fundamental paper, "Functional Analysis and Applied Mathematics," *Uspekhi Mat. Nauk* 8, 1948 (in Russian), translated by C. D. Bender and edited by G. E. Forsythe as *Nat. Bureau Standards Rept.* 1509, 1952. The discussion in these and in the preceding chapters can be easily extended to more general classes of norms.

Chapter 27

The idea of the method was indicated by Cauchy in a note in *Compt. rend.* **25**, pp. 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], 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. Cf. Householder [1] 48–49 and A. A. Goldstein, *Numerische Mathematik* 4, pp. 146–150, 1962, where further bibliography can be found. The concept of convergence to the set Ω^* in this connection is apparently new.

Chapter 28

The results of this chapter are new.

Chapter 29

The results of this chapter are new, but the special case that $\psi_\mu = \Phi_\mu$ was treated in Goldstein's paper, quoted above.

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**, 1942–43, p. 740, 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.* **μ**, 313–324, 1960/61.
2. J. W. Schmidt, "Eine Übertragung der Regula Falsi auf Gleichungen in Banach-Räumen," *Z. angew. Math. Mech.* pp. 1–8, 97–110, 1963.
3. L. Bittner, "Mehrpunktverfahren zur Auflösung von Gleichungssystemen," *Z. angew. Math. Mech.* pp. 111–120, 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. Rumbshiski in ref. [8].

Appendix H

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* **86**, 103–106, 1918) and E. T. Whittaker and G. Robinson, *The Calculus of Observation*, Blackie and Son, London and Glasgow, 1928, § 60. 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," *Jahresber. deut. Mat.-Ver.* **60**, 40–42, 1957. A translation of this note by G. C. Bump, edited by G. E. Forsythe, appeared as Tech. Report. No. 2, Applied Mathematics and Statistics Laboratories, Stanford University, Stanford, California.

Appendix L

The formulas of this appendix are classical but the complete formulas (L.7), (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 Mathematiques*, 2 serie, XIX, pp. 161–172, 193–202, 1880, reprinted in *Oeuvres de Laguerre*, Vol. I, pp. 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*, XLIX, p. 911–921, 1946, and J. G. van der Corput, *Proc. Acad. Amsterdam*, XLIX, pp. 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. (Cf. 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.

Index

A

- Accelerating convergence*, methods for ...
40, 241, 251, 267
Aitkin's transformation, 241
Algebraic equation in one unknown, solution by the gradient method, 208
Ampère, 327
Asymptotic behavior of solutions of linear differential equations, 92, ... errors, 39, 51, 67, 71, 76, 86, 96, 106, 108, 119, 121, 144, 252–254, 312, 326
Attraction, points of ... 38, 161

B

- Bauer, F. L., 331
Bauer, M., 328
Bernoulli-L'Hôpital rule, 294
Bernoullian method for solution of equations, 280
Bilinear form, 169
Bittner, L., 333
Bodewig, E., 334
Bödewadt, U. T., 332
Bussmann, K., 331

C

- Cauchy, A. L., 327, 329, 332, 334
Cauchy's theorem, 53, 96
Cauchy-Schwarz inequality, 151, 168
Center of an iteration, 38
Characteristic equation, root, vector of a matrix, 149
Characteristic polynomial of a linear differential equation, 88, special ... 102, table of zeros of ... 108

- Closed interval, 1, 3, 20
Coincident interpolation points, 56
Common zeros in numerator and denominator, 289
Complex variable, functions of ... 21, 47, 60, 63, 110, 127, 136, 141, 302
Conformal mapping, 109
Connected zeros, 222, 228
Continuity, relative, of roots, 225
Convergence, cubic, 119, linear, 30, 44, weakly linear, 204, superlinear, 31, 36, 267
Convex polygonal line, 3, 234
van der Corput, J. G., 334

D

- Dandelin, G. P., 329, ... bounds, 75
Darboux's theorem, 16, 327
Definiteness of a quadratic form, 169
Degree of convergence, 44
Derivatives of the inverse function, 17–19, 235, 332
Diagonal form of a matrix, 150
Difference equation, linear, 31, 88
Divided differences, 1, in the confluent case (with repeated arguments), 7
Division of power series, 90
Double precision, 37

E

- Efficiency index, 32, 58, 106
Eigenvalue, eigenvector of a matrix, 149
Eneström, G., 105, 329
Entire functions, 138
Error estimates *a priori*, *a posteriori*, 59

- Euler, L., 197
Examples of computation, 36, 37, 87, 116, 251, 281
Extrapolation versus interpolation, 33
- F**
- Faà di Bruno, 300
Faddeyeva, V. N., 330
Fejér, L., 328
Fibonacci sequence, 31
Fixed point, *see* Center of an iteration
Fourier, J. B., 328, 329, Fourier bounds, 70, Fourier conditions, 29, 70
Fricke, R., 334
Frobenius, 330
Functional Analysis, 58, 328
Fundamental equation, root, vector of a matrix, 149
- G**
- Gauss, C. F., 239
Geiringer-v. Mises, H., 328
Generating series, 89
Goldstein, A. A., 332
Günther, S., 278
- H**
- Hadamard, J., 280, ...'s estimate of a determinant, 283
Hardy, G. H., 232
Hausholder, A., 242, 244, 327, 330, 332
Hessian matrix, 201
Horner unit, 32
- I**
- Inner product of vectors, 151
Integral representation of divided difference, 3
Interior of an interval, 1
Interpolating function, 11
Interpolation abscissas, points 11, multiple ... 82, 109
Interval, closed, 1, 3, 20, open, 1
Inverse function, 17, 109, 235
Inverse interpolation, 15, 27, 57
- Isobaric polynomial, 18
Iterating function, 38
Iteration, 38, 87, ... of order k , 267
- J**
- Jacobian, 207, ... matrix, 162, 183
Jordan's canonical form of a matrix, 150
- K**
- Kakaya, S., 105, 329, 330
Kantorovitch, L. V., 331
Karamata, J., 328
Kincaid, W. M., 333
Kronecker, L., 276
- L**
- Lagrange, J. L., 220
Lagrange interpolation, 12
Laguerre, 313, 334, ... iteration, 305
Leibniz' formula for $(uv)^{(v)}$, 68
Length of arc in $|\xi|_p$ -metric, 173
Linear, weakly linear convergence in the gradient method, 204, 210
Linear functions as interpolating functions, 82
Littlewood, J. E., 232
- M**
- Matrix, 145, "size" of ... 147
Mean value formula for divided difference, 5
Metric, $|\xi|_p$ -... 173
Milne-Thomson, L. M., 327, 329
Minimum, regular ... 204
v. Mises, R., 328
Monotonic iterating functions, 49
de Morgan, A., 333
Multiple roots, 51, 66, 67, 121, 288, 289, 293, ... interpolation abscissas, 82, 109
- N**
- Negative definite, semidefinite quadratic forms, 169
Neighborhood, 176, one-sided ... 40

- Newton, I., 56, 232, 330
 Newton's interpolation formula, 7
Newton-Raphson method, 43, 56, 58, 59, 70, 75, its generalization for several variables, 183, 186, its analog for multiple roots, 67
 Nörlund, N. E., 329
 Norms of vectors, 145
 Notations, (J_x) 1, $[x_1, \dots, x_m]$ 2, $\langle x_1, \dots, x_m \rangle$ 3, $P(x_1, \dots, x_m)$ 3, \prec, \succ 5, $[t_1^{m_1}, \dots, t_k^{m_k}]$ 7, \otimes 9, $=15$, \uparrow, \downarrow 23, $O(\cdot)$ 24, $o(\cdot)$ 24, (a, b, c) 26, 21.2(\pm 6) 43, \ll 91, $\operatorname{sgn} a$ 134, $|\xi|_p$, $|\xi|_\infty$ 145, $|A|_\infty$ 146, ξ' 146, $|A|_1$ 147, λ_A 149, (ξ, η) 151, $(\partial(f_i)/\partial(z_j))$ 162, $A_1(A)$, $A_2(A)$, $A_\infty(A)$ 167, $K(\xi_1, \dots, \xi_n)$ 170, $A_q(A)$, $A_q^*(A)$ 183, $f'(\xi)$ 196, $f''(\xi)$ 197, $f''(\xi_1, \dots, \xi_n)$ 197, $\xi \rightarrow Q^*$ 198

O

- Obreschkoff, N., 334
 One-sided neighborhood, 40
 Open interval, 1

P

- Poincaré, H., 91
Points of attraction, 38, ... of repulsion, 161, 164, of definite repulsion, 39
 Polya, G., 232
 Polynomial interpolation, 12
Positive definite, semidefinite quadratic forms, 169
Power series, use of, 90, 271
 Principle of permanence, 21
Products, infinite, 78

Q

- Quadratic convergence, 67, 328
 Quadratic forms, 169

R

- Raphson, 56
Regula falsi, 26, 54, 58, 87, 94, 328, 332, 333, for two equations, 239

- Regular matrix*, 149, ... minimum, 204
 Rehbock, F., 331
 Relative continuity, 225
 Rella, T., 330
 Remainder terms of interpolation formulas, 9, 10, 12–13, 288
Repulsion, points of ... 161, 164, points of definite ... 38, 80
 Ritt, J. F., 328
 Robinson, G., 42, 327
 Rolle's theorem, 103, 117, 288
 Roots of special algebraic equations, 90, 102
 Rouché's theorem, 111, 222
 Rounding-off rule, 33, 256
 Rumshiski, L. S. and B. I., 333
 Runge's notation, 9

S

- Scarborough, J., 331
 Schmidt, J. W., 333
Schröder, E., 327, 329, 330, ...'s series, 24, 25, 312, its analog for multiple roots, 293
 Schröder, J., 328
 Schulz, G., 331
 Square root iteration, 117
 Stability of convergence of A^n , 155
Steepest descent, method of ... 197
 Steffensen, J. F., 42, 241, 327
 Stieltjes, 327
 Sylvester's determinant formula, 273
 Symbols, see Notations
 Symmetric matrices, 169
 Symmetry of divided differences, 2

T

- Taylor approximation* to the roots, errors of ... 112
 Taylor development of the root, 109, ... expansion, 68
 Theremin, 330
 Total differentiability, 161
 Transform of a matrix, 149
 Transpose of a vector, 146

Traub, J. F., 327, 330

Triangular inequality, 167

Triangular scheme for divided differences,
13

V

Vector, 145

W

Weakly linear convergence, *see* Conver-
gence

Weber, H., 334

Weight in a polynomial, 18

Weissinger, J., 328

Whittaker, E. T., 42, 273, 327, 333

Willers, F. A., 241, 327, 331

Wronski, H., 272

Z

Zeros of interpolating polynomials, 136,

. . . of special polynomials, 102

