



**STUDIES IN COMPUTATIONAL MATHEMATICS 14**

editors: **C.K. CHUI** and **L. WUYTACK**

# **NUMERICAL METHODS FOR ROOTS OF POLYNOMIALS**

## **Part I**

---

**J.M. McNAMEE**

NUMERICAL METHODS FOR ROOTS OF POLYNOMIALS  
PART I

# STUDIES IN COMPUTATIONAL MATHEMATICS 14

*Editors:*

C.K. CHUI

*Stanford University  
Stanford, CA, USA*

L. WUYTACK

*University of Antwerp  
Antwerp, Belgium*



ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford • Paris  
San Diego • San Francisco • Singapore • Sydney • Tokyo

NUMERICAL METHODS FOR ROOTS  
OF POLYNOMIALS  
PART I

J.M. MCNAMEE

*Department of Computer Science, York University,  
4700 Keele Street,  
Toronto, Ontario,  
M3J 1P3 Canada*



ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford • Paris  
San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier

Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands  
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

First edition 2007

Copyright © 2007 Elsevier B.V. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: [permissions@elsevier.com](mailto:permissions@elsevier.com). Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

#### **Library of Congress Cataloging-in-Publication Data**

A catalog record for this book is available from the Library of Congress

#### **British Library Cataloguing in Publication Data**

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-444-52729-5

ISBN-10: 0-444-52729-X

ISSN: 1570-579X

For information on all Elsevier publications visit our website at <a href="http://books.elsevier.com">books.elsevier.com</a>
---------------------------------------------------------------------------------------------------------------------------------

Printed and bound in The Netherlands

07 08 09 10 11 10 9 8 7 6 5 4 3 2 1

# Dedication

This work is dedicated to my wife, Jean Frances McNamee, who over the last five years has spent innumerable hours driving around the Haliburton Highlands while I read hundreds of journal articles in preparation for the book.

This page intentionally left blank

# Preface

This book constitutes the first part of two volumes describing methods for finding roots of polynomials. In general most such methods are numerical (iterative), but one chapter in Part II will be devoted to “analytic” methods for polynomials of degree up to four.

It is hoped that the series will be useful to anyone doing research into methods of solving polynomials (including the history of such methods), or who needs to solve many low- to medium-degree polynomials and/or some or many high-degree ones in an industrial or scientific context. Where appropriate, the location of good computer software for some of the best methods is pointed out. The book(s) will also be useful as a text for a graduate course in polynomial root-finding.

Preferably the reader should have as pre-requisites at least an undergraduate course in Calculus and one in Linear Algebra (including matrix eigenvalues). The only knowledge of polynomials needed is that usually acquired by the last year of high-school Mathematics.

The book(s) cover most of the traditional methods for root- finding (and numerous variations on them), as well as a great many invented in the last few decades of the twentieth and early twenty-first centuries. In short, it could well be entitled: “ A Handbook of Methods for Polynomial Root-Solving”.



This page intentionally left blank

# Contents

<b>Preface</b>	<b>vii</b>
<b>Contents</b>	<b>ix</b>
<b>Introduction</b>	<b>xiii</b>
<b>1 Evaluation, Convergence, Bounds</b>	<b>1</b>
1.1 Horner's Method of Evaluation . . . . .	1
1.2 Rounding Errors and Stopping Criteria . . . . .	4
1.3 More Efficient Methods for Several Derivatives . . . . .	7
1.4 Parallel Evaluation . . . . .	8
1.5 Evaluation at Many Points . . . . .	11
1.6 Evaluation at Many Equidistant Points . . . . .	14
1.7 Accurate Evaluation . . . . .	17
1.8 Scaling . . . . .	22
1.9 Order of Convergence and Efficiency . . . . .	23
1.10 A Priori Bounds on (Real or Complex) Roots . . . . .	26
1.11 References for Chapter 1. . . . .	33
<b>2 Sturm Sequences and Greatest Common Divisors</b>	<b>37</b>
2.1 Introduction . . . . .	37
2.2 Definitions and Basic Theorem . . . . .	37
2.3 Application to Locating Roots . . . . .	38
2.4 Elimination of Multiple Roots . . . . .	40
2.5 Detection of Clusters of Zeros (Near-Multiple) . . . . .	43
2.6 Sturm Sequences (or gcd's) Using Integers . . . . .	47
2.7 Complex Roots (Wilf's Method) . . . . .	49
2.8 References for Chapter 2 . . . . .	51
<b>3 Real Roots by Continued Fractions</b>	<b>53</b>
3.1 Fourier and Descartes' Theorems . . . . .	53
3.2 Budans's Theorem . . . . .	54
3.3 Vincent's Theorem . . . . .	55

3.4	Akritas' Improvement of Vincent's Theorem . . . . .	56
3.5	Applications of Theorem 3.4.1 . . . . .	60
3.6	Complexity of $m$ . . . . .	60
3.7	Choice of the $a_i$ . . . . .	61
3.8	Cauchy's Rule . . . . .	62
3.9	Appendix to Chapter 3. Continued Fractions . . . . .	62
3.10	References for Chapter 3 . . . . .	65
<b>4</b>	<b>Simultaneous Methods</b>	<b>67</b>
4.1	Introduction and Basic Methods . . . . .	67
4.2	Conditions for Guaranteed Convergence . . . . .	72
4.3	Multiple Roots . . . . .	80
4.4	Use of Interval Arithmetic . . . . .	84
4.5	Recursive Methods . . . . .	90
4.6	Methods Involving Square Roots, Second Order Derivatives, etc . . . . .	92
4.7	Effect of Rounding Errors . . . . .	100
4.8	Gauss-Seidel and SOR Variations . . . . .	102
4.9	Real Factorization Methods . . . . .	108
4.10	Comparison of Efficiencies . . . . .	114
4.11	Implementation on Parallel Computers . . . . .	117
4.12	Miscellaneous Methods . . . . .	120
4.13	A Robust and Efficient program . . . . .	121
4.14	References for Chapter 4 . . . . .	123
<b>5</b>	<b>Newton's and Related Methods</b>	<b>131</b>
5.1	Definitions and Derivations . . . . .	131
5.2	Early History of Newton's Method . . . . .	135
5.3	Computable Conditions for Convergence . . . . .	137
5.4	Generalizations of Newton's Method . . . . .	141
5.5	Methods for Multiple Roots . . . . .	151
5.6	Termination Criteria . . . . .	158
5.7	Interval Methods . . . . .	161
5.8	Parallel Methods . . . . .	173
5.9	Hybrid Methods Involving Newton's Method . . . . .	175
5.10	Programs . . . . .	189
5.11	Miscellaneous Methods Related to Newton's . . . . .	190
5.12	References for Chapter 5 . . . . .	196
<b>6</b>	<b>Matrix Methods</b>	<b>207</b>
6.1	Methods Based on the Classical Companion Matrix . . . . .	207
6.2	Other Companion Matrices . . . . .	214
6.3	Methods with $O(N^2)$ Operations . . . . .	231
6.4	Methods Designed for Multiple Roots . . . . .	257
6.5	Methods for a Few Roots . . . . .	289

*Contents*

6.6	Errors and Sensitivity . . . . .	294
6.7	Miscellaneous Methods and Special Applications . . . . .	304
6.8	Programs and Packages . . . . .	314
6.9	References for Chapter 6 . . . . .	316

**Index**

This page intentionally left blank

# Introduction

A polynomial is an expression of the form

$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0 \quad (1)$$

If the highest power of  $x$  is  $x^n$ , the polynomial is said to have degree  $n$ . It was proved by Gauss in the early 19th century that every polynomial has at least one zero (i.e. a value  $\zeta$  which makes  $p(\zeta)$  equal to zero), and it follows that a polynomial of degree  $n$  has  $n$  zeros (not necessarily distinct). Often we use  $x$  for a real variable, and  $z$  for a complex. A zero of a polynomial is equivalent to a “root” of the equation  $p(x) = 0$ . A zero may be real or complex, and if the “coefficients”  $c_i$  are all real, then complex zeros occur in conjugate pairs  $\alpha + i\beta$ ,  $\alpha - i\beta$ . The purpose of this book is to describe methods which have been developed to find the zeros (roots) of polynomials.

Indeed the calculation of roots of polynomials is one of the oldest of mathematical problems. The solution of quadratics was known to the ancient Babylonians, and to the Arab scholars of the early Middle Ages, the most famous of them being Omar Khayyam. The cubic was first solved in closed form by G. Cardano in the mid-16th century, and the quartic soon afterwards. However N.H. Abel in the early 19th century showed that polynomials of degree five or more could not be solved by a formula involving radicals of expressions in the coefficients, as those of degree up to four could be. Since then (and for some time before in fact), researchers have concentrated on numerical (iterative) methods such as the famous Newton’s method of the 17th century, Bernoulli’s method of the 18th, and Graeffe’s method of the early 19th. Of course there have been a plethora of new methods in the 20th and early 21st century, especially since the advent of electronic computers. These include the Jenkins-Traub, Larkin’s and Muller’s methods, as well as several methods for simultaneous approximation starting with the Durand-Kerner method. Recently matrix methods have become very popular. A bibliography compiled by this author contains about 8000 entries, of which about 50 were published in the year 2005.

Polynomial roots have many applications. For one example, in control theory we are led to the equation

$$y(s) = G(s)u(s) \quad (2)$$

where  $G(s)$  is known as the “transfer function” of the system,  $u(s)$  is the Laplace transform of the input, and  $y(s)$  is that of the output.  $G(s)$  usually takes the form  $\frac{P(s)}{Q(s)}$  where  $P$  and  $Q$  are polynomials in  $s$ . Their zeros may be needed, or we may require not their exact values, but only the knowledge of whether they lie in the left-half of the complex plane, which indicates stability. This can be decided by the Routh-Hurwitz criterion. Sometimes we need the zeros to be inside the unit circle. See Chapter 15 in Volume 2 for details of the Routh-Hurwitz and other stability tests.

Another application arises in certain financial calculations, e.g. to compute the rate of return on an investment where a company buys a machine for, (say) \$100,000. Assume that they rent it out for 12 months at \$5000/month, and for a further 12 months at \$4000/month. It is predicted that the machine will be worth \$25,000 at the end of this period. The solution goes as follows: the present value of \$1 received  $n$  months from now is  $\frac{1}{(1+i)^n}$ , where  $i$  is the monthly interest rate, as yet unknown. Hence

$$100,000 = \sum_{j=1}^{12} \frac{5000}{(1+i)^j} + \sum_{j=13}^{24} \frac{4000}{(1+i)^j} + \frac{25,000}{(1+i)^{24}} \quad (3)$$

Hence

$$100,000(1+i)^{24} - \sum_{j=1}^{12} 5000(1+i)^{24-j} - \sum_{j=13}^{24} 4000(1+i)^{24-j} - 25,000 = 0 \quad (4)$$

a polynomial equation in  $(1+i)$  of degree 24. If the term of the lease was many years, as is often the case, the degree of the polynomial could be in the hundreds.

In signal processing one commonly uses a “linear time-invariant discrete” system. Here an input signal  $x[n]$  at the  $n$ -th time-step produces an output signal  $y[n]$  at the same instant of time. The latter signal is related to  $x[n]$  and previous input signals, as well as previous output signals, by the equation

$$y[n] = b_0x[n] + b_1x[n-1] + \dots + b_Nx[n-N] + a_1y[n-1] + \dots + a_My[n-M] \quad (5)$$

To solve this equation one often uses the “z-transform” given by:

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n} \quad (6)$$

A very useful property of this transform is that the transform of  $x[n-i]$  is

$$z^{-i}X(z) \quad (7)$$

Then if we apply 6 to 5 using 7 we get

$$\begin{aligned} Y(z) &= b_0X(z) + b_1z^{-1}X(z) + \dots + b_Nz^{-N}X(z) + \\ &a_1z^{-1}Y(z) + \dots + a_Mz^{-M}Y(z) \end{aligned} \quad (8)$$

and hence

$$Y(z) = X(z) \frac{[b_0 + b_1z^{-1} + \dots + b_Nz^{-N}]}{[1 - a_1z^{-1} - \dots - a_Mz^{-M}]} \quad (9)$$

$$= X(z)z^{M-N} \frac{[b_0z^N + b_1z^{N-1} + \dots + b_N]}{[z^M - a_1z^{M-1} - \dots - a_M]} \quad (10)$$

For stability we must have  $M \geq N$ . We can factorize the numerator and denominator polynomials in the above (or equivalently find their zeros  $z_i$  and  $p_i$  respectively). Then we may expand the right-hand-side of 10 into partial fractions, and finally apply the inverse z-transform to get the components of  $y[n]$ . For example the inverse transform of  $\frac{z}{z-a}$  is

$$a^n u[n] \quad (11)$$

where  $u[n]$  is the discrete step-function, i.e.

$$\begin{aligned} u[n] &= 0 \quad (n < 0) \\ &= 1 \quad (n \geq 0) \end{aligned} \quad (12)$$

In the common case that the denominator of the partial fraction is a quadratic (for the zeros occur in conjugate complex pairs), we find that the inverse transform is a sin- or cosine- function. For more details see e.g. van den Emden and Verhoeckx (1989).

As mentioned, this author has been compiling a bibliography on roots of polynomials since about 1987. The first part was published in 1993 (see McNamee (1993)), and is now available at the web-site

<http://www.elsevier.com/locate/cam>

by clicking on ‘‘Bibliography on roots of polynomials’’. More recent entries have been included in a Microsoft Access Database, which is available at the web-site

[www.yorku.ca/mcnamee](http://www.yorku.ca/mcnamee)

by clicking on ‘‘Click here to download it’’ (under the heading ‘‘Part of my bibliography on polynomials is accessible here’’). For further details on how to use this database and other web components see McNamee (2002).



We will now briefly review some of the more well-known methods which (along with many variations) are explained in much more detail in later chapters. First we mention the bisection method (for real roots): we start with two values  $a_0$  and  $b_0$  such that

$$p(a_0)p(b_0) < 0 \quad (13)$$

(such values can be found e.g. by Sturm sequences -see Chapter 2). For  $i = 0, 1, \dots$  we compute

$$d_i = \frac{a_i + b_i}{2}, \quad (14)$$

then if  $f(d_i)$  has the same sign as  $f(a_i)$  we set  $a_{i+1} = d_i$ ,  $b_{i+1} = b_i$ ; otherwise  $b_{i+1} = d_i$ ,  $a_{i+1} = a_i$ . We continue until

$$|a_i - b_i| < \epsilon \quad (15)$$

where  $\epsilon$  is the required accuracy (it should be at least a little larger than the machine precision, usually  $10^{-7}$  or  $10^{-15}$ ). Alternatively we may use

$$|p(d_i)| < \epsilon \quad (16)$$

Unlike many other methods, we are guaranteed that 15 or 16 will eventually be satisfied. It is called an **iterative** method, and in that sense is typical of most of the methods considered in this work. That is, we repeat some process over and over again until we are close enough to the required answer (we hardly ever reach it exactly). For more details of the bisection method, see Chapter 7.

Next we consider the famous Newton's method. Here we start with a single initial guess  $x_0$ , preferably fairly close to a true root  $\zeta$ , and apply the iteration:

$$z_{i+1} = z_i - \frac{p(z_i)}{p'(z_i)} \quad (17)$$

Again, we stop when

$$\frac{|z_{i+1} - z_i|}{|z_{i+1}|} < \epsilon \quad (18)$$

or  $|p(z_i)| < \epsilon$  (as in 16). For more details see Chapter 5.

In Chapter 4 we will consider simultaneous methods, such as

$$z_i^{(k+1)} = z_i^{(k)} - \frac{p(z_i^{(k)})}{\prod_{j=1, j \neq i}^n (z_i^{(k)} - z_j^{(k)})} \quad (i = 1, \dots, n) \quad (19)$$

starting with initial guesses  $z_i^{(0)}$  ( $i = 1, \dots, n$ ). Here the notation is a little different from before, that is  $z_i^{(k)}$  is the  $k$ -th approximation to the  $i$ -th zero  $\zeta_i$  ( $i = 1, \dots, n$ ).

Another method which dates from the early 19th century, but is still often used, is Graeffe's. Here 1 is replaced by another polynomial, still of degree  $n$ , whose zeros are the squares of those of 1. By iterating this procedure, the zeros (usually) become widely separated, and can then easily be found. Let the roots of  $p(z)$  be  $\zeta_1, \dots, \zeta_n$  and assume that  $c_n = 1$  (we say  $p(z)$  is "monic") so that

$$f_0(z) \equiv p(z) = (z - \zeta_1) \dots (z - \zeta_n) \quad (20)$$

Hence

$$f_1(w) \equiv (-1)^n f_0(z) f_0(-z) \quad (21)$$

$$= (w - \zeta_1^2) \dots (w - \zeta_n^2) \quad (22)$$

with  $w = z^2$ .

We will consider this method in detail in Chapter 8 in Volume II.

Another popular method is Laguerre's:

$$z_{i+1} = z_i - \frac{np(z_i)}{p'(z_i) \pm \sqrt{(n-1)\{(n-1)[p'(z_i)]^2 - np(z_i)p''(z_i)\}}} \quad (23)$$

where the sign of the square root is taken the same as that of  $p'(z_i)$  (when all the roots are real, so that  $p'(z_i)$  is real and the expression under the square root sign is positive). A detailed treatment of this method will be included in Chapter 9 in Volume II.

Next we will briefly describe the Jenkins-Traub method, which is included in some popular numerical packages. Let

$$H^{(0)}(z) = p'(z) \quad (24)$$

and find a sequence  $\{t_i\}$  of approximations to a zero  $\zeta_1$  by

$$t_{i+1} = s_i - \frac{p(s_i)}{\tilde{H}^{(i+1)}(s_i)} \quad (25)$$

For details of the choice of  $s_i$  and the construction of  $\tilde{H}^{(i+1)}(s_i)$  see Chapter 12 in Volume II.

There are numerous methods based on interpolation (direct or inverse) such as the secant method:

$$x_{i+1} = \frac{p(x_i)}{p(x_i) - p(x_{i-1})} x_{i-1} + \frac{p(x_{i-1})}{p(x_{i-1}) - p(x_i)} x_i \quad (26)$$

(based on linear inverse interpolation) and Muller's method (not described here) based on quadratic interpolation. We consider these and many variations in Chapter 7 of Volume II.

Last but not least we mention the approach, recently popular, of finding zeros as eigenvalues of a “companion” matrix whose characteristic polynomial coincides with the original polynomial. The simplest example of a companion matrix is (with  $c_n = 1$ ):

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 1 \\ -c_0 & -c_1 & \dots & \dots & -c_{n-1} \end{bmatrix} \quad (27)$$

Such methods will be treated thoroughly in Chapter 6.

## References for Introduction

McNamee, J.M. (1993), A bibliography on roots of polynomials, *J. Comput. Appl. Math.* **47**, 391-394

————— (2002), A 2002 update of the supplementary bibliography on roots of polynomials, *J. Comput. Appl. Math.* **142**, 433-434

van den Emden, A.W.M. and Verhoeckx, N.A.M. (1989), *Discrete- Time Signal Processing*, Prentice-Hall, New York

This page intentionally left blank

# Chapter 1

## Evaluation, Convergence, Bounds

### 1.1 Horner's Method of Evaluation

Evaluation is, of course, an essential part of any root-finding method. Unless the polynomial is to be evaluated for a very large number of points, the most efficient method is Horner's method (also known as nested multiplication) which proceeds thus:

Let

$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_r x^r + \dots + c_0 \quad (1.1)$$

$$b_n = c_n; \quad b_k = x b_{k+1} + c_k \quad (k = n-1, n-2, \dots, 0) \quad (1.2)$$

Then

$$p(x) = b_0 \quad (1.3)$$

**Outline of Proof**  $b_{n-1} = x c_n + c_{n-1}$ ;  $b_{n-2} = x(x c_n + c_{n-1}) + c_{n-2} = x^2 c_n + x c_{n-1} + c_{n-2} \dots$  Continue by induction

**Alternative Proof** Let

$$p(z) = (z - x)(b_n z^{n-1} + b_{n-1} z^{n-2} + \dots + b_{n-k} z^{n-k-1} + \dots + b_1) + b_0 \quad (1.4)$$

Comparing coefficients of  $z^n, z^{n-1}, \dots, z^{n-k}, \dots, z_0$  gives

$$\begin{array}{ll}
c_n = b_n & \text{so } b_n = c_n \\
c_{n-1} = b_{n-1} - xb_n & \text{so } b_{n-1} = xb_n + c_{n-1} \\
\cdots & \cdots \\
\cdots & \cdots \\
c_{n-k} = b_{n-k} - xb_{n-k+1} & \text{so } b_{n-k} = xb_{n-k+1} + c_{n-k} \quad (k = 2, \dots, n-1) \\
\cdots & \cdots \\
c_0 = b_0 - xb_1 & \text{so } b_0 = xb_1 + c_0
\end{array} \tag{1.5}$$

Now setting  $z = x$  we get  $p(x) = 0 + b_0$

Note that this process also gives the coefficients of the quotient when  $p(z)$  is divided by  $z-x$ , (i.e.  $b_n, \dots, b_1$ )

Often we require several or all the derivatives, e.g. some methods such as Laguerre's require  $p'(x)$  and  $p''(x)$ , while the methods of Chapter 3 involving a shift of origin  $z = y+x$  use the Taylor Series expansion

$$p(z) = p(x+y) = p(x) + p'(x)y + \frac{p''(x)}{2!}y^2 + \dots + \frac{p^{(k)}(x)}{k!}y^k + \dots + \frac{p^{(n)}(x)}{n!}y^n \tag{1.6}$$

If we re-write 1.4 as

$$P_n(z) = (z-x)P_{n-1}(z) + P_n(x) \tag{1.7}$$

and apply the Horner scheme as many times as needed, i.e.

$$P_{n-1}(z) = (z-x)P_{n-2}(z) + P_{n-1}(x) \tag{1.8}$$

....

....

$$P_{n-k+1}(z) = (z-x)P_{n-k}(z) + P_{n-k+1}(x) \quad (k = 2, \dots, n) \tag{1.9}$$

then differentiating 1.7  $k$  times using Leibnitz' theorem for higher derivatives of a product gives

$$P_n^{(k)}(z) = (z-x)P_{n-1}^{(k)}(z) + kP_{n-1}^{(k-1)}(z) \tag{1.10}$$

Hence

$$P_n^{(k)}(x) = kP_{n-1}^{(k-1)}(x) = k(k-1)P_{n-2}^{(k-2)}(x) = \dots = k!P_{n-k}(x) \tag{1.11}$$

Hence

$$P_{n-k}(x) = \frac{1}{k!}P_n^{(k)}(x) \tag{1.12}$$

These are precisely the coefficients needed in 1.6

**EXAMPLE** Evaluate  $p(x) = 2x^3 - 8x^2 + 10x - 4$  and its derivatives at  $x = 1$ .  
 Write  $P_3(x) = c_3x^3 + c_2x^2 + c_1x + c_0$  and  $P_2(x) = b_3x^2 + b_2x + b_1$  with  $p(1) = b_0$   
 Then  $b_3 = c_3 = 2$ ;  $b_2 = xb_3 + c_2 = -6$ ;  $b_1 = xb_2 + c_1 = 4$ ;  $b_0 = xb_1 + c_0 = 0$

Thus quotient on division by  $(x-1) = 2x^2 - 6x + 4$

Writing  $P_1(x) = d_3x + d_2$  (with  $d_1 = p'(1)$ ):

$d_3 = b_3 = 2$ ,  $d_2 = xd_3 + b_2 = -4$ ,  $d_1 = xd_2 + b_1 = 0 = p'(1)$

Finally write  $P_0(x) = e_3$ , with  $e_2 = \frac{1}{2}p''(1)$

i.e.  $e_3 = d_3 = 2$ ,  $e_2 = xe_3 + d_2 = -2$ ,  $p''(1) = 2e_2 = -4$

**CHECK**  $p(1) = 2-8+10-4 = 0$ ,  $p'(1) = 6-16+10 = 0$ ,  $p''(1) = 12-16 = -4$ , OK

The above assumes real coefficients and real  $x$ , although it could be applied to complex coefficients and argument if we can use complex arithmetic. However it is more efficient, if it is possible, to use real arithmetic even if the argument is complex. The following shows how this can be done.

Let  $p(z) = (z-x-iy)(z-x+iy)Q(z)+r(z-x)+s =$

$$(z^2+pz+q)(b_nz^{n-2}+b_{n-1}z^{n-3}+\dots+b_{n-k}z^{n-k-2}+\dots+b_2)+b_1(z-x)+b_0 \tag{1.13}$$

where  $p = -2x$ ,  $q = x^2 + y^2$ , and thus  $p$ ,  $q$ ,  $x$ , and the  $b_i$  are all real.

Comparing coefficients as before:

$$\begin{array}{ll} c_n = b_n & \text{so } b_n = c_n \\ c_{n-1} = b_{n-1} + pb_n & \text{so } b_{n-1} = c_{n-1} - pb_n \\ \dots & \dots \\ c_{n-k} = b_{n-k} + pb_{n-k+1} + qb_{n-k+2} & \text{so } \dots \\ b_{n-k} = c_{n-k} - pb_{n-k+1} - qb_{n-k+2} & \\ (k = 2, \dots, n-1) & \\ \dots & \dots \\ c_0 = b_0 - xb_1 + qb_2 & \text{so } b_0 = c_0 + xb_1 - qb_2 \end{array} \tag{1.14}$$

Now setting  $z = x+iy$  gives

$$p(x + iy) = b_0 + iyb_1 = R(x, y) + iJ(x, y), \text{ say} \tag{1.15}$$

Wilkinson (1965) p448 shows how to find the derivative  $p'(x + iy) = RD + iJD$ ; we let

$$\begin{array}{l} d_n = b_n, d_{n-1} = b_{n-1} - pd_n, \dots, \\ d_{n-k} = b_{n-k} - pd_{n-k+1} - qd_{n-k+2} \quad (k = 2, \dots, n-3), \dots, \end{array} \tag{1.16}$$



, ...,  $d_2 = b_2 - qd_4$  (but if  $n = 3$ ,  $d_2 = b_2$ )

Then

$$RD = -2y^2d_3 + b_1, JD = 2y(xd_3 + d_2) \quad (1.17)$$

**EXAMPLE** (As before), at  $z=1+i$ ,  $p = -2$ ,  $q = 2$ ,  $b_3 = 2$ ,  $b_2 = -8 - (-2)2 = -4$ ,  $b_1 = 10 - (-2)(-4) - 2 \times 2 = -2$ ,  $b_0 = -4 + (-2) - 2(-4) = 2$ ;  $p(1+i) = 2 - 2i$

**Check**  $p(1+i) = 2(1+i)^3 - 8(1+i)^2 + 10(1+i) - 4 = 2(1+3i-3-i) - 8(1+2i-1) + 10(1+i) - 4 = 2-2i$  OK.

For  $p'(1+i)$ ,  $d_3 = 2$ ,  $d_2 = -4$ ,  $RD = -4 - 2 = -6$ ,  $JD = 2(2-4) = -4$ ;

**Check**  $p'(1+i) = 6(1+i)^2 - 16(1+i) + 10 = 6(1+2i-1) - 16(1+i) + 10 = -6-4i$ . OK.

## 1.2 Rounding Errors and Stopping Criteria

For an iterative method based on function evaluations, it does not make much sense to continue iterations when the calculated value of the function approaches the possible rounding error incurred in the evaluation.

Adams (1967) shows how to find an upper bound on this error. For real  $x$ , he lets

$$h_n = \frac{1}{2}c_n; \dots, h_i = |x|h_{i+1} + |s_i| \quad (i = n-1, \dots, 0) \quad (1.18)$$

where the  $s_i$  are the computed values of the  $b_i$  defined in Sec. 1.

Then the rounding error  $\leq RE =$

$$\beta^{1-t}(h_0 - \frac{1}{2}|s_0|) \quad (1.19)$$

where  $\beta$  is the base of the number system and  $t$  the number of digits (usually bits) in the mantissa.

The proof of the above, from Peters and Wilkinson (1971), follows:-

Equation 1.2 describes the **exact** process, but computationally (with rounding) we have

$$s_n = c_n; s_i = fl(xs_{i+1} + c_i) \quad (i = n-1, \dots, 0); \bar{p}(x) = s_0$$

where  $\bar{p}(x)$  is the computed value of  $p(x)$ .

Now it is well known that  $fl(x+y) = (x+y)/(1+\epsilon)$  and  $fl(xy) = xy(1+\epsilon)$

where  $\epsilon \leq \frac{1}{2}\beta^{1-t} \equiv E$

Hence

$$s_i = \{xs_{i+1}(1 + \epsilon_i) + c_i\}/(1 + \eta_i) \quad (i = n - 1, \dots, 0)$$

where  $|\epsilon_i|, |\eta_i| \leq E$ .

Hence  $s_i = xs_{i+1}(1 + \epsilon_i) + c_i - s_i\eta_i$

Now letting  $s_i = b_i + e_i$  (N.B.  $s_n = c_n = b_n$  and so  $e_n = 0$ ) we have

$$\begin{aligned} b_i + e_i &= x(b_{i+1} + e_{i+1}) + xs_{i+1}\epsilon_i + c_i - s_i\eta_i \\ &= b_i + xe_{i+1} + xs_{i+1}\epsilon_i - s_i\eta_i \end{aligned}$$

and so  $|e_i| \leq |x|\{|e_{i+1}| + |s_{i+1}|E\} + |s_i|E$

Now define

$$g_n = 0; \quad g_i = |x|\{g_{i+1} + |s_{i+1}|\} + |s_i| \quad (i = n - 1, \dots, 0) \quad (1.20)$$

Then we claim that

$$|e_i| \leq g_i E \quad (1.21)$$

For we have

$$\begin{aligned} |e_{n-1}| &\leq |x|\{|e_n| + |s_n|E\} + |s_{n-1}|E \\ &= \{|x||s_n| + |s_{n-1}|\}E \quad (\text{since } e_n = 0) = g_{n-1}E \\ \text{i.e. the result is true for } i=n-1. \end{aligned}$$

Now suppose it is true as far as  $r+1$ , i.e.  $|e_{r+1}| \leq g_{r+1}E$

Then

$$\begin{aligned} |e_r| &\leq |x|\{|e_{r+1}| + |s_{r+1}|E\} + |s_r|E \\ &\leq |x|\{g_{r+1}E + |s_{r+1}|E\} + |s_r|E \\ &= \{|x|(g_{r+1} + |s_{r+1}|) + |s_r|\}E \\ &= g_r E \end{aligned}$$

i.e. it is true for  $r$ . Hence by induction it is true for all  $i$ , down to 0.

The amount of work can be reduced by letting  $h_n = \frac{1}{2}|s_n| = \frac{1}{2}|c_n|$  and  $h_i = \frac{g_i + |s_i|}{2}$  ( $i = n - 1, \dots, 0$ )  
or  $2h_i - |s_i| = g_i = |x|\{g_{i+1} + |s_{i+1}|\} + |s_i| = |x|2h_{i+1} + |s_i|$   
Hence

$$2h_i = 2(|x|h_{i+1} + |s_i|) \quad (i = n - 1, \dots, 0) \quad (1.22)$$

and finally  $g_0 = 2h_0 - |s_0|$  i.e.

$$|e_0| = |s_0 - b_0| \leq g_0 E = (h_0 - \frac{1}{2}|s_0|)\beta^{1-t} \quad (1.23)$$

An alternative expression for the error, derived by many authors such as Oliver (1979) is

$$E \leq \left[ \sum_{k=0}^{n-1} (2k+1)|c_k||x|^k + 2n|c_n||x|^n \right] \frac{1}{2}\beta^{1-t} \quad (1.24)$$

Adams suggest stopping when

$$|p| = |s_0| \leq 2RE \quad (1.25)$$

For complex z, he lets

$$h_n = .8|b_n|, \dots, h_i = \sqrt{q}h_{i+1} + |s_i| \quad (i = n-1, \dots, 0) \quad (1.26)$$

Then

$$RE = \{2|xs_1| - 7(|s_0| + \sqrt{q}|s_1|) + 9h_0\} \frac{1}{2}\beta^{1-t} \quad (1.27)$$

and we stop when

$$|R + iJ| = \sqrt{b_0^2 + y^2 b_1^2} \leq 2RE \quad (1.28)$$

**EXAMPLE** As before, but with  $\beta = 10$  and  $t=7$ .

$$h_3 = 1.6, h_2 = \sqrt{2} \times 1.6 + 4 = 6.3, h_1 = \sqrt{2} \times 6.3 + 2 = 10.8, h_0 = \sqrt{2} \times 10.8 + 2 = 17.1$$

$$RE = \{2 \times 1 \times 2 - 7(2 + 1.4 \times 2) + 9 \times 17.1\} \times \frac{1}{2} \times 10^{-6} = 124 \times \frac{1}{2} \times 10^{-6} = .000062$$

Igarashi (1984) gives an alternative stopping criteria with an associated error estimate:

Let  $A(x)$  be  $p(x)$  evaluated by Horner's method, and let

$$G(x) = (n-1)c_n x^n + (n-2)c_{n-1} x^{n-1} + \dots + c_2 x^2 - c_0 = xp'(x) - p(x) \quad (1.29)$$

and

$$H(x) = xp'(x) \quad (1.30)$$

finally

$$B(x) = H(x) - G(x) \quad (1.31)$$

represents another approximation to  $p(x)$  with different rounding errors. He suggests stopping when

$$|A(x_k) - B(x_k)| \geq \min\{|A(x_k)|, |B(x_k)|\} \quad (1.32)$$

and claims that then the difference between

$$x_k - A(x_k)/p'(x_k) \text{ and } x_k - B(x_k)/p'(x_k) \quad (1.33)$$

represents the size of the error in  $x_{k+1}$  (using Newton's method). Presumably similar comparisons could be made for other methods.

A very simple method based on Garwick (1961) is:

Iterate until  $\Delta_k = |x_k - x_{k-1}| \leq 10^{-2}|x_k|$ .

Then iterate until  $\Delta_k \geq \Delta_{k-1}$  (which will happen when rounding error dominates). Now  $\Delta_k$  gives an error estimate.

### 1.3 More Efficient Methods for Several Derivatives

One such method, suitable for relatively small  $n$ , was given by Shaw and Traub (1974). To evaluate all the derivatives their method requires the same number of additions, i.e.  $\frac{1}{2}n(n+1)$ , as the iterated Horner method described in Sec. 1. However it requires only  $3n-2$  multiplications and divisions, compared to (also)  $\frac{1}{2}n(n+1)$  for Horner. It works as follows, to find  $m \leq n$  derivatives (N.B. it is only worthwhile if  $m$  is fairly close to  $n$ ):

$$\text{Let } T_i^{-1} = c_{n-i-1}x^{n-i-1} \quad (i = 0, 1, \dots, n-1) \quad (1.34)$$

$$T_j^j = c_n x^n \quad (j = 0, 1, \dots, m) \quad (1.35)$$

$$T_i^j = T_{i-1}^{j-1} + T_{i-1}^j \quad (j = 0, 1, \dots, m; i = j+1, \dots, n) \quad (1.36)$$

$$\frac{p^{(j)}}{j!} = \frac{T_n^j}{x^j} \quad (j = 0, 1, \dots, m) \quad (1.37)$$

This process requires  $(m+1)(n - \frac{m}{2})$  additions and  $2n+m-1$  multiplications and divisions. If  $m = n$ , no calculation is required for  $\frac{p^{(n)}}{n!} = c_0$ , so it takes  $\frac{n}{2}(n+1)$  additions and  $3n-2$  multiplications/divisions.

Wozniakowski (1974) shows that the rounding error is bounded by

$$\sum_{k=j}^n C(k, j) |c_k| (2k+1) |x|^{k-j} \frac{1}{2} \beta^{1-t} \quad (1.38)$$

Aho et al (1975) give a method more suitable for large  $n$ . In their Lemma 3 they use in effect the fact that

$$p^{(k)}(r) = \sum_{i=k}^n c_i i(i-1)\dots(i-k+1) r^{i-k} \quad (1.39)$$

$$= \sum_{i=k}^n c_i \frac{i!}{(i-k)!} r^{i-k} \quad (1.40)$$

$$= \sum_{i=0}^n f(i)g(k-i) \quad (1.41)$$

if we define

$$f(i) = c_i i! \quad (i = 0, \dots, n) \quad (1.42)$$

$$g(j) = \begin{cases} r^{-j}/(-j)! & (j = -n, -(n-1), \dots, 0) \\ 0 & (j = 1, \dots, n) \end{cases} \quad (1.43)$$

Then  $f(i)$  and  $g(j)$  can be computed in  $O(n)$  steps, while the right side of 1.41, being a convolution, can be evaluated in  $O(n \log n)$  steps by the Fast Fourier Transform.

## 1.4 Parallel Evaluation

Dorn (1962) and Kiper (1997A) describe a parallel implementation of Horner's Method as follows:

$$\text{Let } p(x) = c_0 + c_1 x + \dots + c_N x^N \quad (1.44)$$

and  $n \geq 2$  be the number of processors operating in parallel. The method is simplified if we assume that  $N=kn-1$  (otherwise we may 'pad'  $p(x)$  with extra 0 coefficients). Define  $n$  polynomials in  $x^n$ ,  $p_i(x^n)$  of degree

$$\lfloor \frac{N}{n} \rfloor = k-1 \quad (1.45)$$

thus:

$$p_0(x^n) = c_0 + c_n x^n + c_{2n} x^{2n} + \dots + c_{\lfloor \frac{N}{n} \rfloor n} x^{\lfloor \frac{N}{n} \rfloor n} \quad (1.46)$$

$$p_1(x^n) = c_1 + c_{n+1} x^n + c_{2n+1} x^{2n} + \dots + c_{\lfloor \frac{N}{n} \rfloor n + 1} x^{\lfloor \frac{N}{n} \rfloor n}$$

...

$$p_i(x^n) = c_i + c_{n+i} x^n + c_{2n+i} x^{2n} + \dots + c_{\lfloor \frac{N}{n} \rfloor n + i} x^{\lfloor \frac{N}{n} \rfloor n} \quad (i = 0, \dots, n-1)$$

...

...

Then  $p(x)$  may be expressed as

$$p(x) = p_0(x^n) + xp_1(x^n) + \dots + x^i p_i(x^n) + \dots + x^{n-1} p_{n-1}(x^n) \quad (1.47)$$

Note that the highest power of  $x$  here is  $n-1 + \lfloor \frac{N}{n} \rfloor n = n-1 + (k-1)n = kn-1 = N$ , as required.

Now the powers  $x, x^2, x^3, x^4, \dots, x^n$  may be computed in parallel as follows:

time step 1: compute  $x^2$

time step 2: multiply  $x, x^2$  by  $x^2$  in parallel. Now we have  $x, x^2, x^3, x^4$ .

time step 3: multiply  $x, x^2, x^3, x^4$  by  $x^4$  in parallel; thus we have all powers up to  $x^8$ .

Continue similarly until at step  $\lceil \log n \rceil$  we have all powers up to  $x^{2^{\lceil \log n \rceil}}$ , i.e. at least  $x^n$ . The maximum number of processors required is at the last step where we need  $\frac{n}{2}$  processors.

Next we compute  $p_i(x^n)$  for  $i=0,1,\dots,n-1$  in parallel with  $n$  processors, each one by Horner's rule in  $2\lfloor \frac{N}{n} \rfloor$  steps.

Finally, multiply each  $p_i(x^n)$  by  $x^i$  in 1 step, and add them by associate fan-in in  $\lceil \log n \rceil$  steps (and  $\frac{n}{2}$  processors).

Thus the total number of time steps are

$$T(N, n) = 2\lceil \log n \rceil + 2\lfloor \frac{N}{n} \rfloor + 1 \quad (1.48)$$

For  $n \geq N+1$ , this method reduces to finding  $x^j$  for  $j=1,2,\dots,N$ , multiplying  $c_j$  by  $x^j$  in 1 step, and adding the products in  $\lceil \log(N+1) \rceil$  steps, for a total of

$$T(N, N+1) = \lceil \log N \rceil + \lceil \log(N+1) \rceil + 1 \quad (1.49)$$

if we define  $T^*(N)$  as

$$\min_{1 \leq n \leq N+1} T(N, n) \quad (1.50)$$

then Lakshmivarahan and Dhall (1990) pp255-261 show that

$$T^*(N) = T(N, N+1) \quad (1.51)$$

They also show that the minimum number of processors  $n^*$  required to attain this minimum is

$$\begin{aligned} N+1 & \text{ if } N = 2^g \\ \lceil \frac{N+1}{3} \rceil & \text{ if } 2^g < N < 2^g + 2^{g-1} \\ \lceil \frac{N+1}{2} \rceil & \text{ if } 2^g + 2^{g-1} \leq N < 2^{g+1} \end{aligned} \quad (1.52)$$

where  $g = \lfloor \log N \rfloor$

Kiper (1997B) describes an elaboration of Dorn's method based on a decoupling algorithm of Kowalik and Kumar (1985). However, although they describe it as an improvement of Dorn's method, it appears to take slightly longer for the same number of processors.

Lakshmivarahan describes a "binary splitting" method due to Estrin (1960) which computes  $p(x)$  in  $2\lceil \log N \rceil$  time steps using  $\frac{N}{2} + 1$  processors. This is only slightly faster than optimum Dorn's, but sometimes uses fewer processors.

They also describe a "folding" method, due to Muraoka (1971-unpublished), which takes approximately  $1.44 \log N$  steps—significantly better than Dorn. It works as follows:

Let  $F_i$  be the  $i$ 'th Fibonacci number defined by

$$F_0 = F_1 = 1, F_i = F_{i-1} + F_{i-2} \quad (i \geq 2) \quad (1.53)$$

Let

$$p(x) = c_{F_{t+1}-1}x^{F_{t+1}-1} + c_{F_{t+1}-2}x^{F_{t+1}-2} + \dots + c_1x + c_0 \quad (1.54)$$

(if the degree of  $p(x)$  is not of the required form it may be padded with extra terms having 0 coefficients)

Now we may write

$$p(x) = p_1(x) \times x^{F_t} + p_2(x) \quad (1.55)$$

where  $p_2$  has degree  $F_t - 1$  and  $p_1$  degree  $F_{t-1} - 1$

In turn we write

$$p_1 = p_{11}x^{F_{t-2}} + p_{12} \quad (1.56)$$

where  $p_{11}$  has degree  $F_{t-3} - 1$  and  $p_{12}$  degree  $F_{t-2} - 1$ .

Similarly  $p_2 = p_{21}x^{F_{t-1}} + p_{22}$ ,

where  $p_{21}$  has degree  $F_{t-2} - 1$  and  $p_{22}$  degree  $F_{t-1} - 1$ , and the process is continued until we have to evaluate terms such as  $c_i x$ , which can be done in parallel, as well as evaluating powers of  $x$ . A building up process is then applied, whereby  $p(x)$  of degree  $N$  where  $F_t \leq N \leq F_{t+1}$  can be computed in  $t+1$  steps. Since  $F_t \approx \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^{t+1}$  we have  $\log_2 F_t \approx \log_2 \frac{1}{\sqrt{5}} + (t+1) \log_2 \left( \frac{1+\sqrt{5}}{2} \right)$ . Hence  $t+1 \approx 1.44 \log_2 F_t \leq 1.44 \log_2 N$ .

## 1.5 Evaluation at Many Points

This problem arises, for example, in Simultaneous Root-Finding methods (see Chap. 4). Probably the best method for large  $n$  is that given by Pan et al (1997), based on the Fast Fourier Transform. He assumes that the evaluation is to be done at  $n$  points  $\{x_0, x_1, \dots, x_{n-1}\}$ , but if we have more than  $n$  points we may repeat the process as often as needed. He assumes the polynomial  $p(x)$  is of degree  $n-1$ , with coefficient vector  $\mathbf{c} = [c_0, c_1, \dots, c_{n-1}]$ . Let the value of the polynomial at  $x_i$  be  $v_i$ .

We will interpolate  $p(u)$  at all the  $n$ 'th roots of unity given by

$$w_k = \exp(2\pi k\sqrt{-1}/n) \quad (k = 0, \dots, n-1) \quad (1.57)$$

$$i.e. \ p(u) = \sum_{k=0}^{n-1} \frac{p(w_k) \prod_{i=0, \neq k}^{n-1} (u - w_i)}{\prod_{i=0, \neq k}^{n-1} (w_k - w_i)} \quad (1.58)$$

$$= \Gamma(u) \sum_{k=0}^{n-1} \frac{p(w_k)}{\Gamma'(w_k)(u - w_k)} \quad (1.59)$$

where

$$\Gamma(u) = \prod_{i=0}^{n-1} (u - w_i) = u^n - 1 \quad (1.60)$$

$$\Gamma'(u) = nu^{n-1} \quad (1.61)$$

Hence

$$\Gamma(x_i) = x_i^n - 1, \quad \Gamma'(w_i) = nw_i^{n-1} = n/w_i \quad (1.62)$$

Putting  $u = x_i$  ( $i=0, \dots, n-1$ ) in 1.59 gives

$$v_i = p(x_i) = \Gamma(x_i) \sum_{k=0}^{n-1} \frac{1}{x_i - w_k} \frac{1}{\Gamma'(w_k)} (\sqrt{n} \mathbf{F}\mathbf{c})_k \quad (1.63)$$

where

$$\mathbf{F} = \frac{1}{\sqrt{n}} [w_k^j]_{j,k=0}^{n-1} \quad (1.64)$$

Hence

$$v_i = (x_i^n - 1) \sum_{k=0}^{n-1} \frac{1}{x_i - w_k} \frac{1}{n/w_k} (\sqrt{n} \mathbf{F}\mathbf{c})_k \quad (1.65)$$

$$= (1 - x_i^n) \sum_{k=0}^{n-1} \frac{1}{w_k - x_i} w_k \left( \frac{1}{\sqrt{n}} \mathbf{F}\mathbf{c} \right)_k \quad (1.66)$$



$$= \left(\frac{1}{x_i} - x_i^{n-1}\right) \sum_{k=0}^{n-1} \frac{1}{\frac{1}{x_i} - \frac{1}{w_k}} u_k \quad (1.67)$$

$$\text{where } u_k = \left(\frac{1}{\sqrt{n}} \mathbf{F}\mathbf{c}\right)_k \quad (1.68)$$

$$\text{But } \frac{1}{\frac{1}{x_i} - \frac{1}{w_k}} = x_i \left(1 - \frac{x_i}{w_k}\right)^{-1} = x_i \sum_{j=0}^{\infty} \left(\frac{x_i}{w_k}\right)^j \quad (1.69)$$

$$\text{and so } v_i = (1 - x_i^n) \sum_{j=0}^{\infty} A_j x_i^j \quad (1.70)$$

$$\text{where } A_j = \sum_{k=0}^{n-1} \frac{u_k}{w_k^j} \quad (1.71)$$

Now suppose

$$1 > q \geq \max_k |x_k| \quad (1.72)$$

(later we will see that this can be arranged),

$$\text{and } \alpha = \max_k |u_k| \quad (1.73)$$

and note that  $|w_k| = 1$  all  $k$ , so that

$$|A_j| \leq \alpha n \quad (1.74)$$

Hence if we approximate  $v_i$  by

$$v_i^* = (1 - x_i^n) \sum_{j=0}^{L-1} A_j x_i^j \quad (1.75)$$

the error

$$E_L = \|\mathbf{v}^* - \mathbf{v}\| = \max_i |v_i^* - v_i| \leq \frac{\alpha n b q^L}{1 - q} \quad (1.76)$$

$$\text{where } b = \max_k |x_k^n - 1| \leq 1 + q^n \quad (1.77)$$

Now  $E_L \leq \text{some given } \epsilon$  if  $\left(\frac{1}{q}\right)^L \geq \frac{\alpha n b}{(1 - q)\epsilon}$

$$\text{i.e. if } L \geq \lceil \log\left(\frac{\alpha n b}{(1 - q)\epsilon}\right) / \log\left(\frac{1}{q}\right) \rceil \quad (1.78)$$

Evaluation of  $\mathbf{F}\mathbf{c}$  and hence  $\mathbf{u} = [u_0, \dots, u_{n-1}]$  requires  $O(n \log n)$  operations; while a single  $x_i^n$  can be evaluated by repeated squaring in  $\log n$  operations, so  $x_i^n$  and

$1 - x_i^n$  for all  $i$  can be done in  $O(n \log n)$  operations.  $A_j$  for  $j=0, \dots, L-1$  requires  $L(2n-2)$  operations, and finally  $v_i^*$  for all  $i$  need  $(1+2L)n$  operations. Thus the total number is

$$L(4n - 2) + O(n \log n) + n \quad (1.79)$$

Now the numerator of the right-hand-side of 1.78 can be written  $\log(\frac{\alpha}{\epsilon}) + \log n + \log b - \log(1 - q)$ . It often happens that  $\log(\frac{\alpha}{\epsilon}) = O(\log n)$ , so that  $L = O(\log n)$ , and the number of operations is  $O(n \log n)$  (N.B.  $b$  and  $q$  are fixed constants).

However all the above depends on 1.72, which is often not satisfied. In that case we may partition  $X = \{x_0, \dots, x_{n-1}\}$  into 3 subsets:  $X_-$ ,  $X_0$ ,  $X_+$ , where  $|x_i| < 1$  for  $X_-$ ,  $= 1$  for  $X_0$ , and  $> 1$  for  $X_+$ .  $X_-$  presents no problem (but see below), while for  $X_+$  we have  $\frac{1}{|x_i|} < 1$ , so we apply our process to the reverse polynomial  $q(x) = x^n p(\frac{1}{x})$ .

For  $X_0$ , apart from the trivial cases  $x = \pm 1$ , we have  $-1 < R = Re(x) < 1, I = Im(x) = \pm \sqrt{1 - R^2}$ . Thus we may rewrite  $p(x)$  as  $p_0(R) + Ip_1(R)$  for  $|R| < 1$ .

For example, consider a quadratic  $p(x) = c_0 + c_1(R + iI) + c_2(R + iI)^2$

$$= c_0 + c_1R + c_2(R^2 - I^2) + i(c_1I + 2c_2RI)$$

$$= c_0 + c_1R + c_2(2R^2 - 1) + iI(c_1 + 2c_2R),$$

This takes the stated form with  $p_0 = c_0 - c_2 + c_1R + 2c_2R^2$ ,  $p_1 = ic_1 + 2ic_2R$ .

Despite the 'trick' used above, we may still have a problem if  $q$  is very close to 1, for then  $L$  will need to be very large to satisfy 1.78, i.e. it may be larger than  $O(\log n)$ . We may avoid this problem by using the transformation

$$x = \gamma y + \delta, \text{ or } y = \frac{(x - \delta)}{\gamma} \quad (1.80)$$

where  $\delta$  is the centroid of the  $x_i$ ,

$$= \frac{\sum_{i=0}^{n-1} x_i}{n} \quad (1.81)$$

$$\text{and } \gamma = (e.g.) 1.2 \text{Max}|x_i - \delta| \quad (1.82)$$

$$\text{Then } \text{max}_i |y_i| < .833 = q \quad (1.83)$$

1.80 may be executed in two stages; first  $x = z + \delta$ , which requires  $O(n \log n)$  operations using the method of Aho et al referred to in section 3. Then we let  $z = \gamma y$ , leading to a new polynomial whose  $i$ 'th coefficient is  $\gamma^i$  times the old one. Since the  $\gamma^i$  can be evaluated in  $n$  operations, and also multiplied by the old coefficients in  $n$  operations, the overall time complexity for the transformation is  $O(n \log n)$ . So the entire multipoint evaluation will be of that order.

## 1.6 Evaluation at Many Equidistant Points

This problem is quite common, for example in signal processing. Nuttall (1987) and Dutta Roy and Minocha (1991) describe a method that solves this problem in  $nm$  additions and no multiplications, where  $n$  = degree and  $m$  = number of evaluation points. That is, apart from initialization which takes  $O(n^3)$  operations. The method compares favourably in efficiency with the repeated Horner method for small  $n$  and moderate  $m$ . For example, for  $n = 3, 4, 5$  Nuttall's method is best for  $m > 12, 17$ , and  $24$  respectively.

The polynomial

$$p_n(x) = \sum_{j=0}^n c_j x^j \quad (1.84)$$

is to be evaluated at equidistant points

$$x_s = x_0 + s\Delta \quad (s = 0, 1, 2, \dots, m) \quad (1.85)$$

Combining 1.84 and 1.85 gives

$$\begin{aligned} p_n(x_s) &= Q_n(s) = \sum_{j=0}^n c_j (x_0 + s\Delta)^j = \sum_{j=0}^n c_j \sum_{k=0}^j x_0^{j-k} \binom{j}{k} \Delta^k s^k \\ &= \sum_{k=0}^n \left( \sum_{j=k}^n \binom{j}{k} c_j x_0^{j-k} \right) \Delta^k s^k \\ &= \sum_{k=0}^n a_k s^k \end{aligned} \quad (1.86)$$

where

$$a_k = \Delta^k \sum_{j=k}^n \binom{j}{k} c_j x_0^{j-k} \quad (k = 0, 1, \dots, n) \quad (1.87)$$

Now we define the backward differences

$$Q_k(s) = Q_{k+1}(s) - Q_{k+1}(s-1) \quad (k = n-1, n-2, \dots, 1, 0) \quad (1.88)$$

We will need initial values  $Q_k(0)$ ; these can be obtained as follows:-

by 1.88

$$Q_{n-1}(s) = Q_n(s) - Q_n(s-1)$$

$$Q_{n-2}(s) = Q_{n-1}(s) - Q_{n-1}(s-1) = Q_n(s) - Q_n(s-1) - [Q_n(s-1) - Q_n(s-2)] =$$

$Q_n(s) - 2Q_n(s-1) + Q_n(s-2)$ , and so on, so that in general (by induction)

$$Q_{n-r}(s) = \sum_{i=0}^r (-1)^i \binom{r}{i} Q_n(s-i) \quad (r = 1, 2, \dots, n) \quad (1.89)$$

Putting  $s = 0$  above gives

$$Q_{n-r}(0) = \sum_{i=0}^r (-1)^i \binom{r}{i} Q_n(-i) \quad (1.90)$$

Also putting  $s = -i$  in 1.86 gives

$$Q_n(-i) = \sum_{k=0}^n a_k (-i)^k \quad (1.91)$$

Hence

$$\begin{aligned} Q_{n-r}(0) &= \sum_{i=0}^r \sum_{k=0}^n (-1)^{i+k} \binom{r}{i} (i)^k a_k = \\ &= \sum_{k=0}^n \left[ \sum_{i=1}^r (-1)^{i+k} (i)^k \binom{r}{i} \right] a_k \end{aligned} \quad (1.92)$$

since  $i = 0$  gives  $(i)^k = 0$ .

However, combining 1.88 for  $k = n-1$  and 1.86, we have

$$\begin{aligned} Q_{n-1}(s) &= [a_0 + \sum_{k=1}^n a_k s^k] - [a_0 + \sum_{k=1}^n a_k (s-1)^k] = \\ &= \sum_{k=1}^n a_k [s^k - (s-1)^k] = \\ &= a_1 + \sum_{k=2}^n a_k (k s^{k-1} + \dots) \end{aligned}$$

i.e.  $Q_{n-1}(s)$  contains no term in  $a_0$ .

Similarly  $Q_{n-2} = [a_1 + \sum_{k=2}^n b_k s^{k-1}] - [a_1 + \sum_{k=2}^n b_k (s-1)^{k-1}]$

where the  $b_k$  are functions of  $a_2, \dots, a_n$ , i.e.  $Q_{n-2}$  contains no term in  $a_1$  or  $a_0$ . Hence by induction we may show that  $Q_{n-r}$  contains no terms in  $a_0, a_1, \dots, a_{r-1}$ , and that it is of degree  $n-r$  in  $s$ .

Thus 1.92 may be replaced by

$$Q_{n-r}(0) = \sum_{k=r}^n \left[ \sum_{i=1}^r (-1)^{i+k} (i)^k \binom{r}{i} \right] a_k \quad (1.93)$$

Also from 1.86

$$Q_n(0) = a_0 \quad (1.94)$$

and, since  $Q_0$  is a polynomial of degree 0 in  $s$ , (and by the above inductive proof)

$$Q_0(s) = n!a_n, \text{ for all } s \quad (1.95)$$

Finally 1.88 may be re-arranged to give the recursion

$$Q_{k+1}(s) = Q_{k+1}(s-1) + Q_k(s) \quad (k = 0, 1, \dots, n-1) \quad (1.96)$$

whereby we may obtain in turn  $Q_n(1), Q_n(2), \dots$  at a cost of  $n$  additions per sample point.

Volk (1988) shows empirically that the above method is unstable for large  $m$  (number of evaluation points). For this reason, as well as for efficiency considerations, it is probably best to use the method of Pan (section 5) in the case of large  $m$ .

It would be a useful research project to determine where the break-even point is.

**AN EXAMPLE** Let  $p_3(x) = 1 + 2x + 3x^2 + 4x^3$ , i.e.  $c_0 = 1, c_1 = 2, c_2 = 3, c_3 = 4$ , and let us evaluate it at 2,4,6,8, i.e. with  $x_0 = 2$  and  $\Delta = 2$ .

$$\begin{aligned} \text{Then in 1.87, } a_0 &= 2^0 \sum_{j=0}^3 \binom{j}{0} c_j 2^j = 2^0(1 \times 2^0 + 2 \times 2^1 + 3 \times 2^2 + 4 \times 2^3) \\ &= 1(1 + 4 + 12 + 32) = 49 \end{aligned}$$

$$\begin{aligned} a_1 &= 2^1 \sum_{j=1}^3 \binom{j}{1} c_j 2^{j-1} = 2(2 \times 2^0 + 2 \times 3 \times 2^1 + 3 \times 4 \times 2^2) \\ &= 2(2 + 12 + 48) = 2 \times 62 = 124 \end{aligned}$$

$$a_2 = 2^2 \sum_{j=2}^3 \binom{j}{2} c_j 2^{j-2} = 4(3 \times 2^0 + 3 \times 4 \times 2^1) = 4(3 + 24) = 4 \times 27 = 108$$

$$a_3 = 2^3 \times \binom{3}{3} c_3 2^0 = 8 \times 4 = 32$$

$$\text{i.e. } Q_3(s) = 49 + 124s + 108s^2 + 32s^3$$

**check** by direct method

$$\begin{aligned} p_3(x_0 + \Delta s) &= p_3(2 + 2s) = 1 + 2(2 + 2s) + 3(2 + 2s)^2 + 4(2 + 2s)^3 = \\ &= 1 + 4 + 4s + 3(4 + 8s + 4s^2) + 4(8 + 24s + 24s^2 + 8s^3) = \\ &= 49 + 124s + 108s^2 + 32s^3 \quad (\text{agrees with above}) \end{aligned}$$

Next we use 1.93 to give

$$Q_2(0) = \sum_{k=1}^3 [\sum_{i=1}^1 (-1)^{i+k} (i)^k \binom{1}{i}] a_k =$$

$$\sum_{k=1}^3 (-1)^{1+k} a_k = a_1 - a_2 + a_3 = 124 - 108 + 32 = 48$$

$$\text{check } Q_2(0) = Q_3(0) - Q_3(-1) = 49 - (49 - 124 + 108 - 32) = 48$$

$$Q_1(0) = \sum_{k=2}^3 [\sum_{i=1}^2 (-1)^{i+k} (i)^k \binom{2}{i}] a_k = \sum_{k=2}^3 [(-1)^{1+k} 2 + (-1)^{2+k} 2^k] a_k =$$

$$[-2 + 4]a_2 + [2 - 8]a_3 = 2a_2 - 6a_3 = 2 \times 108 - 6 \times 32 = 216 - 192 = 24$$

Also  $Q_3(0) = a_0 = 49$ , while for all  $s$ ,  $Q_0(s) = 3! \times 32 = 192$

Finally 1.96 gives, for  $s = 1$ ,

$$Q_1(1) = Q_1(0) + Q_0(1) = 24 + 192 = 216$$

$$Q_2(1) = Q_2(0) + Q_1(1) = 48 + 216 = 264$$

$$Q_3(1) = Q_3(0) + Q_2(1) = 49 + 264 = 313$$

$$\text{check } p_3(2 + 2 \times 1) = p_3(4) = 1 + 2 \times 4 + 3 \times 4^2 + 4 \times 4^3 =$$

$$1 + 8 + 48 + 256 = 313 \text{ (agrees).}$$

while for  $s = 2$

$$Q_1(2) = Q_1(1) + Q_0(2) = 216 + 192 = 408$$

$$Q_2(2) = Q_2(1) + Q_1(2) = 264 + 408 = 672$$

$$Q_3(2) = Q_3(1) + Q_2(2) = 313 + 672 = 985$$

$$\text{check } p_3(2 + 2 \times 2) = p_3(6) = 1 + 2 \times 6 + 3 \times 6^2 + 4 \times 6^3 =$$

$$1 + 12 + 108 + 864 = 985 \text{ (agrees).}$$

## 1.7 Accurate Evaluation

In implementing iterative methods we usually need (at some point) to evaluate  $p(x_i)$  where  $x_i$  is close to a root. In that case the rounding error in the evaluation may be bigger than the actual value, so that the latter cannot be calculated, at least not in normal floating-point arithmetic. A popular solution to this problem is to utilize multi-precision arithmetic, but this is quite expensive.

Paquet (1994) describes a method which can evaluate  $p(x_i)$  correct to machine accuracy, while using only 'normal' floating-point arithmetic (hereafter referred to as 'float' arithmetic). He assumes that the underlying float operations are optimal, i.e.

$$x \hat{\circ} y = \text{round}(x \circ y) \tag{1.97}$$

i.e. the computed result of an operation  $\circ = (+, -, *, \text{ or } /)$  equals the result of rounding the exact value. Let  $u = \frac{1}{2}\beta^{1-t}$ ,  $\omega = 1 + u$ . then

$$|x \hat{\circ} y - x \circ y| \leq \frac{u}{\omega} |x \circ y| \tag{1.98}$$

In an example, using his exact evaluation method on a 19 decimal digit machine he gets a result correct to machine accuracy, whereas ordinary float arithmetic gave only 11 digits correct.

Paquet's precise method depends on having available a precise scalar product. For this he recommends the method of Dekker (1971) to obtain an exact product

of two float numbers as a sum of two float numbers. He also recommends a method of Pichat (1972) or a similar method due to Kahan (1965) to sum a set of numbers exactly. However this author believes that some of Dekker's procedures would be adequate for this purpose. Dekker's techniques will be described towards the end of this section.

Paquet's method allows for evaluation at a point given in 'staggered correction format', i.e. where  $x_i$  is a sum of several float numbers (usually each one much smaller than the previous). The result will be rounded to the nearest float number, although theoretically it could also be given in staggered format. The method will be explained assuming the staggered format for  $x_i$ , although most commonly  $x_i$  will consist of only one float number.

We wish to evaluate  $p(x)$  at the point

$$\tau = \sum_{s=0}^r t^{(s)} \quad (1.99)$$

where  $t^{(0)}, t^{(1)}, \dots, t^{(s)}$  are float numbers. Set

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \dots & & & 0 \\ -\tau & 1 & 0 & \dots & & 0 \\ 0 & -\tau & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -\tau & 1 \end{bmatrix}, \quad (1.100)$$

$$\mathbf{A}_0 = \begin{bmatrix} 1 & 0 & \dots & & & 0 \\ -t^{(0)} & 1 & 0 & \dots & & 0 \\ 0 & -t^{(0)} & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & & -t^{(0)} & 1 \end{bmatrix} \quad (1.101)$$

$$\mathbf{p} = \begin{bmatrix} c_n \\ c_{n-1} \\ \dots \\ \dots \\ c_0 \end{bmatrix} \quad (1.102)$$

The elements of  $\mathbf{A}_0$  are float numbers, but generally those of  $\mathbf{A}$  are not.

In **exact** arithmetic the solution of  $\mathbf{Ax} = \mathbf{p}$  (or  $\mathbf{A}_0\mathbf{x} = \mathbf{p}$ ) is equivalent to the Horner scheme 1.2 with  $x$  replaced by  $\tau$  (or  $t^{(0)}$ ). Paquet shows how we can obtain, in float arithmetic, a series  $\sum_{l=0}^{\infty} x_0^l$  of float numbers convergent to the value of  $p(\tau)$ .

The algorithm follows:

(i) Initialize

$$\mathbf{r}^{(0)} = \mathbf{p} = [c_n, c_{n-1}, \dots, c_0]^T \quad (1.103)$$

(ii) For  $j = 0, 1, 2, \dots$

(a) Solve

$$\mathbf{A}_0 \mathbf{y} = \mathbf{r}^{(j)} = [r_n^{(j)}, \dots, r_0^{(j)}]^T \quad (1.104)$$

in float arithmetic, by forward substitution, to give

$$\mathbf{x}^{(j)} = [x_n^{(j)}, \dots, x_0^{(j)}]^T, \text{ i.e.}$$

$$x_n^{(j)} = r_n^{(j)} \quad (1.105)$$

and

$$x_i^{(j)} = r_i^{(j)} \hat{+} (t^{(0)} \hat{*} x_{i+1}^{(j)}) \quad (i = n-1, n-2, \dots, 0) \quad (1.106)$$

(b) Compute the residual

$$\mathbf{r}^{(j+1)} = \text{round}(\mathbf{p} - \mathbf{A} \sum_{l=0}^j \mathbf{x}^{(l)}) \quad (1.107)$$

or in more detail

$$r_i^{(j+1)} = \text{round}(c_i - \sum_{l=0}^j x_i^{(l)} + \sum_{s=0}^r t^{(s)} \sum_{l=0}^j x_{i+1}^{(l)}) \quad (i = n, n-1, \dots, 0) \quad (1.108)$$

where we have set

$$x_{n+1}^{(0)} = \dots = x_{n+1}^{(j)} = 0 \quad (1.109)$$

The above 1.108 needs a precise scalar product, namely that between the vectors

$$[1, -1, -1, \dots, -1, t^{(0)}, t^{(1)}, \dots, t^{(r)}, t^{(0)}, t^{(1)}, \dots, t^{(r)}, \dots, t^{(0)}, t^{(1)} \dots t^{(r)}] \quad (1.110)$$

and

$$[c_i, x_i^{(0)}, x_i^{(1)}, \dots, x_i^{(j)}, x_{i+1}^{(0)}, \dots, x_{i+1}^{(0)}, x_{i+1}^{(1)}, \dots, x_{i+1}^{(1)}, \dots, x_{i+1}^{(j)}, \dots, x_{i+1}^{(j)}] \quad (1.111)$$

Paquet proves that provided

$$|\tau| < 1, |t^{(0)}| < 1 \quad (1.112)$$

and

$$q = 2(n^2 + 2n)u\omega^{2n-1} + 2(n+1)\frac{u}{\omega} + \omega n(n+1)|\tau - t^{(0)}| < 1 \quad (1.113)$$



and with  $\mathbf{x}$  the exact solution of  $\mathbf{Ax} = \mathbf{p}$ . then

$$\left\| \mathbf{x} - \sum_{l=0}^j \mathbf{x}^{(l)} \right\|_{\infty} \leq q \left\| \mathbf{x} - \sum_{l=0}^{j-1} \mathbf{x}^{(l)} \right\|_{\infty} \quad (1.114)$$

leading to

$$\left| x_0 - \sum_{l=0}^j x_0^{(l)} \right| \leq q^j \left\| \mathbf{x} - \mathbf{x}^{(0)} \right\|_{\infty} \quad (1.115)$$

where

$$x_0 = p(\tau) \quad (1.116)$$

and  $\mathbf{x}^{(0)}$  is the float-number solution of

$$\mathbf{A}_0 \mathbf{x}^{(0)} = \mathbf{p} \quad (1.117)$$

Accordingly

$$\sum_{l=0}^j x_0^{(l)} \rightarrow p(\tau) \text{ as } j \rightarrow \infty \quad (1.118)$$

Moreover he shows that by the change of variables

$$t^{(s)} = \beta^a \xi^{(s)} \quad (1.119)$$

(a an integer) we may drop the conditions 1.112 and the last term in the middle expression in 1.113.

In this context Paquet gives more details of the previously mentioned example: if  $\tau = a$  a simple float number =  $t^{(0)}$  = root of the equation correct to 19 decimal places (machine accuracy), then  $\sum_{l=0}^j x_0^{(l)}$  converges to machine accuracy in 3 iterations ( $j = 2$ ). The individual  $x_0^{(l)}$  for  $l = 0, 1, 2, 3, 4$  are approximately (in base 16)  $+8 \times 16^{-14}$ ,  $-8 \times 16^{-14}$ ,  $+3 \times 16^{-30}$ ,  $-3.7 \times 16^{-46}$ , and  $-A.A \times 16^{-63}$ .

Dekker's method of multiplying two 'single-length' float numbers  $x$  and  $y$  gives a pair  $(z, zz)$  of float numbers such that

$$z + zz = x \times y \quad (1.120)$$

where  $zz$  is almost negligible within machine precision compared to  $x \times y$ , i.e.

$$|zz| \leq |z + zz| \frac{2^{-t}}{1 + 2^{-t}} \quad (1.121)$$

He refers to the pair  $(z, zz)$  as a 'double-length' number. In practise  $x$ ,  $y$ ,  $z$ , and  $zz$  may be double precision numbers in the usual sense, so that  $(z, zz)$  is really a

quadruple-precision number. But we will refer to  $x$ ,  $y$ ,  $z$ ,  $zz$ , and other intermediate variables as ‘single-length’ for the present discussion. His algorithm for multiplication is as follows (in Algol 60):

```

procedure mul12(x,y,z,zz);
value x,y; real x,y,z,zz;
begin real hx,tx,hy,ty,p,q;
p:= x × constant;
comment constant = 2 ↑ (t - t ÷ 2) + 1
hx:= (x-p)+p; tx:= x-hx;
p:= y × constant;
hy:= (y-p)+p; ty:= y-hy;
p:= hx × hy; q:= hx × ty + tx × hy;
z:= p+q; zz:= (p-z)+q+tx × ty;
end mul12;

```

He also gives an algorithm for adding two ‘double-length’ (in his special sense) numbers  $(x,xx)$  and  $(y,yy)$ . It follows:

```

comment add2 calculates the double-length sum of  $(x,xx)$  and  $(y,yy)$ , the result
being  $(z,zz)$ ;
procedure add2(x,xx,y,yy,z,zz);
value x,xx,y,yy; real x, xx, y, yy, z, zz;
begin real r,s;
r:= x+y;
s:= if abs(x) > abs(y) then
((x-r)+y)+yy+xx else ((y-r)+x)+xx+yy;
z:=r+s; zz:=(r-z)+s;
end add2;

```

Dekker gives a detailed proof of the accuracy of these procedures, and remarks that they ‘have been used extensively for calculating double-length scalar products of single-length scalar products of vectors of single-length float numbers’. If, as he implies, these tests have been successful, it would seem that the methods of Pichat etc mentioned by Paquet are not really needed.

Hammer et al (1995) and Kulisch and Miranker (1983) describe methods very similar to Paquet’s, except that they use special hardware for the accurate scalar product instead of Dekker’s float method. This special hardware may be generally inaccessible.

## 1.8 Scaling

A common problem in evaluating polynomials in float arithmetic is overflow or underflow. These phenomena, especially the former, may result in highly inaccurate values. Linnainmaa (1981) gives several examples where both of these effects cause problems. One solution is scaling; i.e. we may multiply all the coefficients, and/or the argument, by a scaling factor so that overflow is prevented and the incidence of underflow reduced. One work in which the coefficients are scaled is by Hansen et al (1990).

They assume that a number is represented by  $m\beta^c$  where  $c \in [a, b]$  and  $a, b$ , are respectively negative and positive integers. Also as usual the numbers are normalized, i.e.

$$\frac{1}{\beta} \leq |m| < 1 \quad (1.122)$$

Let

$$\tilde{a} = \text{floor}[(a+1)/2], \tilde{b} = \text{floor}[b/2] \quad (1.123)$$

and

$$\gamma = \min\{-\tilde{a}, \tilde{b}\}, \bar{I} = [-\gamma, \gamma] \quad (1.124)$$

Then  $z_1 z_2$  can be computed without under- or overflow provided

$$c(z_i) \in \bar{I} \quad (i = 1, 2) \quad (1.125)$$

The same is true for  $z_1 + z_2$

The polynomial will be evaluated by Horner's rule

$$f_n = c_n, f_k = x f_{k+1} + c_k \quad (k = n-1, \dots, 0) \quad (1.126)$$

Assume  $c(x) \in \bar{I}$ , and initially scale  $f_n = c_n$  so that  $c(f_n) \in \bar{I}$ , recording the scale factor. Now, as an inductive hypothesis, assume that  $c(f_{k+1}) \in \bar{I}$ . Since also  $c(x) \in \bar{I}$ ,  $x f_{k+1}$  can be computed without under- or overflow. Next we will have to add  $x f_{k+1}$  to  $c_k$ . We scale  $x f_{k+1}$  and  $c_k$  so that the larger of  $|x f_{k+1}|$  and  $|c_k|$  has exponent in  $\bar{I}$  (note that scaling  $x f_{k+1}$  implicitly scales  $c_n, c_{n-1}, \dots, c_{k+1}$ ). This may cause the smaller to underflow, but according to Hansen this is only a rounding error.

The scaling is done by multiplying by  $s = \beta^r$  ( $r$  an integer, possibly negative). Thus the mantissa of the scaled number is not disturbed, so there will be no rounding error from this cause. We choose  $r$  to minimize underflow, i.e.

$$r = \gamma - c(\max\{|x f_{k+1}|, |c_k|\}) \quad (1.127)$$

Thus the larger of  $|xf_{k+1}|$  and  $|c_k|$  will have exponent  $\gamma$ .

Suppose we have done a sequence of scalings using  $s_i = \beta^{r_i}$ , for  $i = 1, 2, \dots, M$ . Then the resulting scaling for all the coefficients at this stage is the same as if we had scaled only once using

$$r = \sum_{i=1}^M r_i \quad (1.128)$$

We record  $r$ . We need not scale any particular coefficient until it is used in 1.126.

Hansen gives a detailed pseudo-code for the method and summarizes briefly a FORTRAN 77 implementation. The output of that is given in the form of a real number  $F$  and integer  $SCALE$  such that  $p(x) = F * \beta^{SCALE}$ . Also the input and intermediate data are converted to that form. This allows for a very wide range of values of  $x$  and coefficients to be accommodated without over- or underflow problems.

Linnainmaa describes a very similar method, except that for  $|x| > 1$  he makes the transformation

$$P(x) = x^n Q\left(\frac{1}{x}\right) \quad (1.129)$$

where

$$Q(z) = c_0 z^n + c_1 z^{n-1} + \dots + c_{n-1} z + c_n \quad (1.130)$$

Then

$$\frac{P}{P'} = \frac{x}{n - \frac{Q'(\frac{1}{x})}{xQ(\frac{1}{x})}} \quad (1.131)$$

which lends itself to use in Newton's or similar methods. His algorithm scales  $P'$  as well as  $P$ . As mentioned earlier he gives numerous examples where under- or over- flow cause serious errors in the standard Horner method, but where correct results are obtained with his modification.

## 1.9 Order of Convergence and Efficiency

Except for low-degree polynomials (up to degree 4) nearly all methods for finding roots involve iteration. That is, we make a guess  $x_0$  for one of the roots  $\alpha$  (or a series of guesses for all the roots), improve it by applying some "iteration function"  $\phi(x_0, \dots)$  to give  $x_1$  (hopefully closer to  $\alpha$ ), and so on, so that in general

$$x_{i+1} = \phi(x_i, \dots) \quad (1.132)$$

Under suitable conditions the iterations will converge towards  $\alpha$ .

Traub (1971) classifies various types of iteration function. In the simplest, one-point iteration (without memory),  $\phi$  is a function only of  $x_i, f(x_i), f'(x_i), \dots, f^{(s-1)}(x_i)$ . An example is Newton's method,

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (1.133)$$

The main part of the work involved in an iteration (of whatever type) is the evaluation of  $f(x_i)$  and its derivatives.

Another type of iteration is one-point with memory. Here we **re**-use old values of  $x_{i-1}, \dots, x_{i-m}, f(x_{i-1}), \dots, f(x_{i-m}), \dots, f^{(s-1)}(x_{i-1}), \dots, f^{(s-1)}(x_{i-m})$ , which have previously been calculated. The work is the same as for one-point without memory, i.e. the cost of the **new** evaluations (the cost of combining the various values of  $x, f, \dots, f^{(s-1)}$  is usually relatively small). An example is the Secant method

$$x_{i+1} = x_i - f(x_i) \frac{(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})} \quad (1.134)$$

A further class of iteration function is multipoint (with or without memory). For the case without memory, it proceeds thus:

$$z_1 = \phi_1(x_i, f(x_i), \dots, f^{(s-1)}(x_i)) \quad (1.135)$$

$$z_2 = \phi_2(x_i, f(x_i), \dots, f^{(s-1)}(x_i), z_1, f(z_1), \dots, f^{(s-1)}(z_1)) \quad (1.136)$$

·  
·  
·

$$z_j = \phi_j(x_i, f(x_i), \dots, f^{(s-1)}(x_i), z_1, f(z_1), \dots, f^{(s-1)}(z_1), \dots, z_{j-1}, f(z_{j-1}), \dots, f^{(s-1)}(z_{j-1})) \quad (j = 3, \dots, n) \quad (1.137)$$

$$x_{i+1} = z_n \quad (1.138)$$

The case with memory would also use old values

$$x_{i-l}, f(x_{i-l}), \dots, f^{(s-1)}(x_{i-l}) \quad (l = 1, \dots, m) \quad (1.139)$$

We would like to have a measure of how fast a method converges. Such a measure is given by the **order**  $p$ , defined by

$$\text{Lim}_{x_i \rightarrow \alpha} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|^p} = C \neq 0, \infty \quad (1.140)$$

Traub shows that one-point iterations without memory, using  $s-1$  derivatives, are of order at most  $s$ . Even with memory the order is at most  $s+1$ .

Werschultz (1981) describes a class of multipoint methods with memory using Hermite information, i.e. we compute  $x_{i+1}$  from  $x_i$  using

$$f^{(j)}(z_{i-s,l}), 0 \leq j \leq r_l - 1, 1 \leq l \leq k, 0 \leq s \leq m \quad (1.141)$$

The number of new function evaluations per iteration is

$$n = \sum_{l=1}^k r_l \quad (1.142)$$

and the memory is  $m$ .  $z_{i,l+1}$  depends on  $z_{i,q}$ ,  $f(z_{i,q})$ ,  $f^{(j)}(z_{i,q})$ ,  $j = 1, \dots, r_l - 1$ ,  $q = 1, \dots, l$ ,  $z_{i-s,q}$ ,  $f(z_{i-s,q})$ ,  $f^{(j)}(z_{i-s,q})$   $q = 1, \dots, k$ ,  $s = 1, \dots, m$ ,  $j = 1, \dots, r_l - 1$ . if  $k = 1$ , we have one-point iterations with memory, or if  $m = 0$  we have multipoint methods without memory. Werschultz shows that the order of such methods is bounded by  $2^n$ . This bound is attained (very nearly) for  $k = n$ ,  $r_1 = \dots = r_n = 1$  (i.e. no derivatives used) and  $m$  moderately large, by a Hermite Interpolatory method which he describes.

For another example he takes  $k = 1$ ,  $r_1 = s$ . He shows that the order  $p$  is the unique positive root of

$$p^{m+1} - s \sum_{i=0}^m p^i = 0 \quad (1.143)$$

e.g. the secant method has  $m = s = 1$ , giving

$$p^2 - p - 1 = 0 \quad (1.144)$$

i.e.  $p = \frac{1+\sqrt{5}}{2} = 1.618$ .

Another important consideration, which enables us to compare different methods, is efficiency, which is defined as the inverse of total work needed to reduce the error by a specified amount. It can be shown that this is

$$E = \log p^{\frac{1}{n}} = \frac{\log p}{n} \quad (1.145)$$

(apart from a constant factor independent of the method) where  $p$  is order and  $n$  is the number of new function evaluations (including derivatives) per iteration. The base of the log is arbitrary; this author uses base 10 but some authors use  $e$ . For example, for Newton's method  $p = n = 2$ , so  $E = \frac{1}{2} \log 2 = .1505$ .; for the secant method  $n = 1$ ,  $p = 1.618$ , so  $E = .2090$ .

According to 1.145 and Werschultz' bound, the maximum efficiency of multi-point methods with memory is

$$\log_{10}2 = .3010 \tag{1.146}$$

This author is not aware of any method exceeding this efficiency.

## 1.10 A Priori Bounds on (Real or Complex) Roots

Many methods for finding roots of polynomials (e.g. Sturm sequence methods) start with an estimate of an upper bound on the largest absolute value of the (real or complex) roots. If one can obtain a more accurate estimate for the bound, one can reduce the amount of work used in searching within the range of possible values (e.g. using a one- or two-dimensional bisection method). Thus it would be useful to know which of the available formulas is most accurate.

McNamee and Olhovsky (2005) report that, using the bibliography by McNamee (2002), they have located over 50 articles or books which give bounds on polynomial roots. They rejected those which were concerned only with real roots, or gave formulas which appeared too complicated (and hence might take too much time to compute). Also a few were rejected which worked only for special cases. They were left with 45 distinct formulas, which are listed in the appendix to this section.

The authors wrote a Java program to compare these formulas as follows: for each degree from 3 to 10 inclusive, 10 polynomials were generated with random real roots in the range (-1,+1) (and another set in the range -10,+10), and their coefficients computed. The 45 formulas (in terms of the coefficients) were applied, and a ratio determined in each case between the estimated bound and the actual maximum-magnitude root. These ratios were averaged over the 80 distinct polynomials, and the results output, along with a note as to which method gave the minimum ratio. The program was run 10 times for each set, giving slightly different results each time, as expected. The above process was repeated for complex roots with degrees 4,6,8,10. Thus we have 4 sets of polynomials (2 sets of 800 for real roots and 2 sets of 400 for complex)

Let the polynomial be

$$P(z) = z^n + c_{n-1}z^{n-1} + \dots + c_1z + c_0$$

and

$$\zeta = \text{Max}_{i=1,\dots,n} |\zeta_i|$$

where the  $\zeta_i$  are the roots of  $P(z) = 0$ . The best result (i.e. minimum ratio), by a margin of 20%-50%, was for the formula 1.147 below, due to Kalantari (2004,2005). The average ratio for this formula over the 2400 polynomials was close to 2.0, with a standard deviation of about .08. Then there were a set of 3 formulas due to Deutsch (1970) which gave ratios in the range 2.5-3.5. There were several other formulas which gave relatively low ratios for some sets of tests, but not for others. These will not be mentioned further.

Although the Deutsch formulas give bounds greater than those of Kalantari by about 30%, in many cases it may be preferable to use them instead, for they require considerably less work than Kalantari's formula (equivalent of about 1 function evaluation compared to at least 4 for Kalantari). In fact Kalantari gives a series of formulas which he believes approach closer and closer to the true bound, but at the price of increased work. The simplest of the Deutsch formulas is given by 1.148 below.

#### KALANTARI'S FORMULA

$$|\zeta| \leq \frac{1}{.682328} \text{Max}_{k=4, \dots, n+3} \{ |c_{n-1}^2 c_{n-k+3} - c_{n-1} c_{n-k+2} - c_{n-2} c_{n-k+3} + c_{n-k+1}|^{\frac{1}{k-1}} \} \quad (1.147)$$

$$(c_{-1} = c_{-2} = 0)$$

#### DEUTSCH'S 'SIMPLE' FORMULA

$$|\zeta| \leq |c_{n-1}| + \text{Max}_{i=0, \dots, n-2} \left\{ \left| \frac{c_i}{c_{i+1}} \right| \right\} \quad (1.148)$$

In the above, where denominators such as  $c_{i+1}$  are involved, it is assumed that these denominators are not 0. This was the case in all the tests run. But it was realized that if they were 0, the bounds would be infinite.



## Appendix for Section 10

### List of Formulas for Bounding Roots of Polynomials

(For detailed references see below).

#### Part A. Formulas based on Maximum of some function of the $c_i$

A1. (Guggenheimer 1978)  $\zeta \leq 2 \text{Max}_{i=1, \dots, n} |c_{n-i}|^{\frac{1}{7}}$

A2. (Deutsch 1981)  $\zeta \leq \text{Max}[1, |c_0| + |c_1| + \dots + |c_k|, 1 + |c_{k+1}|, \dots, 1 + |c_{n-1}|]$   
for each  $k = 0, 1, \dots, n-1$

A3. (Reich and Losser 1971) Let  $Q = [\text{Max}_{k=0, \dots, n-1} |c_k|]^{\frac{1}{n}}$  then  $\zeta \leq Q + Q^2 + \dots + Q^{n-1}$

A4. (ibid)  $\zeta \leq \text{Max}_{0 \leq j \leq k \leq n-1 (j \neq k)} [(1 + |c_k|)(1 + |c_j|)]^{\frac{1}{2}}$

A5. (Joyal et al 1967)  $\zeta \leq \frac{1}{2}[1 + \sqrt{1 + 4B'}]$  where  
 $B' = \text{Max}_{k=0}^{n-1} |c_{n-1}c_k - c_{k-1}|$  ( $c_{-1} = 0$ )

A6. (ibid)  $\zeta \leq 1 + \sqrt{B''}$  where  $B'' = \text{Max}_{k=0}^{n-1} |(1 - c_{n-1})c_k + c_{k-1}|$  ( $c_{-1} = 0$ )

A7. (Deutsch 1970)  $\zeta \leq \frac{1}{2}[1 + |c_{n-1}| + \sqrt{(c_{n-1})^2 + 4M}]$  where  $M = \text{Max}_{i=0}^{n-2} |c_i|$

A8. (ibid)  $\zeta \leq \text{Max}[2, |c_0| + |c_{n-1}|, |c_1| + |c_{n-1}|, \dots, |c_{n-2}| + |c_{n-1}|]$

A9. (ibid)  $\zeta \leq \sqrt{2 + M^2 + |c_{n-1}|^2}$  (M as in A7).

A10. (ibid)  $\zeta \leq \frac{1}{2}[\beta' + |c_{n-1}| + \sqrt{(|c_{n-1}| - \beta')^2 + 4\gamma'|c_{n-1}|}]$   
where  $\beta' = \text{Max}_{i=1}^{n-2} \frac{|c_i|}{|c_{i+1}|}$ ,  $\gamma' = \text{Max}_{i=0}^{n-2} \frac{|c_i|}{|c_{i+1}|}$ ,

A11. (ibid)  $\zeta \leq |c_{n-1}| + \gamma'$ , ( $\gamma'$  as above)

A12. (ibid)  $\zeta \leq \sqrt{2|c_{n-1}|^2 + (\beta')^2 + (\gamma')^2}$  ( $\beta'$ ,  $\gamma'$  as above)

A13. (ibid)  $\zeta \leq |c_{n-1}| + (\delta')^2$  where  $\delta' = \text{Max}_{i=0}^{n-2} \left[ \frac{|c_i|}{|c_{n-1}|^{n-2-i}} \right]$

A14. (ibid)  $\zeta \leq |c_{n-1}| + \text{Max} \left[ |c_{n-1}|, \frac{\delta'}{|c_{n-1}|} \right]$

A15. (ibid)  $\zeta \leq \sqrt{3|c_{n-1}|^2 + \frac{(\delta')^2}{|c_{n-1}|^2}}$  ( $\delta'$  as above)

A16. (ibid)  $\zeta \leq \frac{1}{2}[N + |c_{n-1}| + \sqrt{(|c_{n-1}| - N)^2 + 4N^2}]$   
where  $N = \text{Max}_{i=0}^{n-2} [|c_i|^{\frac{1}{n-i}}]$

A17. (ibid)  $\zeta \leq N + \text{Max}[N, |c_{n-1}|]$  (N as above)

A18. (ibid)  $\zeta \leq \sqrt{3N^2 + |c_{n-1}|^2}$  (N as above)

A19. (Takeya 1912)  $\zeta \leq \text{Max}_{i=0}^{n-1} \left| \frac{c_i}{c_{i+1}} \right|$  ( $c_n = 1$ )

A20. (Mignotte 1991)  $\zeta \leq 1 + \text{Max}[1, |c_0|, |c_1|, \dots, |c_{n-1}|]$

A21. (ibid)  $\zeta \leq n \text{Max}[1, |c_0|, |c_1|, \dots, |c_{n-1}|] + 1$

A22. (Datt and Govil 1978)  $\zeta \leq 1 + \left( 1 - \frac{1}{(1+M)^n} \right) M$   
where  $M = \text{Max}_{i=0}^{n-1} |c_i|$

A23. (Boese and Luther 1989) With M as above,

(i) If  $M < \frac{1}{n}$ ,  $\zeta \leq \left[ \frac{M(1-nM)}{(1-nM)^{\frac{1}{n}}} \right]^{\frac{1}{n}}$ ,

(ii) If  $M \geq \frac{1}{n}$ ,  $\zeta \leq \text{Min} \left[ (1+M) \left( 1 - \frac{M}{(1+M)^{n+1} - nM} \right), 1 + 2 \left( \frac{nM-1}{n+1} \right) \right]$

N.B. first result useful for small M, for then roots are small.

A24. (Birkhoff 1914)  $\zeta \leq \frac{\text{Max}_{i=1}^n |c_{n-i}|^{\frac{1}{i}}}{2^{\frac{1}{n}-1}}$  (here and below  $C_i^n$  is the binomial coefficient)

A25. (Diaz-Barrero 2002)  $\zeta \leq \text{Max}_{k=1}^n \left[ \frac{2^{n-1} |c_{n-k}| C_2^{n+1}}{k^2 C_k^n} \right]^{\frac{1}{k}}$

A26. (ibid)  $\zeta \leq \text{Max}_{k=1}^n \left[ \frac{F_{3n} |c_{n-k}|}{2^k F_k C_k^n} \right]^{\frac{1}{k}}$  where  $F_i$  = the i'th Fibonacci number, given by  $F_0 = 0$ ,  $F_1 = 1$ ,  $F_n = F_{n-1} + F_{n-2}$  ( $n \geq 2$ )

A27. (Riddell 1974)  $\zeta \leq \text{Max}\{\text{Max}_{i=0}^{n-1} \left| \frac{c_i}{c_{i+1}} \right| + |c_{n-1}|, \frac{c_0}{c_1}\}$  ( $c_n = 1$ )

A28. (Kojima 1917)  $\zeta \leq \text{Max}[2 \left| \frac{c_{n-1}}{c_n} \right|, \frac{k+1}{k-1} \left| \frac{c_{n-k}}{c_{n-k+1}} \right|$  ( $k = 2, \dots, n-1$ ),  $\frac{|c_0|}{(n-1)|c_1|}$ ]

A29. (Mignotte 1999)  $\zeta \leq \text{Max}_{i=1}^n [n|c_{n-i}|]^{\frac{1}{i}}$

A30. (Simuenovic 1991). If  $M_2 = \text{Max}_{i=2}^n \frac{|c_{n-2}| + \dots + |c_{n-i}|}{i-1}$  then  $\zeta \leq \frac{|c_{n-1}| + 1 + \sqrt{(|c_{n-1}| - 1)^2 + 4M_2}}{2}$

A31. (Mishra 1993)  $\zeta \leq \frac{\text{Min}_{c_i \neq 0} |c_i| + \text{Max}_{c_i \neq 0} |c_i|}{\text{Min}_{c_i \neq 0} |c_i|}$

### Part B. Formulas based on sums of some function of the $c_i$

B1. (Riddell 1974)  $\zeta \leq \text{Max}[1, \sum_{i=0}^{n-1} |c_i|]$

B2. (ibid, also Walsh 1924)  $\zeta \leq \sum_{i=0}^{n-1} |c_i|^{\frac{1}{n-i}}$

B3. (ibid)  $\zeta \leq |c_{n-1}| + \sum_{i=0}^{n-2} \frac{|c_i|}{|c_{i+1}|}$

B4 (Williams 1922)  $\zeta \leq \sqrt{1 + \sum_{i=0}^{n-1} |c_i|^2}$

B5. (ibid)  $\zeta \leq \sqrt{1 + (c_{n-1} - 1)^2 + \sum_{i=0}^{n-2} |c_i - c_{i+1}|^2 + c_0^2}$

B6. (Kittaneh 1995) With  $\alpha = \sum_{i=0}^{n-1} |c_i|^2$ ,  $\zeta \leq \sqrt{\frac{\alpha + 1 + \sqrt{(\alpha + 1)^2 - 4|c_0|^2}}{2}}$

B7. (Fujii and Kubo 1993)  $\zeta \leq \cos \frac{\pi}{n+1} + \frac{\sqrt{\sum_{i=0}^{n-1} |c_i|^2 + |c_{n-1}|}}{2}$

B8. (Rahman 1970)  $|\zeta + \frac{1}{2}c_{n-1}| \leq \frac{1}{2}|c_{n-1}| + \alpha M$  where  $M = \sum_{i=2}^n |c_{n-i}|^{\frac{1}{i}}$  and  $\alpha = \text{Max}_{i=2}^n [M^{-1}|c_{n-i}|^{\frac{1}{i}}]^{\frac{i-1}{i}}$  ( $\alpha = 0$  if all  $c_i = 0$ ,  $i = 0, \dots, n-2$ )

B9. (Alzer 1995)  $\zeta \leq |c_{n-1}| + \sqrt{\sum_{i=2}^n |c_{n-i}| \alpha^{i-2}}$  where  $\alpha = \frac{1}{\text{Max}_{i=2}^n |c_{n-i}|^{\frac{1}{i}}}$

B10. (Guggenheimer 1978)  $\zeta \leq |c_{n-1}| + \sum_{i=2}^n \left| \frac{c_{n-i}}{c_{n-1}^{i-1}} \right|$

B11. (ibid)  $\zeta \leq 1 + \sum_{i=1}^n b_i^{\frac{1}{i}}$  where  $b_i = \text{Max}_{j < i} (0, |c_{n-i}| - |c_{n-j}|)$  ( $i = 2, \dots, n$ );  $b_1 = |c_{n-1}|$

B12. (Kuniyeda 1916) (Only if  $\sum |c_i|^2 < \frac{1}{n-1}$ )

$$\zeta \leq \left[ \frac{n^n}{(n-1)^{n-1}} \sum_{i=0}^{n-1} |c_i|^2 \right]^{\frac{1}{2n}}$$

B13. (Parodi 1949)  $\zeta \leq \text{Max}[1, \frac{1}{2}[|c_{n-1}| + \sqrt{|c_{n-1}|^2 + 4 \sum_{j=0}^{n-2} |c_j|}]]$

B14. (Mignotte 1999) (Only if  $\sum < 1$ )  $\zeta \leq [\sum_{i=0}^{n-1} |c_i|]^{\frac{1}{n}}$

## References for formulas for bounds

Alzer, H. (1995), On the zeros of a polynomial, *J. Approx. Theory* **81**, 421-424

Birkhoff, G.D. (1914), An elementary double inequality for the roots of an algebraic equation having greatest absolute value, *Bull. Amer. Math. Soc.*, **21**, 494-495

Boese, F.G. and Luther, W.J. (1989), A Note on a Classical Bound for the Moduli of All Zeros of a Polynomial, *IEEE Trans. Autom. Control* **34**, 998-1001

Datt, B. and Govil, N.K. (1978), On the Location of the Zeros of a Polynomial, *J. Approx. Theory* **24**, 78-82

Deutsch, E. (1970), Matricial Norms and the Zeros of Polynomials, *Lin. Alg. Appls.* **3**, 483-489

Deutsch, E. (1981), Bounds for the Zeros of Polynomials, *Amer. Math. Monthly* **88**, 205-206

Diaz-Barrero, J.L. (2002), Note on bounds of the zeros, *Missouri J. Math. Sci.* **14**, 88-91

Fujii, M. and Kubo, F. (1993) Buzano's inequality and bounds for roots of algebraic equations, *Proc. Amer. Math. Soc* **117**, 359-361

Guggenheimer, H. (1962), Bounds for the Roots of Algebraic Equations, *Amer. Math. Monthly* **69**, 915-916

————— (1978), Bounds for roots of algebraic equations, *Arch. Math.* **31**, 568-569

Joyal, A., Labelle, G. and Rahman, Q.I. (1967), On the location of zeros of polynomials, *Canad. Math. Bull.* **10**, 53-63

Takeya, S. (1912), On the Limits of the Roots of an Algebraic Equation with Positive Coefficients, *Tohoku Math. J.* **2**, 140-142

Kalantari, B. (2004), Private Communication

————— (2005), An infinite family of bounds on zeros of analytic functions and relationship to Smale's bounds, *Math. Comp.* **74**, 841- 852

Kittaneh. F. (1995), Singular values of companion matrices and bounds on zeros of polynomials, *SIAM J. Matrix Anal. Appl.* **16**, 333-340

Kojima (1917), On the Limits of the Roots of an Algebraic Equation, *Tohoku Math. J.* **11**, 119-127

Kuniyeda (1916), Note on the Roots of Algebraic Equations, *Tohoku Math. J.* **9**, 187-188

McNamee, J.M. (2002), A 2002 update of the supplementary bibliography on roots of polynomials, *J. Comput. Appl. Math.* **142** (2002), 433-434. **Note** this contains instructions for accessing the bibliography via the web.

Mignotte, M. (1991), An inequality on the greatest roots of a polynomial, *Elem. Math.* **46**, 85-86

————— (1999), *Polynomials, An Algorithmic Approach*, Springer- Verlag

Mishra, B. (1993), Bounds on the Roots, in *Algorithmic Algebra*, Springer-Verlag, 306-308

Parodi, M. (1949), Sur les limites des modules des racines des équations algébriques, *Bull. Sci. Math. Ser. 2* **73**, 135-144

Rahman, Q.I. (1970), A bound for the moduli of the zeros of polynomials, *Canad. Math. Bull* **13**, 541-542

Reich, S. and Losser, O.P. (1971), Locating Zeros of Polynomials, *Amer. Math. Monthly* **78**, 681-683

Riddell, R.C. (1974), Upper bounds on the moduli of the zeros of a polynomial, *Math. Mag.* **47**, 267-273

Simeunovic, D.M. (1991), A Remark on the Zeros of a Polynomial, *Zeit. Angew. Math. Mech.* **71**, T832-835

van der Sluis, A. (1970), Upperbounds for Roots of Polynomials, *Numer. Math.* **15**, 250-262

Walsh, J.L. (1924), An inequality for the roots of an algebraic equation, *Ann. Math. Ser. 2* **25**, 285-286

Williams, K.P. (1922), Note concerning the roots of an equation, *Bull. Amer. Math. Soc.* **29**, 394-396

## 1.11 References for Chapter 1.

Adams, D.A. (1967), A Stopping Criterion for Polynomial Root Finding, *Comm. Ass. Comput. Mach.* **10**, 655-658

Aho, A.V., Steiglitz, K., and Ullman, J.D. (1975), Evaluating polynomials at fixed sets of points, *SIAM J. Comput.* **4**, 533-539

Dekker, T.J. (1971), A floating-point technique for extending the available precision, *Numer. Math.* **18**, 224-242

Dorn, W.S. (1962), Generalization of Horner's Rule for Polynomial Evaluation, *IBM J. Res. Dev.* **6**, 239-245

Dutta Roy, S.C. and Minocha, S. (1991) A Note on "Efficient Evaluation of Polynomials and Exponentials of Polynomials at Equispaced Arguments", *IEEE Trans. Acoustics, Speech, Sig. Proc.* **39**, 2554-2556

Estrin, G. (1960), Organization of a Computer System-The Fixed Plus Variable Structure Computer, *Proc. Western Joint Computer Conference*, AFIPS Press, Montvale, N.J., 33-40

Garwick, J.V. (1961), The limit of a converging sequence, *BIT* **1**, 64

Hammer, R. et al (1995), *C++ Toolbox for Verified Computing: Basic Numerical Problems* (Chap. 4), Springer-Verlag, Berlin

Hansen, E.R. et al (1990), Polynomial Evaluation with Scaling, *ACM Trans. Math. Software* **16**, 86-93

Igarashi, M. (1984), A Termination Criterion for Iterative Methods Used to Find the Zeros of Polynomials, *Math. Comp.* **42**, 165-171

Kahan, W. (1965), Further remarks on reducing truncation errors, *Comm. Ass. Comput. Mach.* **8**, 40

Kiper, A. (1997), Parallel Polynomial Evaluation by Decoupling Algorithm, *Par. Algs. Appl.* **9**, 145-152

Kowalik, J.S., and Kumar, S.P. (1985), Parallel Algorithms for Recurrence and Tridiagonal Equations, in *Parallel MIMD Computation: HEP Supercomputer and its Applications*, ed. J.S. Kowalik, MIT Press, 295-307

Kulisch, U.W. and Miranker, W.L. (1983), *A New Approach to Scientific Computation*, Academic Press, New York

Lakshmivarahan, S., and Dhall, S.K. (1990), *Analysis and Design of Parallel Algorithms*, McGraw-Hill, New York

Linnainmaa, S. (1981), Combatting the Effects of Underflow and Overflow in Determining Real Roots of Polynomials, *SIGNUM Newsletter* **16 No. 2**, 11-15

McNamee, J.M. (2002), A 2002 update of the supplementary bibliography on roots of polynomials, *J. Comput. Appl. Math* **142**, 433-434

————— and Olhovsky, M. (2005), A Comparison of a Priori Bounds on (Real or Complex) Roots of Polynomials, *Proceedings of the 37-th IMACS Conference*, Paris

Nuttall, A.H. (1987), Efficient Evaluation of Polynomials and Exponentials of Polynomials at Equispaced Arguments, *IEEE Trans. Acoustics, Speech, Sig. Proc.* **35**, 1486-1487

Oliver, J. (1979), Rounding error propagation in polynomial evaluation schemes, *J. Comput. Appl. Math.* **5**, 85-95

Paquet, L. (1994), Precise evaluation of a polynomial at a point given in staggered correction format, *J. Comput. Appl. Math.* **50**, 435-454

Pan, V.Y. et al (1997), Fast multipoint evaluation and interpolation via computations with structured matrices, *Ann. Numer. Math.* **4**, 483-510

Peters, G., and Wilkinson, J.H. (1971), Practical Problems Arising in the Solution of Polynomial Equations, *J. Inst. Math. Appl.* **8**, 16-35

Pichat, M. (1972), Correction d'une somme en arithmétique à virgule flottante, *Numer. Math.* **19**, 400-406

Shaw, M. and Traub, J.F. (1974), On the Number of Multiplications for the Evaluation of a Polynomial and Some of its Derivatives, *J. Ass. Comp. Mach.* **21**, 161-167

Stewart, G.W. III (1971), Error Analysis of the Algorithm for Shifting the Zeros of a Polynomial by Synthetic Division, *Math. Comp.* **25**, 135-139

Traub, J.F. (1971), Optimal iterative processes: theorems and conjectures, *Proc 1971 IFIP Congress Booklet TA-1*, 1273-1277

Volk, W. (1988), An Efficient Raster Evaluation Method for Univariate Polynomials, *Computing* **40**, 163-173

Werschultz, A.G. (1981), Maximal Order for Multipoint Methods with Mem-



ory Using Hermitian Information, *Intern. J. Comput. Math. Sec. B* **9**, 223-241

Wilkinson, J.H. (1965), *The Algebraic Eigenvalue Problem*, Clarendon Press

Wozniakowski, H. (1974), Rounding error analysis for the evaluation of a polynomial and some of its derivatives, *SIAM J. Numer. Anal* **11**, 780-787

## Chapter 2

# Sturm Sequences and Greatest Common Divisors

### 2.1 Introduction

Sturm sequences, derived from a polynomial and its derivative, provide a way of determining how many real roots lie in a specified interval, and ultimately, by means of a bisection process, of determining a range within which a single root lies. They have also been used by Wilf (1978) in locating complex roots. Ralston (1978) gives a good treatment, on which the following three sections are based.

### 2.2 Definitions and Basic Theorem

**Definition 1.** A sequence of polynomials  $f_1(x), f_2(x), \dots, f_m(x)$  is called a **STURM SEQUENCE** on an interval  $(a, b)$  if

$$(i) \quad f_m(x) \neq 0 \text{ in } (a, b) \quad (2.1)$$

(ii) at any zero of  $f_k(x)$ ,  $(k = 2, \dots, m - 1)$

$$f_{k-1}(x)f_{k+1}(x) < 0 \quad (2.2)$$

Note that this implies

$$f_{k-1} \neq 0 \text{ and } f_{k+1} \neq 0 \quad (2.3)$$

**Definition 2.** Let  $f_i(x)$ ,  $i = 1, \dots, m$ , be a Sturm sequence on  $(a, b)$ , and let  $x_0$  be a point at which  $f_1(x) \neq 0$ . We define  $V(x_0)$  = number of changes of

sign in  $f_i(x_0)$ , zeroes being ignored. If  $a = -\infty$ ,  $V(a)$  is defined as number of sign changes in  $\lim_{x \rightarrow -\infty} f_i(x)$ . Similarly for  $V(b)$  if  $b = +\infty$ .

**Definition 3.** Let  $R(x)$  be a rational function. We define the **CAUCHY INDEX**  $I_a^b R(x) = (\text{number of jumps from } -\infty \text{ to } +\infty) - (\text{number of jumps from } +\infty \text{ to } -\infty)$  as  $x$  goes from  $a$  to  $b$ , excluding endpoints.

Now we can prove:-

**STURM'S THEOREM.** If  $f_i(x)$ , ( $i = 1, \dots, m$ ) is a Sturm sequence on  $(a,b)$ , and if  $f_1(a) \neq 0$ ,  $f_1(b) \neq 0$ , then:-

$$I_a^b \frac{f_2(x)}{f_1(x)} = V(a) - V(b) \tag{2.4}$$

**Proof.**  $V(x)$  does not change when  $x$  passes through a zero  $x_0$  of  $f_k(x)$ , ( $k = 2, \dots, m$ ), because of 2.2 [e.g. if the sign of  $f_k$  is the same as that of  $f_{k-1}$  just left of  $x_0$ , it will be the same as that of  $f_{k+1}$  just right of  $x_0$ , so there is only one sign change in the sequence  $f_{k-1}, f_k, f_{k+1}$  both left and right of  $x_0$ ]. Thus  $V(x)$  can only change at a zero of  $f_1(x)$ .

Now if  $x_0$  is a zero of  $f_1(x)$ , it is not a zero of  $f_2(x)$ , by 2.3 with  $k=2$  (for if it were,  $f_1(x) \neq 0$ ). Hence  $f_2(x)$  has the same sign on both sides of  $x_0$ . So if  $x_0$  is a zero of  $f_1(x)$  of even multiplicity, then  $V(x)$  does not change as  $x$  passes through  $x_0$  (for  $f_1(x)$  does not change sign), while there is no contribution to the Cauchy index (for  $\frac{f_2}{f_1}$  remains equal to  $+\infty$ , or  $-\infty$ ). But, if  $x_0$  is of odd multiplicity,  $f_1(x)$  changes sign at  $x_0$ . If  $f_1$  and  $f_2$  have the same sign to the left of  $x_0$ ,  $V(x)$  increases by 1 while the Cauchy index has a contribution of -1 ( $\frac{f_2}{f_1}$  jumps from  $+\infty$  to  $-\infty$ ). Likewise if  $f_1$  and  $f_2$  have opposite sign,  $V(x)$  decreases by 1 while the Cauchy index incurs a contribution of +1. Thus  $I_a^b \frac{f_2}{f_1} = -(V(b) - V(a))$ . **Q.E.D.**

### 2.3 Application to Locating Roots

To find the real roots of  $f(x)$  in  $(a,b)$  we let  $f_1(x) \equiv f(x)$ ,  $f_2(x) = f'(x)$  and compute  $f_j(x)$ ,  $j=3, \dots, m$ , by

$$f_{j-1}(x) = q_{j-1}(x)f_j(x) - f_{j+1}(x), \quad j = 2, \dots, m - 1 \tag{2.5}$$

$$f_{m-1}(x) = q_{m-1}(x)f_m(x) \tag{2.6}$$

i.e.  $f_{j+1}$  is - remainder when  $f_{j-1}$  divided by  $f_j$ . The sequence  $f_j$  is of decreasing degree and hence must terminate with  $f_m(x)$ , possibly a constant, which divides  $f_{m-1}$  and hence all  $f_j$  ( $j=m-1, \dots, 1$ ), i.e.  $f_m$  is the greatest common divisor of  $f_1$  and  $f_2$ . If  $f_m \neq 0$  in (a,b) condition i) of Defn. 1 is satisfied, while by 2.5 condition ii) is satisfied, i.e.  $f_j$  is a Sturm sequence . If  $f_m=0$  in (a,b) we work with  $\{\frac{f_j}{f_m}\}$  which has the same values of  $I_a^b$ ,  $V(a)$ ,  $V(b)$  as  $\{f_j\}$ .

Now suppose

$$f = (x - \zeta_1)^{m_1}(x - \zeta_2)^{m_2} \dots (x - \zeta_p)^{m_p} Q(x), \tag{2.7}$$

where  $Q(x)$  has no real roots. Then

$$f'(x) = \sum_{i=1}^p m_i(x-\zeta_i)^{m_i-1} \prod_{j=1, \neq i}^p (x-\zeta_j)^{m_j} Q(x) + \prod_{j=1}^p (x-\zeta_j)^{m_j} Q'(x) \tag{2.8}$$

Hence

$$\frac{f_2}{f_1} = \frac{f'}{f} = \sum_{i=1}^p \frac{m_i}{x - \zeta_i} + \frac{Q'}{Q} \tag{2.9}$$

**THEOREM 2** Then  $I_a^b \frac{f_2}{f_1} =$  number of **distinct** real zeros in (a,b) =  $V(a)-V(b)$  (provided  $f(a) \neq 0$ ,  $f(b) \neq 0$ ).

If  $f(x)$  has a simple root at a or b the result holds with  $V(x)$ =number of sign changes in  $f_2(x), \dots, f_m(x)$ .

Using this theorem and a bisection process we may isolate one or more of the real roots of  $f(x)$ , i.e. find an interval  $(a_k, b_k)$  containing exactly one root (e.g. the largest). Start with

$$b_0 = 1.1 \times \text{an upper bound on roots of } f, \quad a_0 = -b_0$$

Now for  $k=0,1, \dots$  do:-

Find  $n_k = V(a_k) - V(b_k)$ . If  $n_k > 1$  then:-

Set  $c_k = (a_k + b_k)/2$ . Find  $V(c_k)$ .

If  $n_k = V(c_k) - V(b_k) > 1$  set  $a_k = c_k$  and continue;

else if this  $n_k = 0$  set  $b_k = c_k$  and continue;

else ( $n_k = 1$ ) the interval  $(c_k, b_k)$  contains exactly one root.

Now we may continue the bisection process until  $(b_k - a_k) <$  some error criterion, or we may switch to some more rapidly converging method. Dunaway (1974) describes a program which uses Sturm sequences to get a

first approximation to real roots and Newton's method to improve them.

## 2.4 Elimination of Multiple Roots

It was mentioned in Sec. 3 that special methods must be used if  $f_m = 0$  in (a,b), i.e. we should divide each  $f_j$  by  $f_m$ . Now it may be shown that dividing  $f_1$  ( $f$ ) by the g.c.d.  $f_m$  of  $f$  and  $f'$  leaves us with a polynomial having no multiple roots (if indeed  $f$  has any multiple roots to start with). For we see by 2.7 and 2.8 of Sec. 3 (allowing some or all of the  $\zeta_i$  to be complex and  $Q(x) = \text{constant}$ ) that  $f$  and  $f'$  have common factors

$$\prod_{i=1}^p (x - \zeta_i)^{m_i-1} \quad (2.10)$$

and these are the highest powers of  $(x - \zeta_i)$  which are common factors.

Thus, apart from a constant, 2.10 gives the g.c.d. of  $f$  and  $f'$ . Note that if  $m_i = 1$ ,

$$(x - \zeta_i)^{m_i-1} = 1 \quad (2.11)$$

i.e. this factor does not occur in the g.c.d. Hence if we divide  $f$  by the g.c.d. of  $f$  and  $f'$  we are left with

$$\prod_{i=1}^p (x - \zeta_i) \quad (2.12)$$

i.e. there are no multiple roots. It is recommended that one do this division before applying the Sturm sequence method. Not only is it then easier to apply Sturm, but other methods such as Newton's, which may be used in conjunction with Sturm, converge faster in the absence of multiple roots.

By repeating the process of finding and dividing by the g.c.d. we may obtain a set of polynomials each of which contain only zeros of a specific (known) multiplicity. Thus, with a slight change of notation, let

$$\begin{aligned} P_1 &= f \text{ (or } f_1), P_2 = \gcd(P_1, P_1'), \dots, \\ P_j &= \gcd(P_{j-1}, P_{j-1}') \text{ for } j = 2, \dots, m+1 \end{aligned} \quad (2.13)$$

where  $P_{m+1}$  is the first  $P_j$  to be constant. Then as we have seen  $P_2$  contains  $\prod_{i=1}^p (x - \zeta_i)^{m_i-1}$ , so in general  $P_j$  will contain

$$\prod_{i=1}^p (x - \zeta_i)^{m_i+1-j} \quad (2.14)$$

with the proviso that if

$$m_i + 1 - j \leq 0 \quad (2.15)$$

the  $i$ 'th factor is replaced by 1.

It follows from 2.14 that zeros of  $P_1$  (f) which have multiplicity

$$m_i \geq j \quad (2.16)$$

will appear in  $P_j$ , but not zeros which have

$$m_i < j \quad (2.17)$$

Moreover  $m$  is the greatest multiplicity of any zero of  $P_1$  (for by 2.16 and 2.17  $m+1 >$  all  $m_i$  and hence  $m+1 \geq \text{Max } m_i+1$ ).

Now we let

$$Q_j = \frac{P_j}{P_{j+1}} \quad (j = 1, 2, \dots, m-1) \quad (2.18)$$

$$Q_m = P_m \quad (2.19)$$

Each  $Q_j$  contains only simple zeros, and the zeros of  $Q_j$  appear in  $P_1$  with multiplicity  $\geq j$ . For using 2.14 and 2.15 we see that

$$Q_j = \frac{\prod_{i=1; m_i \geq j}^p (x - \zeta_i)^{m_i+1-j}}{\prod_{i=1; m_i \geq j+1}^p (x - \zeta_i)^{m_i-j}} \quad (2.20)$$

Clearly, for  $m_i > j$ , the factors  $(x - \zeta_i)$  in the numerator and in the denominator cancel except for a power of 1 remaining in the numerator. But for  $m_i = j$  the numerator contains just  $(x - \zeta_i)^1$  while the denominator contains no factor  $(x - \zeta_i)$ . In short

$$Q_j = \prod_{i=1; m_i \geq j}^p (x - \zeta_i) \quad (2.21)$$

We can take the process a further step, by letting

$$\tilde{Q}_j = \frac{Q_j}{Q_{j+1}} \quad (j = 1, 2, \dots, m-1) \quad (2.22)$$

$$\tilde{Q}_m = Q_m \quad (2.23)$$

Then

$$\tilde{Q}_j = \frac{\prod_{i=1; m_i \geq j}^p (x - \zeta_i)}{\prod_{i=1; m_i \geq j+1}^p (x - \zeta_i)} = \prod_{i=1; m_i=j}^p (x - \zeta_i) \quad (2.24)$$

or in words,  $\tilde{Q}_j$  contains all factors (to power 1) which are of multiplicity exactly  $j$ .

This is a very useful result, for if we factorize each  $\tilde{Q}_j$  separately we will not be handicapped by multiple roots, but we will know the multiplicity of each factor of  $\tilde{Q}_j$  in the original  $P_1$ .

### EXAMPLE

$$P_1 = (x - \zeta_1)^5(x - \zeta_2)^3(x - \zeta_3)^3(x - \zeta_4)$$

$$P_2 = (x - \zeta_1)^4(x - \zeta_2)^2(x - \zeta_3)^2$$

$$P_3 = (x - \zeta_1)^3(x - \zeta_2)(x - \zeta_3)$$

$$P_4 = (x - \zeta_1)^2$$

$$P_5 = (x - \zeta_1)$$

$$P_6 = 1$$

Thus we see that the highest multiplicity is 5.

Now

$$Q_1 = (x - \zeta_1)(x - \zeta_2)(x - \zeta_3)(x - \zeta_4)$$

$$Q_2 = (x - \zeta_1)(x - \zeta_2)(x - \zeta_3)$$

$$Q_3 = (x - \zeta_1)(x - \zeta_2)(x - \zeta_3)$$

$$Q_4 = (x - \zeta_1)$$

$$Q_5 = (x - \zeta_1)$$

and finally

$$\tilde{Q}_1 = (x - \zeta_4)$$

$$\tilde{Q}_2 = 1$$

$$\tilde{Q}_3 = (x - \zeta_2)(x - \zeta_3)$$

$$\tilde{Q}_4 = 1$$

$$\tilde{Q}_5 = (x - \zeta_1)$$

and we conclude that  $P_1$  has one simple root, two of multiplicity 3, and one of multiplicity 5. There are no roots of multiplicity 2 or 4. Note that initially  $\tilde{Q}_3$  will be given as a quadratic, not factorized. We will have to factorize it, and of course in general  $\tilde{Q}_i$  could be of still higher degree.

## 2.5 Detection of Clusters of Zeros (Near-Multiple)

The methods of the last section break down in the presence of rounding error, as in floating-point computation, since a small change in the coefficients leads to the disintegration of a  $k$ -fold zero into a cluster of  $k$  distinct (but usually close) zeros. In other words, a g.c.d. of  $(p, p')$  which is of degree  $> 0$ , as would be found for the multiple zero case if infinite precision is used (e.g. in pure integer or rational arithmetic), is replaced by a g.c.d. = 1 in floating-point arithmetic.

One popular solution to this problem is of course to work in rational arithmetic, as in a symbolic algebra system. This will be discussed in the next section. Here we will describe a method discussed by Hribernic and Stetter (1997), which enables us to compute clusters of zeros, with their multiplicities and centers, using a combination of symbolic and floating-point computation.

The key idea is the definition of a cluster in relation to an accuracy level:-

**DEFINITION 2.5.1** “At the accuracy level  $\alpha$ , a complex polynomial  $\tilde{p}$  possesses a  $k$ -cluster of zeros with center  $\zeta$  if there exists a polynomial  $p^*$ , with  $\deg p^* \leq \deg \tilde{p}$ , such that  $p^*$  has an exact  $k$ -fold zero at  $\zeta$ , and

$$\|\tilde{p} - p^*\| \leq \alpha \tag{2.25}$$

or equivalently

$$\tilde{p} = (x - \zeta)^k \tilde{q} + \tilde{r}(x), \text{ with } \|\tilde{r}\| \leq \alpha \tag{2.26}$$

Note that we have defined

$$\|p(x)\| \equiv \sum_{j=0}^n |c_j| \tag{2.27}$$



where  $p(x) = c_0 + c_1x + \dots c_nx^n$ .

Note also that the values of  $\zeta$ ,  $p^*$  or  $\tilde{q}$ ,  $\tilde{r}$  are not uniquely determined; but for a given  $\alpha$  there is a highest possible value of  $k$ —this is the value we assume in 2.26. Theoretically, this  $k$  is an increasing, integer-valued function of  $\alpha$ ; hence it is a discontinuous function. In practise. the **points** of discontinuity are replaced by critical **ranges**.

Since, as we have seen, the computation of the g.c.d.'s used in Sec. 4 breaks down, we replace the gcd by a “near- or  $\alpha$ -gcd” defined as follows:-

**DEFINITION** At the accuracy level  $\alpha$ , two polynomials  $\tilde{f}_1$  and  $\tilde{f}_2$  possess a near-gcd  $\tilde{g}$  if there exist polynomials  $f_1^*$  and  $f_2^*$  such that

$$\gcd(f_1^*, f_2^*) = \tilde{g} \text{ and } \|\tilde{f}_i - f_i^*\| \leq \alpha, \quad i = 1, 2 \quad (2.28)$$

or

$$\tilde{f}_i = \tilde{g}\tilde{q}_i + \tilde{r}_i \text{ where } \|\tilde{r}_i\| \leq \alpha, \quad i = 1, 2 \quad (2.29)$$

We seek a near-gcd  $\tilde{g}^*$  of the maximum degree feasible at accuracy level  $\alpha$ . Again,  $f_i^*$ ,  $\tilde{g}^*$  or  $\tilde{q}_i$ ,  $\tilde{r}_i$  are not uniquely defined, and so  $\deg \tilde{g}^*$  is an increasing integer-valued function of  $\alpha$  whose discontinuities are not sharp. On the other hand, once  $\tilde{g}$  has been computed,  $\text{Max}_{i=1,2} \|\tilde{r}_i\|$  is well-defined and we may assess the validity of our gcd.

The classical Euclidean algorithm (EA) for  $\gcd(f_1, f_2)$  computes  $q_i$  and  $f_{i+1}$  by division:-

$$f_{i-1} = f_i q_i + f_{i+1} \quad (i = 2, 3, \dots, l) \quad (2.30)$$

terminating when  $f_{l+1} = 0$ , so that

$$\gcd(f_1, f_2) = f_l \quad (2.31)$$

But, as mentioned, small perturbations of the  $f_i$  will lower the degree of the gcd dramatically, usually to 0. We expect that replacement of the criterion “ $f_{l+1} = 0$ ” by “ $f_{l+1}$  sufficiently small” will stabilize the gcd and yield a “near-gcd”. But how do we quantify this criterion at a given accuracy level  $\alpha$ ?. It turns out that we can write

$$f_i = s_j^{(i)} f_j + s_{j-1}^{(i)} f_{j+1} \quad (j > i \geq 1) \quad (2.32)$$

with

$$s_j^{(i)} = q_j s_{j-1}^{(i)} + s_{j-2}^{(i)}, \quad j > i \text{ and } s_{i-1}^{(i)} = 0, \quad s_i^{(i)} = 1 \quad (2.33)$$

**PROOF** Assume true up to  $j$ , i.e.

$$f_i = s_j^{(i)} f_j + s_{j-1}^{(i)} f_{j+1} \quad (j > i \geq 1)$$

Also

$$f_j = q_{j+1} f_{j+1} + f_{j+2} \quad (2.34)$$

$$\begin{aligned} \text{hence } f_i &= s_j^{(i)} \{q_{j+1} f_{j+1} + f_{j+2}\} + s_{j-1}^{(i)} f_{j+1} \\ &= \{q_{j+1} s_j^{(i)} + s_{j-1}^{(i)}\} f_{j+1} + s_j^{(i)} f_{j+2} \end{aligned}$$

$$= s_{j+1}^{(i)} f_{j+1} + s_j^{(i)} f_{j+2} \quad (2.35)$$

i.e. it is true for  $j+1$  in place of  $j$ .

$$\begin{aligned} \text{But } f_i &= q_{i+1} f_{i+1} + f_{i+2} \\ &= (q_{i+1} 1 + 0) f_{i+1} + 1 f_{i+2} \end{aligned}$$

$$= s_{i+1}^{(i)} f_{i+1} + s_i^{(i)} f_{i+2} \quad (2.36)$$

according to the definitions in 2.33

i.e. theorem is true for  $j = i+1$ ;

hence by induction for all  $j$ .

The  $s_j^{(1)}$  and  $s_j^{(2)}$  may be computed as we compute the  $f_i$ . Then comparing 2.32 and 2.29 with  $f_j = \tilde{g}$  and  $s_j^{(i)} = \tilde{q}_i$  we see that the criterion

$$\|s_{j-1}^{(i)} f_{j+1}\| \leq \alpha \quad (i = 1, 2) \quad (2.37)$$

means that  $f_j$  is an  $\alpha$ -gcd of  $f_1$  and  $f_2$ .

We may use floating-point arithmetic here provided that the backward error

$$\|s_{j-1}^{(i)} \rho_{j+1}\| \quad (2.38)$$

of the residuals

$$\rho_{j+1} = \tilde{f}_{j-1} - \tilde{f}_j q_j - \tilde{f}_{j+1} \quad (2.39)$$

remain small compared to  $\alpha$ . This leads to a dramatic reduction in execution time when the coefficients are ratios of large integers (as often happens

during an exact g.c.d. calculation).

We may check the adequacy of the near-gcd a posteriori by dividing  $\tilde{f}_1$  and  $\tilde{f}_2$  by  $\tilde{f}_l$  to give residuals  $v_l^{(i)}$

$$\tilde{f}_i = \tilde{f}_l w_l^{(i)} + v_l^{(i)}, \quad (i = 1, 2) \quad (2.40)$$

if  $\|v_l^{(i)}\| \ll \alpha$  ( $i = 1, 2$ ), one may test the previous  $v_{l-1}^{(i)}$  to see if  $\tilde{f}_{l-1}$  is a better near-g.c.d. (note that the higher the degree of  $\tilde{f}_l$ , the better).

**EXAMPLE 1** Consider

$$\tilde{f}_1 = \tilde{p}(x) = x^4 - 1.4143878x^3 + .0001232x^2 + .7071939x - .2500616$$

Take  $\tilde{f}_2 = \frac{1}{4}\tilde{p}'$ , and perform an EA in float-point arithmetic, computing the error bounds on L.H.S. Of 2.37 as we go. Results are

$j$	$\text{deg } \tilde{f}_j$	2.37
1	4	—
2	3	—
3	2	1.09
4	1	$.9 \times 10^{-7}$
5	0	$.6 \times 10^{-9}$

As our polynomial has an accuracy level  $.5 \times 10^{-7}$ , we appear to have a near-gcd  $\tilde{f}_3$  of degree 2.

The sequence  $P_i = \text{gcd}(P_{i-1}, P'_{i-1})$  described in section 4 will now be replaced by

$$\tilde{P}_i = \text{an } \alpha - \text{gcd}(\tilde{P}_{i-1}, \tilde{P}'_{i-1}) \quad (2.41)$$

where by an  $\alpha$ -gcd we mean a near-gcd at accuracy level  $\alpha$ .

Hribernic shows that

“For  $|\zeta| < 1$ , if  $(x - \zeta)^{k-1}$  is a factor of an  $\alpha$ -gcd of  $\tilde{P}$  and  $\tilde{P}'$ , then  $\tilde{P}$  has a  $k$ -cluster of zeros with center  $\zeta$  at an accuracy level  $3\alpha$ ”

Suppose that after applying 2.41 and the further computation of  $Q_j$  and  $\tilde{Q}_j$  as in sec 4 we have

$$\tilde{P}_1 \equiv \tilde{f}_1 = \prod_{j=0}^m (\tilde{Q}_j)^j + \tilde{r} \quad (2.42)$$

then we may form

$$\tilde{P}_1 = (\tilde{Q}_j)^j \tilde{q}_j + \tilde{r}_j \quad (j = 1, \dots, m) \quad (2.43)$$

and check the size of the  $\tilde{r}_j$ . If  $\|\tilde{r}_j\| \approx \alpha$ , ( $j=1, \dots, m$ ), then we accept 2.42 as a valid grouping of the zeros of  $\tilde{P}_1$  into clusters at accuracy level  $\alpha$ . But if  $\|\tilde{r}_j\| \gg (\ll) \alpha$  we have too few (too many) clusters, and we have to lower (raise) degree  $\tilde{P}_l$ , i.e. we have to take at least one step more (less) in the stabilized EA which generates  $\tilde{P}_l$

As in the standard procedure, if some of the  $\tilde{Q}_j$  have degree  $> 1$ , their approximate zeros must be found as centers of the several i-clusters which they represent. These zeros will be well-separated at the specified accuracy level (otherwise they would be considered as **one** cluster), and finding them should be relatively easy.

## 2.6 Sturm Sequences (or gcd's) Using Integers

As pointed out at the start of section 5, errors in floating arithmetic often prevent us from finding the true gcd of two polynomials. Note that the calculation of Sturm sequences is the same as that of gcd's, apart from the sign of the remainder at each step in the Euclidean algorithm (see below). Thus in what follows we will refer to gcd's, as is often done in the literature. (Also the methods of Sec. 4 require the calculation of gcd's, so this topic is important in its own right).

One favorite way of avoiding errors is to multiply the polynomial by the gcd of the denominator of its rational coefficients and work with integers only, if necessary using multiple precision (of course if the true coefficients are irrationals they will necessarily be approximated by rationals on input to the computer).

The traditional Euclid's method for finding the gcd of two polynomials  $P_1$  and  $P_2$  is to divide  $P_1$  by  $P_2$  giving remainder  $P_3$  and repeat that process until  $P_{l+1} = 0$ . Then  $P_l$  is the gcd of  $P_1$  and  $P_2$ . Unfortunately, even if  $P_1$  and  $P_2$  have integral coefficients,  $P_3$  etc generally have non-integral ones.

We can avoid this problem by instead computing the **pseudo-remainder** given for F and G by

$$g_m^{n-m+1}F = QG + R \quad (2.44)$$

where  $g_m$  is the leading coefficient of G,  $n$  and  $m$  are the degrees of F and G respectively, and  $\text{degree}(R) < \text{degree}(G) = m$ . To see why this works, observe that division is actually accomplished by a series of ‘partial divisions’

$$S_{i+1} = S_i - \frac{s_{in}}{g_m} x^{n_i-m} G \quad (2.45)$$

where  $s_{in}$  is the leading coefficient of  $S_i$ ,  $n_i$  is its degree, and  $S_0 = F$ . This eliminates successively lower powers of  $x$  from the partial remainders. However, to avoid non-integral coefficients we modify the above to

$$S_{i+1} = g_m S_i - s_{in} x^{n_i-m} G \quad (2.46)$$

This will be repeated until we reach a value of  $i$  such that  $n_{i+1} < m$ . Since  $n_i \leq n_{i-1} - 1$ , 2.46 is applied at most  $n-m+1$  times,  $S_i$  being multiplied by  $g_m$  each time. As we do not know how many times it will actually be used, we assume the worst case, leading to 2.44

If 2.44 is applied with  $F=P_{i-2}$ ,  $G = P_{i-1}$  and  $R = \beta_i P_i$ , i.e. to give

$$\beta_i P_i = \alpha_i P_{i-2} - Q P_{i-1} \quad (i = 3, \dots, l+1) \quad (2.47)$$

where

$$\alpha_i = l c_{i-1}^{n_{i-2}-n_{i-1}+1} \quad (2.48)$$

and  $n_i = \text{deg}(P_i)$ ,  $l c_i$  =leading coefficient of  $P_i$  and  $\beta_i$  is yet to be chosen, we will have what is called a **polynomial remainder sequence (prs)**.

The choice  $\beta_i = 1$  gives a Euclidean prs, but this leads to an exponential increase in coefficient size and thus is impractical except for low degrees. Or, we may divide the pseudo-remainder by the gcd of its coefficients, so that after division they will be relatively prime (then we say that the prs is **Primitive**). This controls the size, but finding the gcd of the coefficients takes a great deal of work.

A method which is often relatively efficient (that is, when  $\delta_i = n_i - n_{i+1} = 1$  for all  $i$ – the **normal** case) is the **Reduced prs** (Collins (1967)). Here we take

$$\beta_i = \alpha_{i-1} = l c_{i-2}^{\delta_{i-3}+1} \quad (2.49)$$

and it turns out (see Brown and Traub(1971) or Akritas (1987) for proof) that the pseudo-remainder in 2.47 in this case is exactly divisible by  $\beta_i$ , i.e.  $P_i$  has integral coefficients for all  $i$ . But also note that in the non-normal case for this method the coefficients may grow a great deal.

An even better, but slightly more complicated, method is known as the **Subresultant prs**, where a "subresultant" is a polynomial whose coefficients are certain determinants related to the resultant. In this method we take

$$\beta_3 = (-1)^{\delta_1+1}; \beta_i = -lc_{i-2}\psi_i^{\delta_i-2} \quad (i = 4, \dots, l+1) \quad (2.50)$$

where

$$\psi_3 = -1; \psi_i = (-lc_{i-2})^{\delta_i-3}\psi_{i-1}^{1-\delta_{i-3}} \quad (i = 4, \dots, l+1) \quad (2.51)$$

Brown and Traub (1971) give a good derivation, and Brown (1978) proves that the calculated  $P_i$  have integral coefficients. Note that if the prs is normal 2.50 and 2.51 reduce to

$$\psi_i = -lc_{i-2}; \beta_i = +lc_{i-2}^2 \quad (2.52)$$

In that case the reduced and subresultant prs are identical except possibly for signs.

Collins (1967) p139 shows that the coefficients of  $P_i$  in a subresultant prs do not exceed

$$(n - n_{l-1} + 1)(2d + \log_{10}2(n - n_{l-1} + 1)) \quad (2.53)$$

where  $d$  is the maximum number of decimal places in the coefficients of  $P_1$  and  $P_2$ . That is, they grow almost linearly with  $n$ , and no expensive computations of gcd's of coefficients are required. Moreover Brown (1978) p 247 shows that the cost is of order  $(d^2n^4)$ . Thus it appears that the Subresultant prs is the best of its class, although other methods such as heuristic or modular may be even faster.

## 2.7 Complex Roots (Wilf's Method)

We will show a method due to Wilf (1978) by which the number of roots in a given rectangle may be counted - then a two-dimensional extension of

the bisection method can be used to find a rectangle, as small as desired, containing a single root. The method is based on the "Argument Principle", which states:-" suppose no zeroes of  $f(z)$  lie on the boundary  $\partial R$  of  $R$ . Then the number  $N$  of zeroes of  $f(z)$  inside  $R$ =

$$\frac{1}{2\pi} \Delta_{\partial R}(\arg f(z)) = \frac{1}{2\pi} \times \{\text{change in arg } f(z) \text{ around } \partial R\} \quad (2.54)$$

Now consider a straight line from  $a$  to  $b$  which is part of  $\partial R$ . Let  $z=a+(b-a)t$ , so that

$$f(z) = \sum_{\nu=0}^n (\alpha_{\nu} + i\beta_{\nu})t^{\nu} = f_R(t) + if_I(t) \quad (2.55)$$

Consider the curve in the  $w$ -plane given by  $w=f(z)$  as  $z$  moves from  $a$  to  $b$ . If the  $w$ -curve crosses from the 1st quadrant to the 2nd, or from the 3rd to the 4th, the function  $f_I(t)/f_R(t) \equiv R(t)$  jumps from  $+\infty$  to  $-\infty$ , while if it crosses from the 2nd to the 1st, (or from 4th to 3rd),  $R(t)$  jumps from  $-\infty$  to  $+\infty$ . Hence  $-I_0^1 R =$  net excess of counter-clockwise over clockwise crossings by  $w$ -curve as  $z$  traverses  $ab$  and  $t$  goes from 0 to 1. But each extra counter-clockwise crossing contributes  $\pi$  to  $\arg f(z)$ . Hence

$$\Delta_{\partial R} \arg f(z) = -\pi \sum_{\text{edges of } R} I_0^1 f_I(t)/f_R(t) \quad (2.56)$$

Hence

$$N = -\frac{1}{2} \sum_{i=1}^m I_0^1 \{f_I^{(i)}(t)/f_R^{(i)}(t)\} \quad (2.57)$$

Now we may find each  $I_0^1$  by the Sturm's theorem using a sequence with  $f_1 = f_R$  and  $f_2 = f_I$ . That is, if  $Q_1, Q_2, Q_3, Q_4, (Q_5 = Q_1)$  are vertices of a rectangle, then on side  $Q_k Q_{k+1}$  we expand  $f(z)$  about  $Q_k$ , i.e. replace  $z$  by  $Q_k + i^{k-1}t$ ; let

$$\tilde{f}(t) = \sum_{\nu=0}^n (\alpha_{\nu} + i\beta_{\nu})t^{\nu} \quad (2.58)$$

and take

$$f_1(t) = \sum_{\nu=0}^n \alpha_{\nu} t^{\nu}, \quad f_2(t) = \sum_{\nu=0}^n \beta_{\nu} t^{\nu}. \quad (2.59)$$

The usual Sturm sequence will terminate with a **constant**  $f_m$  since  $f_1, f_2$  can have no common factor (for this would give a zero of  $f$  on a side of  $R$ ). Then

$$N = \frac{1}{2} \sum_{k=1}^4 \{V_k(|Q_{k+1} - Q_k|) - V_k(0)\} \quad (2.60)$$

Krishnamurthy and Venkateswaran (1981) describe a parallel version of Wilf's method, which they claim to work in  $O(n^3p)$  sequential time or  $O(n^2p)$  parallel time with  $n$  processors. Here  $2^{-p}$  is the precision required.

Camargo-Brunetto et al. (2000) recommend using integer arithmetic where possible, i.e. in calculating the Sturm sequences by, for example, the subresultant method referred to in sec. 6.

Pinkert (1976) describes a variation using Routh's theorem as well as Sturm's sequences.

A problem in all these methods is that Sturm's (or Routh's) theorem does not work if there is a root or roots on one of the boundaries of a rectangle. The method recommended to deal with this problem is as follows: let the transformed polynomial on a side be given by 2.55 above, i.e.  $f(z) = f_R(t) + if_I(t)$ . We will find the gcd  $h(t)$  of  $f_R(t)$  and  $f_I(t)$ . If this has a root  $\zeta$  on the side, then

$$f_R(\zeta) = f_I(\zeta) = 0 \quad (2.61)$$

i.e.  $\zeta$  is a root of  $f(z)$ . So we will find the real roots of  $h(t)$  (and  $f(z)$ ) on the side, as accurately as desired, by Sturm's theorem and bisection. We record these roots. Then we shift the line a small distance to the left (or up) and repeat the process (more than once if necessary). If and when we reach a line with no roots on it, we apply Wilf's method to the resulting rectangles on left and right (or above and below).

## 2.8 References for Chapter 2

Akritas, A.G. (1987), A Simple Proof of the Validity of the Reduced prs Algorithm, *Computing* **38**, 369-372



Brown, W.S. (1978), The Subresultant PRS Algorithm, *ACM Trans. Math. Software* **4**, 237-249

Brown, W.S. and Traub, J.F. (1971), On Euclid's Algorithm and the Theory of Subresultants, *J. Assoc. Comput. Mach.* **18**, 505-514

Camargo Brunetto, M.A.O., Claudio, D.M., and Trevisan, V. (2000), An Algebraic Algorithm to Isolate Complex Polynomial Zeros Using Sturm Sequences, *Comput. Math. Appls.* **39**, 95-105

Collins, G.E. (1967), Subresultants and Reduced Polynomial Remainder Sequences, *J. Assoc. Comput. Mach.* **14**, 128-142

Dunaway, D.K. (1974), Calculation of Zeros of a Real Polynomial Through Factorization Using Euclid's Algorithm, *SIAM J. Numer. Anal.* **11**, 1087-1104

Hribernic, V. and Stetter, H.J. (1997), Detection and Validation of Clusters of Polynomial Zeros, *J. Symbolic Comput.* **24**, 667-681

Krishnamurthy, E.V. and Venkateswaran, H. (1981), A Parallel Wilf Algorithm for Complex Zeros of a Polynomial, *BIT* **21**, 104-111

Pinkert, J.R. (1976), An Exact Method for Finding the Roots of a Complex Polynomial, *ACM Trans. Math. Software* **2**, 351-363

Ralston, A. (1978), *A First Course in Numerical Analysis*, McGraw-Hill, New York

Wilf, H.S. (1978), A Global Bisection Algorithm for Computing the Zeros of Polynomials in the Complex Plane, *J. Assoc. Comput. Mach.* **25**, 415-420

## Chapter 3

# Real Roots by Continued Fractions

### 3.1 Fourier and Descartes' Theorems

Fourier (1820) gave an easily calculated upper bound on the number of roots of a real polynomial  $p(x)$  between any two values  $a, b$  of  $x$ . Like Sturm's theorem (see Chapter 2) it is based on the number of sign variations in a certain sequence, where we have the definition:-

If  $c_0, c_1, \dots, c_n$  are a sequence of numbers and  $c'_0, c'_1, \dots, c'_r$  the sub-sequence of their non-zero members (re-numbered if necessary), then we say that a sign variation exists if  $c_p$  and  $c_{p+1}$  have opposite signs. Let  $\text{Var}\{c_i\}$  = total number of sign variations in the sequence.

EXAMPLE  $p(x) = x^3 - 5x^2 + 3$ . Sequence of coefficients =  $\{1, -5, 0, 3\}$ ,  $\text{Var} = 2$ .

Now we can state Fourier's theorem:-

**THEOREM 3.1.1** Let  $\text{fseq}(x) = \{p(x), p'(x), p^{(2)}(x), \dots, p^{(n)}(x)\}$ . If we replace  $x$  in the above by two numbers  $a$  and  $b$  ( $a < b$ ) we have that:

(i) the number of real roots of  $p(x) = 0$ , between  $a$  and  $b = \text{Var}\{\text{fseq}(a)\} - \text{Var}\{\text{fseq}(b)\} - 2\lambda$ , where  $\lambda$  is a positive integer or zero.

PROOF. see Akritas (1989) p339.

EXAMPLE (as before)  $p(x) = x^3 - 5x^2 + 3$ ;  $\text{fseq}(x) = \{x^3 - 5x^2 + 3, 3x^2 - 10x, 6x - 10, 6\}$ . Take  $a = 0, b = 1$ , then number of roots between 0 and 1 =  $\text{Var}\{\text{fseq}(0)\} - \text{Var}\{\text{fseq}(1)\} - 2\lambda = \text{Var}\{3, 0, -10, 6\} - \text{var}\{-1, -7, -4, 6\} - 2\lambda = 2 - 1 - 2\lambda = 1$  (for since number of roots  $\geq 0$ ,  $\lambda$  must = 0).

Somewhat easier to apply, although the number of roots is still not given precisely, is Descartes' rule (given in his *Geometrie* in 1637, proved by Gauss in 1828-see Bartolozzi and Franci (1993)).

**THEOREM 3.1.2** Let  $p(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0$  be a real polynomial. Then number of positive roots =  $\text{Var}\{c_i\} - 2\lambda$ , where again  $\lambda$  is an integer  $\geq 0$ .

**PROOF** We apply Fourier's theorem above, using the fact that

$$\begin{aligned} p(x) &= c_n x^n + \dots + c_0 \\ p^{(1)} &= n c_n x^{n-1} + \dots + c_1 \\ p^{(2)} &= n(n-1) c_n x^{n-2} + \dots + 2c_2 \\ &\dots \\ &\dots \\ &\dots \\ p^{(n)} &= (n!) c_n \end{aligned}$$

Then  $\text{fseq}(0) = \{c_0, c_1, 2c_2, \dots, (n!)c_n\}$ ; while for  $x = \infty$   $\text{fseq}(\infty)$  has **no** sign variations, since each member has the sign of  $c_n$ . Hence number of positive roots = number between 0 and  $\infty = \text{Var}\{\text{fseq}(0)\} - \text{Var}\{\text{fseq}(\infty)\} - 2\lambda = \text{Var}\{c_i\} - 0 - 2\lambda$ . QED.

### 3.2 Budans's Theorem

This theorem is the basis of the important Vincent's theorem. It is equivalent to Fourier's theorem, but leads in a different direction. It states (see Budan 1807):-

**THEOREM 3.2.1** Suppose in a real equation  $p(x) = 0$  we make two distinct substitutions  $x = \alpha + x'$  and  $x = \beta + x''$ , where  $\alpha$  and  $\beta$  are real and  $\alpha < \beta$ , giving equations  $A(x') = \sum a_i x'^i = 0$  and  $B(x'') = \sum b_i x''^i = 0$ . Then

- (i)  $\text{Var}\{a_i\} \geq \text{Var}\{b_i\}$
- (ii) The number of real roots of  $p(x) = 0$  between  $\alpha$  and  $\beta = \text{Var}\{a_i\} -$

$\text{Var}\{b_i\} - 2\lambda$ ,

where as usual  $\lambda = \text{an integer} \geq 0$

To see that this is equivalent to Fourier's theorem note that the coefficients  $a_i = \frac{p^{(i)}(\alpha)}{i!}$  (by Taylor's theorem). Thus  $\text{Var}\{a_i\} = \text{Var}\{\text{fseq}(\alpha)\}$  and similarly for  $\beta$  and  $b_i$ .

### 3.3 Vincent's Theorem

This theorem, first published in Vincent (1836), leads to a relatively efficient method for isolating the real roots of a real polynomial. Now in a computer only rational numbers can be stored, and we can multiply a polynomial with rational coefficients by the least common multiplier of their denominators to give integer coefficients. We may then work with exact integer arithmetic to give rational bounds on real (possibly irrational) roots, thus avoiding problems with rounding error endemic in floating point calculations.

As background let us look at Descartes' rule again: it gives the exact number of roots only in two special cases:-

- (i) If there are no variations, there is no positive root.
- (ii) if there is one sign variation, there is exactly one positive root.

The converse of (i) is true according to:

**LEMMA 3.3.1** (Stodola). If  $p(x) = c_n x^n + \dots + c_0$  ( $c_i$  real,  $c_n > 0$ ) has only roots with negative real parts, then  $\text{Var}\{c_i\} = 0$

**PROOF** Let  $-\alpha_i$  ( $i=1, \dots, k$ ) be the real roots, and let  $-\gamma_m \pm i\delta_m$  ( $m = 1, 2, \dots, s$ ) be the complex roots, where  $\alpha_i$  and  $\gamma_m > 0$ , all i,m. Then  $p(x)$  can be written as the product  $c_n \prod_{i=1}^k (x + \alpha_i) \prod_{m=1}^s ([x + \gamma_m]^2 + \delta_m^2)$ , where all the factors have positive coefficients, hence all  $c_i > 0$ , i.e.  $\text{Var}\{c_i\} = 0$ .

The converse of (ii) is not true in general, as we see from the counter- example  $x^3 - 2x^2 + x - 2 = (x - 2)(x - i)(x + i)$  which has one positive root but 3 variations of sign.

However, under certain special conditions the converse of (ii) IS true, namely:

**LEMMA 3.3.2** (Akritas and Danielopoulos 1985). Let  $p(x)$  be a real polynomial of degree  $n$ , with simple roots, having one positive root  $\xi$  and  $n-1$  roots  $\xi_1, \dots, \xi_{n-1}$  with negative real parts, these roots being of the form

$\xi_j = -(1 + \alpha_j)$  with  $|\alpha_j| < \epsilon = (1 + \frac{1}{n})^{\frac{1}{n-1}} - 1$ . Then  $p(x)$  has exactly one sign variation.

**PROOF** see Akritas (1989) or paper referred to above.

**THEOREM 3.3.3** (Vincent 1836). If in a polynomial  $p(x)$  with rational coefficients we make successive transformations

$$x = a_1 + \frac{1}{x'}, x' = a_2 + \frac{1}{x''}, x'' = a_3 + \frac{1}{x'''}, \text{ etc}$$

where  $a_1$  is an arbitrary non-negative integer and  $a_2, a_3, \dots$  are arbitrary positive integers, then eventually the transformed equation has either zero or one sign variation. If one, the equation has exactly one positive root given by the continued fraction

$$a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}$$

If zero, it has no positive roots.

**PROOF**-see Vincent's paper.

### 3.4 Akritas' Improvement of Vincent's Theorem

Vincent's theorem was apparently forgotten until 1948 when Uspensky extended it to give a bound on the number of transformations required to give one or zero sign variation(s). Akritas (1978) corrected Uspensky's proof.

**THEOREM 3.4.1** let  $p(x)$  be a polynomial of degree  $n$  with rational coefficients and simple roots, and let  $\Delta > 0$  be the minimum distance between any two roots. Let  $m$  be the smallest integer such that

$$F_{m-1} \frac{\Delta}{2} > 1 \text{ and } F_{m-1} F_m \Delta > 1 + \frac{1}{\epsilon_n} \quad (3.1)$$

where the  $F_m$  are the Fibonacci numbers given by

$$F_{m+1} = F_m + F_{m-1} \quad (F_1 = F_2 = 1) \quad (3.2)$$

and

$$\epsilon_n = (1 + \frac{1}{n})^{\frac{1}{n-1}} - 1 \quad (3.3)$$

Let  $a_i$  be as in Theorem 3.3.3. then the transformation

$$x = a_1 + \frac{1}{a_2 + \dots + \frac{1}{a_m + \frac{1}{\xi}}} \quad (3.4)$$

transforms  $p(x)$  into  $\hat{p}(\xi)$  which has 0 or 1 sign variation.

**PROOF** We need to show that the real parts of all complex roots and all real roots, except at most one, become negative. Then the theorem follows by Lemmas 3.3.1 and 3.3.2. Now let  $\frac{p_k}{q_k}$  be the  $k$ 'th convergent to the continued fraction in 3.4 (see Appendix for some definitions and proofs regarding continued fractions, especially equations 3.35 and 3.36). Then for  $k \geq 0$ , (with  $p_0 = 1$ ,  $p_{-1} = 0$ ,  $q_0 = 0$ ,  $q_{-1} = 1$ ) we have

$$\begin{aligned} p_{k+1} &= a_{k+1}p_k + p_{k-1} \\ q_{k+1} &= a_{k+1}q_k + q_{k-1} \end{aligned} \quad (3.5)$$

From  $q_1 = 1$  and  $q_2 = a_2 \geq 1$  (and  $a_k \geq 1$ ) we have

$$q_k \geq F_k \quad (3.6)$$

Also 3.4 can be written

$$x = \frac{p_m \xi + p_{m-1}}{q_m \xi + q_{m-1}} \quad (3.7)$$

(by 3.39 with  $x$  in place of  $\xi$  and  $\xi$  in place of  $\xi_n$ )  
and so

$$\xi = -\frac{p_{m-1} - q_{m-1}x}{p_m - q_m x} \quad (3.8)$$

Hence, if  $x_0$  is a root of  $p(x) = 0$ , then  $\xi_0$  given by 3.8 with  $x_0$  in place of  $x$  is a root of  $\hat{p}(\xi) = 0$ .

1) Suppose  $x_0$  is complex =  $a+ib$  ( $b \neq 0$ ). Then

$$\begin{aligned} Re(\xi_0) &= -Re \left[ \frac{(p_{m-1} - q_{m-1}a) - iq_{m-1}b}{p_m - q_m a - iq_m b} \right] \\ &= -Re \left[ \frac{\{(p_{m-1} - q_{m-1}a) - iq_{m-1}b\} \{(p_m - q_m a) + iq_m b\}}{\{(p_m - q_m a) - iq_m b\} \{(p_m - q_m a) + iq_m b\}} \right] \\ &= - \left[ \frac{(p_{m-1} - q_{m-1}a)(p_m - q_m a) + q_{m-1}q_m b^2}{(p_m - q_m a)^2 + q_m^2 b^2} \right] \end{aligned} \quad (3.9)$$

This will be negative if

$$(p_{m-1} - q_{m-1}a)(p_m - q_m a) \geq 0 \quad (3.10)$$

but what if this quantity is  $< 0$  ?

Well, then  $a$  lies between  $\frac{p_{m-1}}{q_{m-1}}$  and  $\frac{p_m}{q_m}$ , whose difference in absolute value is  $\frac{|p_{m-1}q_m - p_m q_{m-1}|}{q_{m-1}q_m} = \frac{|(-1)^m|}{q_{m-1}q_m} = \frac{1}{q_{m-1}q_m}$  (see Appendix, theorem A1, part 1).

Hence

$$\left| \frac{p_{m-1}}{q_{m-1}} - a \right| \text{ and } \left| \frac{p_m}{q_m} - a \right| \text{ are both } < \frac{1}{q_{m-1}q_m} \quad (3.11)$$

and so

$$|(p_{m-1} - aq_{m-1})(p_m - aq_m)| < \frac{1}{q_{m-1}q_m} \leq 1 \quad (3.12)$$

Consequently

$$Re(\xi_0) < 0 \text{ if } q_{m-1}q_m b^2 > 1 \quad (3.13)$$

But by definition of  $\Delta$ ,  $|(a + ib) - (a - ib)| = |2ib| = 2|b| \geq \Delta$ , and the conditions of the theorem give

$$q_{m-1}|b| \geq q_{m-1} \frac{\Delta}{2} \geq F_{m-1} \frac{\Delta}{2} \text{ (by 3.6)} > 1 \text{ (by 3.1)}$$

Moreover  $q_m \geq q_{m-1}$  (by 3.5), so also  $q_m|b| > 1$

Finally

$$q_{m-1}q_m b^2 > 1 \quad (3.14)$$

so  $Re(\xi_0) < 0$

2) Suppose  $x_0$  is a real root. Suppose that for all real roots  $x_i$  we have

$$(p_{m-1} - q_{m-1}x_i)(p_m - q_m x_i) > 0 \quad (3.15)$$

Then by 3.8 all real roots of  $\hat{p}(\xi)$  are  $< 0$ , while from 1) all the complex roots of  $\hat{p}(\xi)$  have negative real parts. Hence by Lemma 3.3.1,  $\hat{p}(\xi)$  has no sign variations.

Now suppose on the other hand that 3.15 is not true. Then  $x_0$  lies between  $\frac{p_{m-1}}{q_{m-1}}$  and  $\frac{p_m}{q_m}$ , and hence as in 3.11,

$$\left| \frac{p_m}{q_m} - x_0 \right| \leq \frac{1}{q_{m-1}q_m} \quad (3.16)$$

Also by 3.8  $\xi_0$  is  $> 0$

Let  $x_k$  ( $k \neq 0$ ) be another root (real or complex) of  $p(x) = 0$ , and  $\xi_k$  the corresponding root of  $\hat{p}(\xi) = 0$ . Then using 3.8

$$\begin{aligned} \xi_k + \frac{q_{m-1}}{q_m} &= - \left( \frac{p_{m-1} - q_{m-1}x_k}{p_m - q_m x_k} \right) + \frac{q_{m-1}}{q_m} \\ &= \frac{-p_{m-1}q_m + q_{m-1}p_m + (q_{m-1}q_m - q_{m-1}q_m)x_k}{q_m(p_m - q_m x_k)} = \\ &= \frac{(-1)^m}{q_m(p_m - q_m x_k)} \end{aligned} \quad (3.17)$$

Hence

$$\begin{aligned} \xi_k &= - \left( \frac{q_{m-1}}{q_m} \right) \left[ 1 - \frac{(-1)^m}{q_{m-1}q_m \left( \frac{p_m}{q_m} - x_k \right)} \right] = \\ &= - \frac{q_{m-1}}{q_m} (1 + \alpha_k) \end{aligned} \quad (3.18)$$

where

$$\alpha_k = \frac{(-1)^{m-1}}{q_{m-1}q_m \left( \frac{p_m}{q_m} - x_k \right)} \quad (3.19)$$

But  $\left| \frac{p_m}{q_m} - x_k \right| = \left| \frac{p_m}{q_m} - x_0 + x_0 - x_k \right| \geq |x_0 - x_k| - \left| \frac{p_m}{q_m} - x_0 \right| \geq \Delta - \frac{1}{q_{m-1}q_m}$   
 (by 3.16)  $= \left( \frac{q_{m-1}q_m \Delta - 1}{q_{m-1}q_m} \right) >$

$$\frac{(F_{m-1}F_m \Delta - 1)}{q_{m-1}q_m} > 0 \text{ by 3.1} \quad (3.20)$$

Hence  $|\alpha_k| \leq \frac{1}{q_{m-1}q_m \Delta - 1} \leq \frac{1}{F_{m-1}F_m \Delta - 1}$ , and then by 3.1

$$|\alpha_k| < \epsilon_n \leq \frac{1}{2} \quad (3.21)$$

So  $\xi_k = - \left( \frac{q_{m-1}}{q_m} \right) (1 + \alpha_k)$  all have negative real parts. Thus  $\hat{p}(\xi)$  has one positive roots and all other roots have negative real parts satisfying conditions of Lemma 3.3.2, so  $\hat{p}(\xi)$  must have exactly one sign variation. See Akritas (1989) p369 for the case where the LHS of 3.15 = 0 exactly.



### 3.5 Applications of Theorem 3.4.1

If  $\hat{p}(\xi) = 0$  has only one variation of signs, then it has only one positive root  $\xi'$ , i.e.  $0 < \xi' < \infty$ . Thus substituting 0 and  $\infty$  in 3.7 we see that the corresponding  $x'$  must lie between  $\frac{p_{m-1}}{q_{m-1}}$  and  $\frac{p_m}{q_m}$ . This is usually a very narrow range of values, giving a good starting range for approximating the root  $x'$  (by the same or different method).

The process must be repeated for each positive root of  $p(x) = 0$  (see section 7). Also, for negative roots we replace  $x$  by  $-x$  in  $p(x)$  and proceed as before.

### 3.6 Complexity of m

$m$  is the smallest index such that both parts of 3.1 hold. Suppose the first one fails for  $m-1$ , i.e.

$$F_{m-2} \frac{\Delta}{2} \leq 1 \quad (3.22)$$

now for large  $m$ ,  $F_k \approx \frac{1}{\sqrt{5}}\phi^k$  where  $\phi = \frac{1+\sqrt{5}}{2} = 1.618$   
Hence  $\phi^{m-2} \leq \frac{2\sqrt{5}}{\Delta}$ , hence  $(m-2) \leq \log_\phi(\frac{2\sqrt{5}}{\Delta})$ , i.e.

$$m \leq 2 + \log_\phi 2 + \frac{1}{2} \log_\phi 5 - \log_\phi \Delta \quad (3.23)$$

But (see Akritas 1989, sec 7.2.4 or Mahler 1964)

$$\Delta \geq \sqrt{3} n^{-\frac{n+2}{2}} |p(x)|_1^{-(n+1)} \quad (3.24)$$

i.e.

$$\log_\phi \Delta \geq \frac{1}{2} \log_\phi 3 - \frac{n+2}{2} \log_\phi n - (n+1) \log_\phi |p(x)|_1 \quad (3.25)$$

Combining 3.23 and 3.25 and using  $L(|p(x)|_1) \approx L(|p(x)|_\infty)$  gives

$$m = O\{nL(n) + nL(|p(x)|_\infty)\} \quad (3.26)$$

But usually  $L(n) = O(1)$  so we get

$$m = O\{nL(|p(x)|_\infty)\} \quad (3.27)$$

in the above  $L(x)$  means “the number of bits in  $x$ ” or  $\log_2(x)$ , and

$$|p(x)|_\infty = \text{Max}_{0 \leq i \leq n} |c_i| \quad (3.28)$$

while

$$|p(x)|_1 = \sum_{i=0..n} |c_i| \leq n|p(x)|_\infty \quad (3.29)$$

### 3.7 Choice of the $a_i$

Vincent used a very simple method, i.e. he took  $a_i = 1$ , applying  $x = 1+y$  repeatedly until he detected a change in sign variation. Unfortunately this takes a very long time if the smallest positive root is very large.

Akritis (see e.g. his book and several papers) uses a much more efficient method— he takes  $a_i = b =$  lower bound on positive roots of some intermediate  $\hat{p}(\xi)$ . This is the same as finding an upper bound on roots of  $\hat{p}(\frac{1}{\xi})$ . This can be done e.g. by Cauchy's rule (see section 8). Then he makes the transformation  $x = b + \frac{1}{y}$  and repeats the process for the next  $a_i$

Rosen and Shallit (1978) use Newton's method to find  $a_i$  as the largest integer  $\leq$  the smallest positive root of  $\hat{p}(\xi)$ . They start with an initial approximation of 1, and find the root with an error  $\leq .5$ , rounding the estimate to the nearest integer  $t$ . Then they test  $t-1$ ,  $t$ ,  $t+1$  to see which is really [root]. The rest as before.

It is not clear whether Akritis' or Rosen and Shallit's method is more efficient, or which is most robust (Newton's method is liable to diverge).

The literature does not appear to explain how we get the second, third, etc., roots, but we think it can be done as follows. When the lowest positive root of  $p(x)$  has been isolated between  $\frac{p_m}{q_m}$  and  $\frac{p_{m-1}}{q_{m-1}}$ , we let  $\beta =$  maximum of these two values. Then in the original  $p(x)$  make the transformation  $x = \beta + y$  to give  $\bar{p}(y)$ . This polynomial will have the second lowest positive root of  $p(x)$  as its smallest positive root. We then apply the usual procedure to  $\bar{p}(y)$ , and so on until all the positive roots are obtained.

When isolating intervals have been found for all the real roots, we may approximate them as accurately as required by continuing the process until  $|\frac{p_h}{q_h} - \alpha| \leq \frac{1}{q_{h-1}q_h} \leq$  required error.

### 3.8 Cauchy's Rule

(For details of this see Obreschkoff 1963 pp50-51).

**THEOREM 3.8.1** Let  $p(x) = x^n + c_{n-1}x^{n-1} + \dots + c_1x + c_0$  with  $c_{n-k} < 0$  for at least one  $k$ , and let  $\lambda$  be the number of its negative  $c_{n-k}$ . Then

$$b = \text{Max}_{1 \leq k \leq n: c_{n-k} < 0} \{ |\lambda c_{n-k}|^{\frac{1}{k}} \} \quad (3.30)$$

**PROOF.** Clearly  $b^k \geq \lambda |c_{n-k}|$  for each  $k$  with  $c_{n-k} < 0$ ,  
i.e.  $b^n \geq \lambda |c_{n-k}| b^{n-k}$

Summing over these  $k$  gives  $\lambda b^n \geq \sum_{c_{n-k} < 0} \lambda |c_{n-k}| b^{n-k}$

Hence in  $p(b)$  we have that positive  $b^n \geq$  sum of absolute values of negative terms. Thus if  $x$  increases above  $b$ , the positive term(s) increase at a greater rate than the negative terms, so  $p(x) \neq 0$  for  $x > b$ , i.e.  $b$  is an upper bound on the largest root.

Akritis gives an efficient way of computing  $b$  (e.g. see his book pp350-351), and shows that its time-complexity is  $O\{n^2 L(|p(x)|_\infty)\}$

### 3.9 Appendix to Chapter 3. Continued Fractions

Given a rational fraction  $\frac{a_0}{a_1}$  in lowest terms, and  $a_1 > 0$ , we may apply the Euclidean Algorithm to give

$$\begin{aligned} a_0 &= a_1 c_0 + a_2 \quad (0 < a_2 < a_1) \\ a_1 &= a_2 c_1 + a_3 \quad (0 < a_3 < a_2) \\ &\dots \\ &\dots \\ a_i &= a_{i+1} c_i + a_{i+2} \quad (0 < a_{i+2} < a_{i+1}) \\ &\dots \\ &\dots \\ a_{k-1} &= a_k c_{k-1} \end{aligned} \quad (3.31)$$

i.e. the process stops when  $a_{k+1} = 0$ , as it eventually must by 3.31.

Letting  $\xi_i = \frac{a_i}{a_{i+1}}$ , 3.31 may be written

$$\xi_i = c_i + \frac{1}{\xi_{i+1}} \tag{3.32}$$

Thus  $\xi_0 = c_0 + \frac{1}{\xi_1} = c_0 + \frac{1}{c_1 + \frac{1}{\xi_2}} = \dots =$

$$c_0 + \frac{1}{c_1 + \frac{1}{c_2 + \frac{1}{\dots + \frac{1}{c_{k-2} + \frac{1}{c_{k-1}}}}}} \tag{3.33}$$

which is the continued fraction expansion of  $\frac{a_0}{a_1}$ . The  $c_i$  are called partial quotients. Note that  $c_0$  may have any sign, but the other  $c_i > 0$ .

**EXAMPLE** Consider  $\frac{13}{8}$ . We have  $13 = 8 \times 1 + 5$ ;  $8 = 5 \times 1 + 3$ ;  $5 = 3 \times 1 + 2$ ;  $3 = 2 \times 1 + 1$ ;  $2 = 1 \times 2 + 0$ . Thus, all  $c_i = 1$ , except the last one, which = 2. Hence  $\frac{8}{5} = 1 + \frac{1}{1 + \frac{1}{\frac{1}{2}}}$ . The last  $c_{k-1}$  (2) may be replaced by  $1 + \frac{1}{1}$ .

Akritis (book, p 46) shows that if we take  $c_{k-1} > 1$  the expansion is unique.

An irrational number  $\xi$  may be expressed by an infinite continued fraction with the  $c_i$  defined by  $\xi_0 = \xi$ ,  $c_0 = [\xi_0]$ ,  $\xi_1 = \frac{1}{\xi_0 - c_0}, \dots$ ,

$$\xi_{i+1} = \frac{1}{\xi_i - c_i}, \text{ where } c_i = [\xi_i], i = 1, 2, \dots \tag{3.34}$$

Defining

$$\begin{aligned} p_{-2} &= 0, p_{-1} = 1, p_i = c_i p_{i-1} + p_{i-2} \quad (i = 0, 1, 2, \dots), \\ q_{-2} &= 1, q_{-1} = 0, q_i = c_i q_{i-1} + q_{i-2} \quad (i = 0, 1, 2, \dots) \end{aligned} \tag{3.35}$$

the continued fraction in 3.33, with  $k-1 = n$ , has the value

$$r_n = \frac{p_n}{q_n}, (p_n, q_n) = 1 \tag{3.36}$$

This is called the  $n$ 'th convergent to  $\xi$ . Since  $c_i \geq 1$  ( $i \geq 1$ ) and  $q_0 = 1, q_1 \geq 1$ , we have

$$q_i > q_{i-1} \quad (i \geq 2) \tag{3.37}$$

If we replace  $c_{k-1} = c_n$  in 3.33 by the infinite continued fraction

$$\xi_n = c_n + \frac{1}{c_{n+1} + \frac{1}{\dots}} \quad (3.38)$$

then we have

$$\xi = \frac{p_{n-1}\xi_n + p_{n-2}}{q_{n-1}\xi_n + q_{n-2}} \quad (3.39)$$

**THEOREM A1.** Let  $r_n = \frac{p_n}{q_n}$  be the  $n$ 'th convergent to a finite or infinite continued fraction expansion of  $\xi$ . Then

1)  $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}$  ( $n = 1, 2, \dots$ )

2)  $r_n - r_{n-1} = \frac{(-1)^{n-1}}{q_{n-1} q_n}$  ( $n = 1, 2, \dots$ )

3)  $r_n - r_{n-2} = (-1)^{n-2} \frac{c_n}{q_n q_{n-2}}$  ( $n = 2, 3, \dots$ )

4) For even  $n$ ,  $r_n \rightarrow \xi$  (and  $r_n$  increasing)

For odd  $n$   $r_n \rightarrow \xi$  (and  $r_n$  decreasing)

$r_{2n} < r_{2n-1}$  all  $n$ ,  $r_n$  lies between  $r_{n-1}$  and  $r_{n-2}$

**PROOF** 1) For  $n = 1$ , by 3.35,  $p_1 q_0 - p_0 q_1 = (c_1 p_0 + p_{-1}) q_0 - (c_0 p_{-1} + p_{-2}) q_1 = (c_1 c_0 + 1)(1) - (c_0 1 + 0)(c_1 1 + 0) = 1$

Assume true for  $k$ , i.e.  $p_k q_{k-1} - p_{k-1} q_k = (-1)^{k-1}$

then  $p_{k+1} q_k - p_k q_{k+1} = (c_{k+1} p_k + p_{k-1}) q_k - p_k (c_{k+1} q_k + q_{k-1}) = -(p_k q_{k-1} - p_{k-1} q_k) = (-1)^k$ , i.e. the result is true for  $k+1$ .

Hence, by induction 1) is true for all  $k$ .

2) Divide 1) by  $q_n q_{n-1}$  to give  $\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n-1}}{q_n q_{n-1}}$  and the result follows by definition of  $r_n$ .

$$\begin{aligned} 3) \quad r_n - r_{n-2} &= \frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(p_n q_{n-2} - p_{n-2} q_n)}{q_n q_{n-2}} \\ &= \frac{(c_n p_{n-1} + p_{n-2}) q_{n-2} - p_{n-2} (c_n q_{n-1} + q_{n-2})}{q_n q_{n-2}} \\ &= \frac{c_n (p_{n-1} q_{n-2} - p_{n-2} q_{n-1})}{q_n q_{n-2}} \\ &= \frac{(-1)^{n-2} c_n}{q_n q_{n-2}}. \text{ i.e. the result is proved.} \end{aligned}$$

4) By 3)  $r_{2n+2} - r_{2n} = \frac{(-1)^{2n} c_{2n}}{q_{2n} q_{2n-2}} > 0$ , i.e.  $r_{2n} < r_{2n+2}$  or  $r_{2n} > r_{2n-2}$

Similarly  $r_{2n-1} > r_{2n+1}$

By 2)  $r_{2n} - r_{2n-1} = \frac{(-1)^{2n-1}}{q_{2n-1} q_{2n}} < 0$  i.e.  $r_{2n-1} > r_{2n}$

Thus the sequence  $\{r_{2n}\}$  is increasing and bounded above by  $r_1$  (for  $r_{2n} < r_{2n-1} < r_{2n-3} < \dots < r_3 < r_1$ ); hence it tends to a limit. Similarly

$\{r_{2n+1}\} \rightarrow$  a limit. But  $|r_{2n} - r_{2n-1}| \rightarrow 0$  as  $n \rightarrow \infty$  by 2) and the fact that the  $q_n$  are increasing as  $n \rightarrow \infty$ . Hence the two limits must be the same. Also we have shown that  $r_{2n} > r_{2n-2}$  and  $r_{2n-1} > r_{2n}$ , i.e. with  $n$  even,  $r_n$  lies between  $r_{n-2}$  and  $r_{n-1}$ . We can show the same result for  $n$  odd.

**THEOREM A2.** Let  $\xi$  be an irrational number, and

$$c_0 + \frac{1}{c_1 + \frac{1}{c_2 + \dots + c_{n-1} + \frac{1}{\xi_n}}}$$

be its continued fraction expansion, where  $\xi_n$  is given by 3.38.

Then 1) Each  $r_n$  is nearer to  $\xi$  than  $r_{n-1}$ .

2)  $\frac{1}{2q_{n+1}q_n} < |\xi - \frac{p_n}{q_n}| < \frac{1}{q_{n+1}q_n} < \frac{1}{q_n^2}$ .

**PROOF** 1) By 3.39  $\xi = \frac{(p_{n-1}\xi_n + p_{n-2})}{(q_{n-1}\xi_n + q_{n-2})}$

Hence  $\xi_n(\xi q_{n-1} - p_{n-1}) = -(\xi q_{n-2} - p_{n-2})$

Hence (dividing by  $\xi_n q_{n-1}$ )  $|\xi - \frac{p_{n-1}}{q_{n-1}}| = |\frac{q_{n-2}}{\xi_n q_{n-1}}| |\xi - \frac{p_{n-2}}{q_{n-2}}|$

But  $\xi_n > 1$ ,  $q_{n-1} > q_{n-2}$ , hence  $0 < |\frac{q_{n-2}}{\xi_n q_{n-1}}| < 1$

Hence  $|\xi - \frac{p_{n-1}}{q_{n-1}}| < |\xi - \frac{p_{n-2}}{q_{n-2}}|$

2) From 2) of Theorem A1 we have  $|r_{n+1} - r_n| = \frac{1}{q_{n+1}q_n}$ , and from 1) above  $\xi$  is closer to  $r_{n+1}$  than to  $r_n$ ; hence  $\frac{1}{2q_{n+1}q_n} < |\xi - \frac{p_n}{q_n}| < \frac{1}{q_{n+1}q_n} < \frac{1}{q_n^2}$

### 3.10 References for Chapter 3

Akritis, A.G. (1978), A correction on a theorem by Uspensky, *Bull. Greek. Math. Soc.* **19**, 278-285

Akritis, A.G. (1989), *Elements of Computer Algebra with Applications*, Wiley, New York.

Akritis, A.G. and Danielopoulos, S.D. (1985), A converse rule of signs for polynomials, *Computing* **34**, 283-286

Bartolozzi, M. and Franci, R. (1993), La regola dei segni dall'enunciato di R. Descartes..., *Arch. Hist. Exact Sci.* **45**, 335-374

Budan, F. (1807), *Nouvelle Méthode pour la Résolution des Équations d'un Degré Quelconque* 2/E, Paris, (1822)

Fourier, J. (1820), *Le bulletin des sciences par la Societe philomatique de Paris*.

Mahler, K. (1964), An inequality for the discriminant of a polynomial, *Michigan Math. J.* **11**, 257-262

Obreschkoff, N. (1963), *Verteilung und Berechnung der Nullstellen reeller Polynome*, VEB Deutscher Verlag der Wissenschaften, Berlin

Rosen, D. and Shallit, J. (1978), A Continued Fraction Algorithm for Approximating all Real Polynomial Roots, *Math. Mag.* **51**, 112-116

Uspensky, J.V. (1948), *Theory of Equations*, McGraw-Hill, New York.

Vincent (1836), Sur la résolution des équations numériques, *J. Math. Pures Appl.* **1**, 341-372

## Chapter 4

# Simultaneous Methods

### 4.1 Introduction and Basic Methods

Until the 1960's all known methods for solving polynomials involved finding one root at a time, say  $\zeta_i$ , and deflating (i.e. dividing out the factor  $x-\alpha_i$ ). This can lead to problems with increased rounding errors, and unlike the simultaneous methods described in this chapter, does not lend itself to parallel computing.

The first-to-be-discovered and simplest method of this class of simultaneous methods is the following:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - z_j^{(k)})} \quad (i = 1, \dots, n; k = 0, 1, \dots) \quad (4.1)$$

where  $z_i^{(0)}$  is some initial guess. The various  $z_i^{(k+1)}$  can be formed independently or in parallel. This formula was first mentioned by Weierstrass (1903) in connection with the Fundamental Theorem of Algebra. It was re-discovered by Ilieff (1948), Docev (1962), Durand (1960), Kerner (1966), and others. We shall call it the WDK method. A straightforward evaluation of 4.1, with  $P(z)$  computed by Horner's method, requires about  $2n^2$  complex multiplications and  $2n^2$  additions or subtractions. However Werner (1982) describes a computational scheme which needs only  $\frac{1}{2}n^2$  multiplications,  $\frac{1}{2}n^2$  divisions, and  $2n^2$  additions/subtractions.

Aberth (1973) gives a simple derivation of 4.1 as follows: let  $z_i$  and  $\hat{z}_i = z_i + \Delta z_i$  be old and new approximations to the roots  $\zeta_i$ . Ideally, we



would like  $\hat{z}_i$  to equal  $\zeta_i$ , so

$$P(z) = \prod_{i=1}^n (z - [z_i + \Delta z_i]) = \quad (4.2)$$

$$\prod_{k=1}^n (z - z_k) - \sum_{i=1}^n \Delta z_i \prod_{k=1, \neq i}^n (z - z_k) + O(\Delta z_i^2) \quad (4.3)$$

Neglecting powers (and products) of  $\Delta z_i$  higher than the first, setting  $z = z_1, z_2, \dots, z_n$  in turn, and solving for  $\Delta z_i$  gives

$$\Delta z_i = \frac{-P(z_i)}{\prod_{k=1, \neq i}^n (z_i - z_k)} \quad (4.4)$$

which leads to 4.1 when iterated. Semerdzhiev (1994) also derives the WDK method by solving a differential equation.

Besides the advantage of lending itself to parallel computation, the WDK method is much more robust than e.g. Newton's method, i.e. it nearly always converges, no matter what the initial guess(es). Experiments by Semerdzhiev (1994) found it to converge for all but 4 out of 4000 random polynomials. He observes that "after slight changes in the initial approximations the process becomes a convergent one". Similarly Docev and Byrnev (1964) find convergence in all but 2 out of 5985 cases. Again, perturbations result in convergence. For  $n=2$ , Small (1976) proves that the method always converges except for starting points on a certain line. It is conjectured that a similar situation is true for all  $n$  (although this is not proved so far). However we shall see in Section 2 that various conditions on the starting approximations and the polynomial values at these points will guarantee convergence.

Hull and Mathon (1996) prove that convergence is quadratic for simple roots. For we have

$$\hat{z}_i - \zeta_i = z_i + \Delta z_i - \zeta_i = z_i - \zeta_i - (z_i - \zeta_i) \prod_{k=1, \neq i}^n \frac{z_i - \zeta_k}{z_i - z_k}; \quad (4.5)$$

let

$$\epsilon = \max_{1 \leq i \leq n} |z_i - \zeta_i| \quad (4.6)$$

and note that

$$\frac{z_i - \check{\zeta}_k}{z_i - z_k} = 1 + \frac{z_k - \check{\zeta}_k}{z_i - z_k} \quad (4.7)$$

Now if  $\zeta_i$  is a simple root, then for small enough  $\epsilon$   $|z_i - z_k|$  is bounded away from zero, and so

$$\frac{z_i - \check{\zeta}_k}{z_i - z_k} = 1 + O(\epsilon) \quad (4.8)$$

and

$$\prod_{k=1, \neq i}^n \frac{z_i - \check{\zeta}_k}{z_i - z_k} = (1 + O(\epsilon))^{n-1} = 1 + O(\epsilon) \quad (4.9)$$

Hence

$$\tilde{z}_i - \zeta_i = (z_i - \zeta_i)(1 - (1 + O(\epsilon))) = O(\epsilon^2) \quad (4.10)$$

Other authors, such as Niell (2001) and Hopkins et al (1994) give more complicated proofs of quadratic convergence. In fact, Hopkins shows that if

$$\| \mathbf{z}^{(0)} - \zeta \| \leq \frac{4^{\frac{1}{n-1}} - 1}{2 * 4^{\frac{1}{n-1}} + 1} \delta \quad (4.11)$$

where

$$\delta = \min(|\zeta_i - \zeta_j|, i, j = 1, \dots, n, i \neq j) \quad (4.12)$$

and  $\zeta$  is the vector  $(\zeta_1, \zeta_2, \dots, \zeta_n)$ , then the iterates converge, and

$$\overline{\lim}_{k \rightarrow \infty} \frac{\| \mathbf{z}^{(k+1)} - \zeta \|}{\| \mathbf{z}^{(k)} - \zeta \|^2} \leq \frac{n-1}{\delta} \quad (4.13)$$

An alternative derivation of the WDK formula, used by Durand and Kerner, is to note that the roots satisfy

$$\sum_{i=1}^n \zeta_i = -c_{n-1}/c_n, \quad \sum_{i=1}^n \sum_{k>i} \zeta_i \zeta_k = c_{n-2}/c_n, \quad \text{etc} \quad (4.14)$$

and apply Newton's method for systems, giving a set of linear equations for the  $\Delta z_i$ . The first of these is

$$\sum_{i=1}^n \Delta z_i + \sum_{i=1}^n z_i = -c_{n-1}/c_n \quad (4.15)$$

$$i.e. \sum \hat{z}_i = -c_{n-1}/c_n \quad (4.16)$$

Thus the sum of approximations in each iteration (except the first) remains constant and equal to the sum of the true roots. The above proof is due to Hull, but Kjelberg (1984), Niell, and Hopkins give alternative proofs. This fact is used by Small to eliminate one of the  $z_i$  in the case of  $n = 2$ , and Kjellberg uses it to explain the observation that approximations towards a multiple root tend to be grouped uniformly on a circle around that root, thus: "...the approximations converging towards the simple zeros reach their goals comparatively quickly, ..., and do not then contribute to any difference between the centres of gravity of the zeros and the approximations. The approximations converging towards the multiple zero must then have their centre of gravity equal to the multiple zero, and therefore be grouped around it". (More about this in Section 3).

Another method, apparently discovered by Borsch-Supan (1963), and also described by Ehrlich (1967) and Aberth (1973), is as follows:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{1}{\frac{P'(z_i^{(k)})}{P(z_i^{(k)})} - \sum_{j=1, \neq i}^n \frac{1}{z_i^{(k)} - z_j^{(k)}}} \quad (4.17)$$

Aberth derives it as follows: let

$$R_i(z) = \frac{P(z)}{\prod_{j \neq i} (z - z_j)} \quad (4.18)$$

and apply Newton's method to  $R_i(z)$  at  $z_i$ . Now

$$R'_i(z) = \frac{P'(z) \prod_{j=1, \neq i}^n (z - z_j) - P(z) \sum_{j=1, \neq i}^n \prod_{l=1, \neq i, \neq j}^n (z - z_l)}{(\prod_{l=1, \neq i}^n (z - z_l))^2} \quad (4.19)$$

Then Newton's formula gives:

$$\frac{R_i(z_i)}{R'_i(z_i)} = \frac{N}{D}$$

where

$$N = \frac{P(z_i)}{\prod_{j=1, \neq i}^n (z_i - z_j)}$$

$$D =$$

$$\frac{\left(P'(z_i) \prod_{j=1, \neq i}^n (z_i - z_j) - P(z_i) \sum_{j=1, \neq i}^n \prod_{l=1, \neq i, \neq j}^n (z_i - z_l)\right)}{\left(\prod_{l=1, \neq i}^n (z_i - z_l)\right)^2} \quad (4.20)$$

leading to 4.17. Aberth, Ehrlich, and Farmer and Loizou (1975) have proved that the above method has cubic convergence for simple roots.

Simeunovic (1989) describes a variation

$$z_i^{(k+1)} = a_i - \frac{1}{\frac{P'(a_i)}{P(a_i)} - \sum_{j=1, \neq i}^n \frac{1}{a_i - z_j^{(k)}}} \quad (4.21)$$

where the  $a_i$  are constants chosen to be fairly close to the roots (assumed real and distinct). He gives some elaborate conditions for convergence. It is possible that this method may be more efficient than the standard Aberth method.

Borsch-Supan (1970) and Nourein (1975) give the formula

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)}) / \prod_{j=1, \neq i}^n (z_i^{(k)} - z_j^{(k)})}{1 + \sum_{j=1, \neq i}^n \frac{P(z_j^{(k)}) / \prod_{l=1, \neq j}^n (z_j^{(k)} - z_l^{(k)})}{z_i^{(k)} - z_j^{(k)}}} \quad (4.22)$$

This method also has cubic convergence. Again, Werner gives an efficient way of evaluating the formula 4.22 which requires only  $\frac{n^2}{2}$  multiplications and  $\frac{3}{2}n^2$  divisions.

In (1977a) Nourein gives an improvement of the above, which may be summarized as

$$z_i^{(k+1)} = z_i^{(k)} - \frac{W_i}{1 + \sum_{j=1, \neq i}^n \frac{W_j}{z_i^{(k)} - W_i - z_j^{(k)}}} \quad (4.23)$$

where

$$W_i = \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - z_j^{(k)})} \quad (4.24)$$

is the WDK correction term. This has fourth order convergence. It requires about the same work as 4.22, or about the same amount as two WDK iterations, which together give order 4.

Similarly in (1977b) Nourein gives the “Improved Durand-Kerner Method”:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - z_j^{(k)} + W_j)} \quad (4.25)$$

which is third order, and the “Improved Ehrlich-Aberth Method”:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{1}{\frac{P'(z_i^{(k)})}{P(z_i^{(k)})} - \sum_{j=1, \neq i}^n \frac{1}{z_i^{(k)} - z_j^{(k)} + \frac{P(z_j^{(k)})}{P'(z_j^{(k)})}}}} \quad (4.26)$$

which is of fourth order.

The book by Petkovic (1989B) gives many more details of the above and other methods.

## 4.2 Conditions for Guaranteed Convergence

As mentioned above, the WDK method seems in practice to converge from nearly all starting points. (Kjurkchiev (1998) gives conditions for NON-convergence of the WDK and several other methods). However this behaviour has not been proved, except for  $n = 2$ . It is desirable to have conditions under which convergence is guaranteed, and several authors, notably Petkovic and his colleagues, have described such conditions, for various methods. In fact Petkovic and Herceg (2001) proceed as follows:

$$\text{Let } W_i^{(k)} = \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - z_j^{(k)})} \quad (i = 1 \dots n, k = 0, 1, \dots) \quad (4.27)$$

$$w^{(0)} = \text{Max}_{1 \leq i \leq n} |W_i^{(0)}|, \quad (4.28)$$

and

$$d^{(0)} = \text{Min}_{j \neq i} |z_i^{(0)} - z_j^{(0)}| \quad (4.29)$$

Conditions for safe convergence are found as

$$w^{(0)} \leq c_n d^{(0)} \quad (4.30)$$

where  $c_n$  depends on  $n$  and some constant(s). Note that a large  $c_n$  enables a larger value of  $w^{(0)}$ , i.e. initial guesses further from the roots.

Most simultaneous methods can be expressed as

$$z_i^{(k+1)} = z_i^{(k)} - C_i(z_1^{(k)}, \dots, z_n^{(k)}) \quad (4.31)$$

where

$$C_i = \frac{P(z_i^{(k)})}{F_i(z_1^{(k)}, \dots, z_n^{(k)})} \quad (4.32)$$

and  $F_i(z_1, \dots, z_n) \neq 0$  for  $z_i = \zeta_i$  (roots of P) or  $z_i$  distinct. Define

$$g(t) = 1 + 2t, \quad 0 < t \leq 1/2; = 1/(1-t), \quad 1/2 < t < 1 \quad (4.33)$$

They state a theorem, which we will number as (4.2.1):

**THEOREM 4.2.1**

“Let an iterative method be defined as above, and let  $z_i^{(0)}$  ( $i=1, \dots, n$ ) be initial approximations to the roots of P. If there exists  $\beta \in (0, 1)$  such that

$$(i) |C_i^{(k+1)}| \leq \beta |C_i^{(k)}| \quad (k = 0, 1, \dots) \quad (4.34)$$

$$(ii) |z_i^{(0)} - z_j^{(0)}| > g(\beta)(|C_i^{(0)}| + |C_j^{(0)}|) \quad (i \neq j, i, j = 1, \dots, n) \quad (4.35)$$

then the method converges”

Proof: see Petkovic, Herceg and Ilic (1998)

Analysis of convergence is based on Theorem 4.2.1 and the following relations named (W-D) etc:

$$(W - D) : w^{(0)} \leq c_n d^{(0)}$$

(already mentioned)

$$(W - W) : |W_i^{(k+1)}| \leq \delta_n |W_i^{(k)}| \quad (i = 1, \dots, n; k = 0, 1, \dots) \quad (4.36)$$

$$(C - C) : |C_i^{(k+1)}| \leq \beta_n |C_i^{(k)}| \quad (i = 1, \dots, n; k = 0, 1, \dots) \quad (4.37)$$

$$(C - W) : |C_i^{(k)}| \leq \lambda_n \frac{|W_i^{(k)}|}{c_n} \quad (i = 1, \dots, n; k = 0, 1, \dots) \quad (4.38)$$

where  $\delta_n, \beta_n, \lambda_n$  are  $> 0$  and depend only on  $n$ .  $c_n$  must be chosen so that

$$\delta_n < 1 \quad (4.39)$$

and hence  $W_i^{(k)}$  converges to 0. This will imply (by 4.31 and 4.38) that

$$|z_i^{(k+1)} - z_i^{(k)}| \rightarrow 0 \quad (4.40)$$

and hence  $z_i^{(k)} \rightarrow \zeta_i$  (for then  $P(\text{Lim } z_i^{(k)}) = 0$ ).

Also the choice of  $c_n$  must provide that

$$\beta_n < 1 \quad (4.41)$$

and

$$\lambda_n < \frac{1}{2g(\beta_n)} \quad (4.42)$$

and 4.38. For if 4.42 and 4.38 are both true then  $|z_i^{(0)} - z_j^{(0)}| \geq d^{(0)}$  (by 4.29)  $\geq \frac{w^{(0)}}{c_n}$  (by 4.30)  $\geq \frac{|W_i^{(0)}| + |W_j^{(0)}|}{2c_n}$  (by 4.28)  $\geq \frac{|C_i^{(0)}| + |C_j^{(0)}|}{2\lambda_n}$  (by 4.38)  $\geq g(\beta_n)(|C_i^{(0)}| + |C_j^{(0)}|)$  (by 4.42) which is (ii) of our theorem 4.2.1. Finally we must also prove 4.37 (which is the same as 4.34 i.e. (i) of theorem 2.1) subject to 4.41: this will ensure that  $|C_i^{(m)}| \rightarrow 0$ , i.e. convergence of whatever method we are considering. We will need

**Lemma 4.2.2** For distinct numbers  $z_1, \dots, z_n, \hat{z}_1, \dots, \hat{z}_n$  let

$$d = \text{Min}_{1 \leq i, j \leq n; i \neq j} |z_i - z_j|, \quad \hat{d} = \text{Min} |\hat{z}_i - \hat{z}_j| \quad (4.43)$$

and assume

$$|\hat{z}_i - z_i| \leq \lambda_n d \quad (i = 1, \dots, n) \quad (4.44)$$

Then

$$|\hat{z}_i - z_j| \geq (1 - \lambda_n) d \quad (4.45)$$

$$|\hat{z}_i - \hat{z}_j| \geq (1 - 2\lambda_n) d \quad (4.46)$$

and

$$\left| \prod_{j \neq i} \frac{\hat{z}_i - z_j}{\hat{z}_i - \hat{z}_j} \right| \leq \left( 1 + \frac{\lambda_n}{1 - 2\lambda_n} \right)^{n-1} \quad (4.47)$$

**PROOF**

$$|\hat{z}_i - z_j| = |z_i - z_j + \hat{z}_i - z_i| \geq$$

$$\begin{aligned}
|z_i - z_j| - |\hat{z}_i - z_i| &\geq d - \lambda_n d = (1 - \lambda_n)d \\
|\hat{z}_i - \hat{z}_j| &= |z_i - z_j + \hat{z}_i - z_i + z_j - \hat{z}_j| \geq \\
|z_i - z_j| - |\hat{z}_i - z_i| - |\hat{z}_j - z_j| &\geq \\
d - \lambda_n d - \lambda_n d &= (1 - 2\lambda_n)d
\end{aligned}$$

Also

$$\begin{aligned}
\left| \prod_{j \neq i} \frac{\hat{z}_i - z_j}{\hat{z}_i - \hat{z}_j} \right| &= \prod_{j \neq i} \left| \left( 1 + \frac{\hat{z}_j - z_j}{\hat{z}_i - \hat{z}_j} \right) \right| \leq \prod_{j \neq i} \left( 1 + \frac{|\hat{z}_j - z_j|}{|\hat{z}_i - \hat{z}_j|} \right) \leq \\
\prod \left( 1 + \frac{\lambda_n d}{(1 - 2\lambda_n)d} \right) &= \\
\left( 1 + \frac{\lambda_n}{1 - 2\lambda_n} \right)^{n-1} & \quad Q.E.D.
\end{aligned}$$

In the case of the WDK method

$$C_i^{(k)} = W_i^{(k)}, \text{ i.e. } F_i = \prod_{j \neq i} (z_i - z_j) \quad (4.48)$$

Below we will denote  $z_i^{(k)}$  etc just by  $z_i$  etc and  $z_i^{(k+1)}$  etc by  $\hat{z}_i$  etc. Also  $w = \max |W_i^{(k)}|$  and  $\hat{w} = \max |W_i^{(k+1)}|$ . We need another lemma, namely **LEMMA 4.2.3.** Let

$$w \leq c_n d \quad (4.49)$$

$$c_n \in (0, 0.5) \quad (4.50)$$

$$\delta_n \leq 1 - 2c_n \quad (4.51)$$

where  $\delta_n$  is defined as

$$\frac{(n-1)c_n}{1-c_n} \left( 1 + \frac{c_n}{1-2c_n} \right)^{n-1} \quad (4.52)$$

Then

$$|\hat{W}_i| \leq \delta_n |W_i| \quad (4.53)$$

and

$$\hat{w} \leq c_n \hat{d} \quad (4.54)$$



**PROOF** Let  $\lambda_n = c_n$ . We have

$$|\hat{z}_i - z_i| = |W_i| \leq w \leq c_n d \quad (4.55)$$

But this is 4.44 with  $\lambda_n = c_n$ . Hence by Lemma 4.2.2 we have

$$|\hat{z}_i - z_j| \geq (1 - c_n)d \quad (4.56)$$

and

$$|\hat{z}_i - \hat{z}_j| \geq (1 - 2c_n)d \quad (4.57)$$

Now the WDK iteration is  $\hat{z}_i = z_i - W_i$ , leading to

$$\frac{W_i}{\hat{z}_i - z_i} = -1 \quad (4.58)$$

so

$$\sum_{j=1}^n \frac{W_j}{\hat{z}_i - z_j} + 1 = \frac{W_i}{\hat{z}_i - z_i} + \sum_{j \neq i} \frac{W_j}{\hat{z}_i - z_j} + 1 = \sum_{j \neq i} \frac{W_j}{\hat{z}_i - z_j} \quad (4.59)$$

But Lagrange's interpolation formula for  $P(z)$  based on the points  $z_1, \dots, z_n$  and  $\infty$  gives

$$P(z) = \left( \sum_{j=1}^n \frac{W_j}{z - z_j} + 1 \right) \prod_{j=1}^n (z - z_j) \quad (4.60)$$

Letting  $z = \hat{z}_i$  and using 4.59 we obtain

$$P(\hat{z}_i) = (\hat{z}_i - z_i) \left( \sum_{j \neq i} \frac{W_j}{\hat{z}_i - z_j} \right) \prod_{j \neq i} (\hat{z}_i - z_j) \quad (4.61)$$

Dividing by  $\prod_{j \neq i} (\hat{z}_i - \hat{z}_j)$  gives

$$\begin{aligned} \hat{W}_i &= \frac{P(\hat{z}_i)}{\prod_{j \neq i} (\hat{z}_i - \hat{z}_j)} = \\ &(\hat{z}_i - z_i) \left( \sum_{j \neq i} \frac{W_j}{\hat{z}_i - z_j} \right) \prod_{j \neq i} \left( 1 + \frac{\hat{z}_j - z_j}{\hat{z}_i - \hat{z}_j} \right) \end{aligned} \quad (4.62)$$

Hence  $|\hat{W}_i| \leq |\hat{z}_i - z_i| \sum_{j \neq i} \frac{|W_j|}{|\hat{z}_i - z_j|} \prod_{j \neq i} \left( 1 + \frac{|\hat{z}_j - z_j|}{|\hat{z}_i - \hat{z}_j|} \right)$  (by 4.55, 4.56, 4.57)

$$\leq |W_i| \frac{(n-1)w}{(1-c_n)d} \left( 1 + \frac{c_n d}{(1-2c_n)d} \right)^{n-1}$$

$$\leq |W_i| \frac{(n-1)c_n}{1-c_n} \left(1 + \frac{c_n}{1-2c_n}\right)^{n-1}$$

=  $\delta_n |W_i|$  (with  $\delta_n$  given by 4.52). This is 4.53.

Now since  $\hat{d} = \min_{1 \leq i, j, \leq n; i \neq j} |\hat{z}_i - \hat{z}_j|$  by 4.57 we get

$$\hat{d} \geq (1 - 2c_n)d, \text{ i.e.} \quad (4.63)$$

$$d \leq \frac{\hat{d}}{1 - 2c_n} \quad (4.64)$$

and hence using 4.51

$$|\hat{W}_i| \leq \delta_n |W_i| \leq \delta_n c_n d \leq \frac{\delta_n c_n}{1 - 2c_n} \hat{d} \leq c_n \hat{d}$$

This gives 4.54. Q.E.D.

Now we have

**THEOREM 4.2.4** With the assumptions 4.49, 4.50, and 4.51 and  $z_1^{(0)}, \dots, z_n^{(0)}$  as initial approximations for which

$$w^{(0)} \leq c_n d^{(0)} \quad (4.65)$$

then the WDK method converges.

**PROOF** We will take  $C_i^{(k)} = W_i^{(k)}$  in Theorem 4.2.1 and seek to prove 4.34 and 4.35. Now by 4.54 and 4.65 we conclude that

$$w^{(1)} \leq c_n d^{(1)} \quad (4.66)$$

Similarly  $w^{(k)} \leq c_n d^{(k)}$  implies

$$w^{(k+1)} \leq c_n d^{(k+1)} \quad (4.67)$$

and so by induction 4.67 is true for all k. Hence by 4.53

$$|W_i^{(k+1)}| \leq \delta_n |W_i^{(k)}| = \beta_n |W_i^{(k)}| \quad (4.68)$$

(Note that as  $C_i = W_i$  in this case 4.36 and 4.37 are identical i.e.  $\delta_n = \beta_n$  with  $\delta_n$  given by 4.52). So 4.34 is true.

Similarly to the derivation of 4.57 we can show that  $|z_i^{(k+1)} - z_j^{(k+1)}| \geq (1 - 2c_n)d > 0$  so that  $F_i(z_1^{(k)}, \dots, z_n^{(k)}) = \prod_{i \neq j} (z_i^{(k)} - z_j^{(k)}) \neq 0$  in each iteration, and hence the WDK method is well-defined.

Now  $\beta_n = \delta_n < 1$  by 4.51, i.e. condition 4.41 is true. Then if  $\beta_n \geq \frac{1}{2}$ , 4.42 becomes  $\frac{1}{1-\beta_n} < \frac{1}{2\lambda_n}$  which is equivalent to 4.51. If  $\beta_n < \frac{1}{2}$ , then 4.42 reduces to

$$1 + 2\beta_n < \frac{1}{2\lambda_n}, \text{ i.e. } \lambda_n = c_n < \frac{1}{2(1 + 2\beta_n)} \in (.25, .5)$$

which is true by 4.50. Also,  $\lambda_n = c_n$  and  $C_i^{(k)} = W_i^{(k)}$  (which is 4.38) is automatically satisfied. Hence 4.35 is true, i.e. by Theorem 4.2.1 the WDK method converges. Finally, if we take

$$c_n = \frac{1}{1.76325n + .8689426} = \frac{1}{An + B} \quad (4.69)$$

we have  $c_n \leq c_3 \approx 0.16238$ , so  $c_n \in (0, 0.5)$  i.e. 4.50 is true.

Let us define

$$\eta_n = \frac{\delta_n}{1 - 2c_n} = \frac{(n-1)c_n}{(1-c_n)(1-2c_n)} \left(1 + \frac{c_n}{1-2c_n}\right)^{n-1} \quad (4.70)$$

But  $\lim_{n \rightarrow \infty} \frac{(n-1)c_n}{1-2c_n} = \frac{1}{A}$ ,  $\lim_{n \rightarrow \infty} \frac{1}{1-c_n} = 1$ , and

$$\lim_{n \rightarrow \infty} \left(1 + \frac{c_n}{1-2c_n}\right)^{n-1} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{An}\right)^n = e^{\frac{1}{A}} \quad (4.71)$$

Hence

$$\lim_{n \rightarrow \infty} \eta_n = \frac{1}{A} e^{\frac{1}{A}} < .99998 < 1 \quad (4.72)$$

and since  $\eta_n$  is monotonically increasing we have  $\delta_n < 1 - 2c_n$  which is 4.51.

Thus the conditions of Theorem 4.2.4 are satisfied and WDK converges for those values of A and B.

A slight variation on 4.28 and 4.30 is to write

$$\sum_{i=1}^n |W_i^{(0)}| \leq \Omega_n d^{(0)} \quad (4.73)$$

where we may take

$$\Omega_n = nc_n \quad (4.74)$$

The  $c_n$  derived in the last several pages (based on Petkovic and Herceg (2001)) gives

$$\begin{aligned} \Omega_3 = 3c_3 = .48712, \Omega_n \in (4c_n, \lim_{n \rightarrow \infty} nc_n) = (.50493, \frac{1}{A}) \approx \\ (.505, .567) \quad (n \geq 4) \end{aligned} \quad (4.75)$$

Earlier authors give lower ranges, e.g. Zhao and Wang (1993) give

$$\Omega_n \in (.171, .257) \quad (4.76)$$

and later in Wang and Zhao (1995) they improved this to

$$\Omega_n \in (.204, .324) \quad (4.77)$$

while Batra (1998) gave  $\Omega_n = .5$ . It can be shown that Petkovic and Herceg's range is the best so far (although not much better than Batra's).

Petkovic also shows that the Borsch-Supan or Nourein's method 4.22 converges certainly if

$$c_n = \frac{1}{n+9/2} \quad (n=3,4); \quad = \frac{1}{309n/200+5} \quad (n \geq 5) \quad (4.78)$$

He also points out, quoting Carstensen (1993), that 4.22 and the Ehrlich-Aberth method 4.17 are mathematically equivalent, so that the same conditions for convergence apply. The proof of this equivalence is as follows: The Lagrange interpolation formula for  $P(t)$  at the points  $(z_1, \dots, z_n, \infty)$  is

$$P(t) = \left( \sum_{j=1}^n \frac{W_j}{t-z_j} + 1 \right) \prod_{j=1}^n (t-z_j) \quad (4.79)$$

Taking the logarithmic derivative of both sides gives

$$\frac{P'(t)}{P(t)} = \sum_{j \neq i} \frac{1}{t-z_j} + \frac{\sum_{j \neq i} \frac{W_j}{t-z_j} + 1 - (t-z_i) \sum_{j \neq i} \frac{W_j}{(t-z_j)^2}}{W_i + (t-z_i) \left( \sum_{j \neq i} \frac{W_j}{t-z_j} + 1 \right)} \quad (4.80)$$

Setting  $t = z_i$  gives

$$\frac{P'(z_i)}{P(z_i)} = \sum_{j \neq i} \frac{1}{z_i - z_j} + \frac{\sum_{j \neq i} \frac{W_j}{z_i - z_j} + 1}{W_i} \quad (4.81)$$

from which the stated equivalence follows.

### 4.3 Multiple Roots

Most methods, such as Newton's and the WDK method, converge only linearly to multiple roots (the ones mentioned converge quadratically to simple roots). Several authors have described modifications which improve the speed of convergence to multiple roots, or clusters of roots. Several of these also determine the multiplicity, or effective multiplicity (the number in the cluster).

We will explain in detail the work by Hull and Mathon (1996). Suppose  $\zeta$  is a root of multiplicity  $m$ , and that  $\zeta_1 = \zeta_2 = \dots = \zeta_m = \zeta \neq \zeta_{m+1}, \dots, \zeta_n$ . The WDK formula gives

$$\hat{z}_j - \zeta = z_j - \zeta - A_j \prod_{k=m+1}^n \left( \frac{z_j - \zeta_k}{z_j - z_k} \right) \quad (4.82)$$

where

$$A_j = \frac{(z_j - \zeta)^m}{\prod_{k=1, \neq j}^m (z_j - z_k)} \quad (4.83)$$

Let  $\epsilon = \max_{1 \leq j \leq n} |z_j - \zeta_j|$ . Then as in 4.9 the product in 4.82 is  $1 + O(\epsilon)$ , so

$$\hat{z}_j - \zeta = z_j - \zeta - A_j(1 + O(\epsilon)) \quad (4.84)$$

Now assume that the method converges at least linearly, so that  $\hat{z}_j - \zeta = O(\epsilon)$ , then  $A_j = O(\epsilon)$  and

$$\hat{z}_j - \zeta = z_j - \zeta - A_j + O(\epsilon^2) \quad (4.85)$$

Summing for  $j = 1$  to  $m$  and dividing by  $m$  gives

$$\begin{aligned} \sum_{j=1}^m \hat{z}_j / m - \zeta &= \sum_{j=1}^m (z_j - \zeta - A_j) + O(\epsilon^2) \\ &= \sum_{j=1}^m \left( w_j - \frac{w_j^m}{\prod_{k=1, \neq j}^m (w_j - w_k)} \right) + O(\epsilon^2) \end{aligned} \quad (4.86)$$

where  $w_j = z_j - \zeta$ . But the sum on the RHS above is precisely the sum of the approximations which we would get by applying the WDK formula to  $w^m = 0$ , and by 4.16 this sum  $= \frac{-c_{n-1}}{c_n} = 0$ . Hence

$$\sum_{j=1}^m \hat{z}_j/m - \zeta = O(\epsilon^2) \quad (4.87)$$

i.e. the convergence of the mean is quadratic.

It will be shown that near a multiple root the approximations behave in a special way. For suppose

$$P(z) = (z - \zeta)^m F(z) \quad (4.88)$$

where  $F(\zeta) \neq 0$ . Then if

$$R(z) = c_n \prod_{j=1}^n (z - \hat{z}_j) \quad (4.89)$$

we may write

$$R(z) = P(z) + p(z) \quad (4.90)$$

where  $p(z)$  is a perturbation of order  $(\hat{z}_j - \zeta)$ . For  $z = \hat{z}_j$  ( $j = 1, \dots, m$ ) we have  $R(\hat{z}_j) = 0$  and so

$$(\hat{z}_j - \zeta)^m F(\hat{z}_j) + p(\hat{z}_j) = 0 \quad (4.91)$$

Hence the  $\hat{z}_j$  which approximate  $\zeta$  can be written

$$\hat{z}_j = \zeta + \left( \frac{-p(\hat{z}_j)}{F(\hat{z}_j)} \right)^{\frac{1}{m}} \quad (4.92)$$

Thus the approximations to a multiple root tend to be spaced uniformly around a circle with centre at the root, and hence it is likely that the mean of these approximations is a good approximation to the root.

Now assume

$$z_j = \zeta + \delta e^{(\frac{2\pi j}{m} + \gamma)i} (1 + O(\epsilon)) \quad (4.93)$$

then as above  $\hat{z}_j - \zeta = z_j - \zeta - A_j(1 + O(\epsilon))$  where now

$$\begin{aligned} A_j &= (z_j - \zeta) \prod_{k=1, \neq j}^m \frac{(z_j - \zeta)}{z_j - z_k} \\ &= (z_j - \zeta) \prod \frac{\delta e^{(\frac{2\pi j}{m} + \gamma)i}}{\delta e^{(\frac{2\pi j}{m} + \gamma)i} - \delta e^{(\frac{2\pi k}{m} + \gamma)i}} (1 + O(\epsilon)) \\ &= (z_j - \zeta) \prod_{k=1, \neq j}^m \frac{1}{1 - e^{\frac{2\pi(k-j)}{m}i}} (1 + O(\epsilon)) \end{aligned}$$

$$\begin{aligned}
&= (z_j - \zeta) \prod_{k=1}^{m-1} \frac{1}{1 - e^{(\frac{2\pi k}{m})i}} (1 + O(\epsilon)) \\
&= (z_j - \zeta) \frac{1}{m} (1 + O(\epsilon))
\end{aligned} \tag{4.94}$$

since  $\prod_{k=1}^{m-1} (1 - e^{(\frac{2\pi k}{m})i}) = \text{Lim}_{z \rightarrow 1} \left( \frac{z^m - 1}{z - 1} \right) = m$

Hence, for  $j = 1, \dots, m$

$$\begin{aligned}
\hat{z}_j - \zeta &= (z_j - \zeta) \left(1 - \frac{1}{m} (1 + O(\epsilon))(1 + O(\epsilon))\right) \\
&= \frac{m-1}{m} (z_j - \zeta) + O(\epsilon^2)
\end{aligned} \tag{4.95}$$

Thus the distribution of the new approximations is the same as before, with a slightly reduced radius, i.e. the *individual* approximations converge linearly. But fortunately the mean  $\frac{\sum_{j=1}^m (z_j - \zeta)}{m} = \frac{m-1}{m^2} \delta e^{\gamma i} \sum_{j=1}^m e^{\frac{2\pi j i}{m}} + O(\epsilon^2)$  and the sum on the right = sum of  $m$ 'th roots of unity = 0, i.e.

$$\frac{1}{m} \sum_{j=1}^m \hat{z}_j = \zeta + O(\epsilon^2) \tag{4.96}$$

i.e. convergence of the mean is quadratic.

Hull and Mathon obtain error bounds thus: by the theory of partial fractions (see e.g. Maxwell (1960)), with  $Q(z) = \prod_{k=1}^n (z - z_k)$  we have

$$\frac{P(z)}{Q(z)} = 1 - \sum_{k=1}^n \frac{\Delta z_k}{z - z_k} \tag{4.97}$$

Putting  $z = \zeta_j$  gives

$$1 = \sum_{k=1}^n \frac{\Delta z_k}{\zeta_j - z_k}$$

and hence  $1 \leq \sum \frac{|\Delta z_k|}{|\zeta_j - z_k|}$

The maximum term in the sum occurs at a particular  $k$ , and for that  $k$

$$\begin{aligned}
1 &\leq n \left| \frac{\Delta z_k}{\zeta_j - z_k} \right|, \text{ i.e.} \\
|z_k - \zeta_j| &\leq n |\Delta z_k|
\end{aligned} \tag{4.98}$$

In a similar manner we can show that

$$|z_k - \zeta_j| \leq \sum_{k=1}^n |\Delta z_k| \tag{4.99}$$

We may take the least of these two bounds, and the bounds together with the  $z_k$  define  $n$  disks. Isolated disks correspond to simple roots, while the others form clusters. A collection of approximations is considered to be a cluster if the union of their discs form a continuous region which is disjoint from all the other discs. As successive iterations proceed, these regions get smaller, and we need a criterion to determine if convergence has occurred. This is taken to be the case (for a cluster of multiplicity  $m$ ) if the estimated error bounds in evaluating  $P(z)$ ,  $P'(z)$ , ...,  $P^{(m-1)}(z)$  are greater than the calculated values of the polynomial and its derivatives. The error can be estimated by the method of Peters and Wilkinson (1971). An experimental program gave results slightly more accurate than the NAG and IMSL routines, and at about the same speed. But the method described here has the great advantage that it is suited for parallelization.

Miyakoda (1989, 1992, 1993) takes account of multiplicity in a slightly different way. Considering 4.93 and 4.95 he deduces the following, for approximations  $i$  and  $j$  belonging to the same cluster and with corrections  $\Delta z_i$  and  $\Delta z_j$ :

- (a)  $m$  approximations are situated on a circle centering on the root and are separated by an angle  $2\pi/m$ .
- (b) Every correction is directed towards the center and has almost the same magnitude.
- (c) The distance between new approximations is much smaller than that between the old ones.

Using the above criteria (a) and (b) we are led to:

$$\begin{aligned}
 (i) \quad & 1 - \alpha < \frac{|\Delta z_i|}{|\Delta z_j|} < 1 + \alpha \\
 (ii) \quad & \cos \theta_{ij} = \frac{(z_j - z_i, \Delta z_i)}{|z_j - z_i| |\Delta z_i|} > 0 \\
 & \cos \theta_{ji} = \frac{(z_i - z_j, \Delta z_j)}{|z_i - z_j| |\Delta z_j|} > 0 \\
 (iii) \quad & |\cos \theta_{ij} - \cos \theta_{ji}| < \beta
 \end{aligned} \tag{4.100}$$

where  $\alpha$  is a small number, and  $\beta$  is also small. When we have determined that the number of roots in a cluster is  $m$ , we make a correction for each of them equal to  $m\Delta z_i$  ( $i = 1, \dots, m$ ). Then (c) gives:

$$(iv) \quad |z_i + m\Delta z_i - z_j - m\Delta z_j| < \gamma |z_i - z_j| \tag{4.101}$$

where as usual  $\gamma$  is small. It is suggested that  $\alpha$ ,  $\beta$ , and  $\gamma$  be about .1. Miyakoda (1993) describes an algorithm based on the above, to determine the multiplicity. It starts by sorting the approximations in order of magnitude of corrections, and adds on to a cluster when all the equations of 4.100 are satisfied. He points out that the use of corrections  $m\Delta z_i$  may lead to a miscalculation of the multiplicity, but the use of the mean value of approximations in a cluster speeds up the convergence and usually gives the correct multiplicities.



Fraignaud (1991) gives a similar treatment.

The above all refer to variations on the WDK method. Other methods for multiple roots were treated by Lo Cascio et al (1989) (method of Pasquini and Trigante), Iliev and Semerdzhiev (1999) (a method similar to Aberth's of order 3 derived by them), Carstensen (1993) (Aberth-like method of order 3), Gargantini (1980) (Square-root iteration usually of order 4), and Farmer and Loizou (1977) (order up to 5). The last-mentioned give, among other methods, a second-order variation on the WDK method, namely

$$\hat{z}_i = z_i - \left( \frac{P(z_i)}{\prod_{j=1, \neq i}^n (z_i - z_j)^{m_j}} \right)^{\frac{1}{m_i}} \quad (4.102)$$

where  $m_i$ ,  $m_j$  are the multiplicities of  $z_i$ ,  $z_j$  respectively. They suggest estimating the multiplicity of a root approximated by a sequence  $z_i$  by

$$\text{Lim}_{z_i \rightarrow \alpha_i} \frac{1}{u'(z_i)} \quad (4.103)$$

where

$$u(z) = \frac{P(z)}{P'(z)} \quad (4.104)$$

Petkovic and Stefanovic (1987) give a variation on 4.102 whereby  $z_j$  is replaced by  $z_j - m_j P(z_j)/P'(z_j)$ . This method is of order 3. They also explain how best to choose one of several  $m_i$ 'th roots in 4.102

Sakurai et al (1991) give a method which is of order 7 for simple roots and 3 for multiple roots. They claim that it is more efficient than Aberth's method in many cases.

#### 4.4 Use of Interval Arithmetic

In finding roots of polynomials it is of course necessary to have reliable bounds on the errors in the estimated solutions. One very satisfactory way of obtaining these is by the use of interval arithmetic. That is, we start with some disjoint disks or rectangles which contain the zeros, and apply some iterative method such that successively smaller intervals are generated, which are guaranteed to still contain the roots. Thus we will obtain good approximations with error bounds given by the radii or semi-diagonal of the smallest interval.

Most of the relevant literature deals with disks, as they are easier to invert than rectangles, so we will give a brief survey of complex disk arithmetic (for more details

see Gargantini and Henrici (1971) or Alefeld and Herzberger (1983)).

We denote a closed circular region or disk

$$Z = z : |z - c| \leq r \quad (4.105)$$

by  $\{c;r\}$ . Here we denote the center  $c$  by  $\text{mid } Z$  and the radius  $r$  by  $\text{rad } Z$ . We define

$$Z_1 \pm Z_2 = \{z_1 \pm z_2 : z_1 \in Z_1, z_2 \in Z_2\} \quad (4.106)$$

where  $Z_k = \text{disk } \{c_k; r_k\}$  ( $k=1,2$ ).

It may be shown that the above is also a disk, namely

$$\{c_1 \pm c_2; r_1 + r_2\} \quad (4.107)$$

Unfortunately, if we were to define

$$Z_1.Z_2 = \{z_1 z_2 : z_1 \in Z_1, z_2 \in Z_2\} \quad (4.108)$$

the result would not in general be a disk, so it is usual to use

$$Z_1.Z_2 = \{c_1 c_2; |c_1| r_2 + |c_2| r_1 + r_1 r_2\} \quad (4.109)$$

which is of course a disk by definition. Gargantini and Henrici show that the RHS of 4.109 contains that of 4.108, although the reverse is not generally true.

Now we will prove that (if  $Z$  does not contain 0)  $Z^{-1} = \{\frac{1}{z} : z \in Z\}$  is also a disk, namely

$$\left\{ \frac{\bar{c}}{|c|^2 - r^2}; \frac{r}{|c|^2 - r^2} \right\} \quad (4.110)$$

For the boundary of a disk with center  $c = a + ib$  and radius  $R$  (note upper case) has cartesian equation

$$(x - a)^2 + (y - b)^2 = R^2 \quad (4.111)$$

or  $x^2 - 2ax + y^2 - 2by + a^2 + b^2 - R^2 = 0$ , or in polar coordinates  $(r, \theta)$  (i.e. with  $z = re^{i\theta}$ )

$$r^2 - 2ar \cos\theta - 2br \sin\theta + C = 0 \quad (4.112)$$

where

$$C = a^2 + b^2 - R^2 = c\bar{c} - R^2 = |c|^2 - R^2 \quad (4.113)$$

But if we let  $w = \frac{1}{z}$  we have  $w = \frac{1}{r}e^{-i\theta}$  or  $w = (\rho, \phi)$  where  $\rho = \frac{1}{r}$ ,  $\phi = -\theta$  (or  $r = \frac{1}{\rho}$ ,  $\theta = -\phi$ ), and so the equation of the transformed boundary is  $\frac{1}{\rho^2} - \frac{2a}{\rho} \cos\phi + \frac{2b}{\rho} \sin\phi + C = 0$ , i.e.

$$\frac{1}{C} - \frac{2a}{C} \rho \cos\phi + \frac{2b}{C} \rho \sin\phi + \rho^2 = 0 \quad (4.114)$$

This is the equation of a circle with center

$$\frac{a - ib}{C} = \frac{\bar{c}}{C} \quad (4.115)$$

and radius  $\sqrt{(\frac{a}{C})^2 + (\frac{b}{C})^2 - \frac{1}{C}} = \sqrt{\frac{c\bar{c}}{C^2} - \frac{1}{C}} = \sqrt{\frac{c\bar{c} - (c\bar{c} - R^2)}{C^2}} = \frac{R}{C} =$

$$\frac{R}{c\bar{c} - R^2} \quad (4.116)$$

However in our original definition of Z we used lower case r for the radius, so replacing R by r above, 4.113, 4.115 and 4.116 establish 4.110.

Next we define

$$Z_1/Z_2 = Z_1 \cdot Z_2^{-1} \quad (4.117)$$

(provided  $Z_2$  does not contain 0.)

and, with  $c = |c|e^{i\phi}$  and  $|c| > r$

$$\begin{aligned} \{c; r\}^{\frac{1}{2}} &= \{\sqrt{|c|}e^{i(\phi/2)}; \sqrt{|c|} - \sqrt{|c| - r}\} \cup \\ &\{-\sqrt{|c|}e^{i(\phi/2)}; \sqrt{|c|} - \sqrt{|c| - r}\} \end{aligned} \quad (4.118)$$

For  $*$  = any of the basic operations +, -, ., / the inclusion property holds i.e.

$$Z_k \subseteq W_k \ (k = 1, 2) \implies Z_1 * Z_2 \subseteq W_1 * W_2 \quad (4.119)$$

Also, if F is a complex circular extension of f (i.e. each complex variable in f is replaced by a disk), then  $w_k \in W_k \ (k = 1, \dots, q; w_k \text{ a complex number})$  implies

$$f(w_1, \dots, w_q) \in F(W_1, \dots, W_q) \quad (4.120)$$

The question of when disks are contained in one another is answered by

$$\{c_1; r_1\} \subseteq \{c_2; r_2\} \text{ iff } |c_1 - c_2| \leq r_2 - r_1 \quad (4.121)$$

and  $Z_1, Z_2$  are disjoint if and only if

$$|c_1 - c_2| > r_1 + r_2 \quad (4.122)$$

We may extend 4.107 and 4.109 to more than 2 disks as follows:

$$Z_1 + Z_2 + \dots + Z_q = \{c_1 + \dots + c_q; r_1 + \dots + r_q\} \quad (4.123)$$

$$\prod_{i=1}^q Z_i = \left\{ \prod_{i=1}^q c_i; \prod_{i=1}^q (|c_i| + r_i) - \prod_{i=1}^q |c_i| \right\} \quad (4.124)$$

In the special case where  $Z_1 = \{z_1; 0\}$  (i.e. a single point, denoted just by  $z_1$ ) we have

$$z_1 \pm Z_2 = \{z_1 \pm c_2; r_2\} \quad (4.125)$$

$$z_1.Z_2 = \{z_1c_2; |z_1|r_2\} \quad (4.126)$$

4.118 may be generalized to the  $k$ 'th root thus:

$$Z^{\frac{1}{k}} = \bigcup_{j=0}^{k-1} \left\{ |c|^{\frac{1}{k}} \exp\left(\frac{\phi + 2j\pi}{k}i\right); |c|^{\frac{1}{k}} - (|c| - r)^{\frac{1}{k}} \right\} \quad (4.127)$$

Alefeld and Herzberger (1983) give a disk version of the WDK method: given initial disjoint disks  $Z_1^{(0)}, \dots, Z_n^{(0)}$  containing the simple zeros  $\zeta_1, \dots, \zeta_n$  respectively, let

$$Z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1; j \neq i}^n (z_i^{(k)} - Z_j^{(k)})} \quad (i = 1, \dots, n; k = 0, 1, \dots) \quad (4.128)$$

where  $z_i^{(k)} = \text{mid}(Z_i^{(k)})$ .

Then  $\zeta_i \in Z_i^{(k)}$  for all  $i, k$ . Petkovic, Carstensen and Trajkovic (1995) prove that if

$$\rho^{(0)} > \frac{7(n-1)}{2} r^{(0)} \quad (4.129)$$

where

$$\rho^{(0)} = \min_{i,j; i \neq j} \{|z_i^{(0)} - z_j^{(0)}| - r_j^{(0)}\} \quad (4.130)$$

$$r^{(0)} = \max_{1 \leq j \leq n} \text{rad} Z_j^{(0)} \quad (4.131)$$

and we apply 4.128, then for all  $i, k$

$$\zeta_i \in Z_i^{(k)}; \quad r^{(k+1)} < \frac{7(n-1)}{4(\rho^{(0)} - 5r^{(0)})} (r^{(k)})^2 \quad (4.132)$$

They recommend a hybrid method, in which the ordinary WDK method is applied  $M$  times, then we take disks

$$D_i^{(M-1)} = \{z_i^{(k)}; \frac{1}{4}|W(z_i^{(M-1)})|\} \quad (4.133)$$

where  $W$  = correction in last WDK step. They show that the  $D_i$  will contain the  $\zeta_i$ . Finally we take one step of 4.128, using  $D_i^{(M-1)}$  in place of  $Z_i^{(m)}$ .

They claim that this is considerably more efficient than the use of the pure disk iteration 4.128 throughout, but still provides a good error estimate.

Sun and Li (1999) combine two steps of the WDK disk method, but with only one function evaluation, i.e. they let:

$$U_i^{(k)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - Z_j^{(k)})} \quad (4.134)$$

$$Z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - U_j^{(k)})} \quad (4.135)$$

They prove that if

$$\rho^{(0)} \geq 2.5(n-1)r^{(0)} \quad (4.136)$$

then the  $z_i^{(k)}$  converge to the  $\zeta_i$  with  $r_i^{(k)}$  tending to zero with order 3. As this method involves the equivalent of about 3 function evaluations per iteration, its efficiency index is about  $\log(\sqrt[3]{3}) = .1590$ , which is a little better than the  $\log(\sqrt{2}) = .150$  of the pure WDK -disk method.

Gargantini and Henrici (1971) give a disk-version of the Ehrlich-Aberth method 4.17 i.e.

$$Z_i^{(k+1)} = z_i^{(k)} - \frac{1}{\frac{P'(z_i^{(k)})}{P(z_i^{(k)})} - \sum_{j=1, \neq i}^n \frac{1}{z_i^{(k)} - Z_j^{(k)}}} \quad (4.137)$$

They show that this converges cubically. Gargantini (1978) gives a version of the above for multiple roots (of known multiplicity). With the usual definitions of  $\rho^{(0)}$  and  $r^{(0)}$ , and with  $m_i$  as the multiplicity of  $\zeta_i$  ( $i = 1, \dots, p \leq n$ ), she defines

$$\nu = \min_{1 \leq i \leq p} m_i; \quad \bar{\nu} = \max m_i; \quad \gamma = \frac{\bar{\nu}(n - \nu)}{\nu} \quad (4.138)$$

Then she lets

$$Z_i^{(k+1)} = z_i^{(k)} - \frac{m_i}{\frac{P'(z_i^{(k)})}{P(z_i^{(k)})} - \sum_{j=1, \neq i}^p \frac{m_j}{z_i^{(k)} - Z_j^{(k)}}} \quad (4.139)$$

She shows that if

$$6\gamma r^{(0)} < \rho^{(0)} \quad (4.140)$$

then the  $r^{(k)}$  tend to zero, with

$$r^{(k+1)} < \frac{3\gamma(r^{(k)})^3}{\nu(\rho^{(0)})^2} \quad (4.141)$$

Petkovic (1982) similarly gives a disk version of the Borsch-Supan method 4.22, namely (with the usual notation)

$$Z_i^{(k+1)} = z_i^{(k)} - \frac{W_i}{1 - \sum_{j=1, \neq i}^n \frac{W_j}{z_j^{(k)} - Z_i^{(k)}}} \quad (4.142)$$

where  $W_i$  = the WDK correction in 4.24. They show that if

$$\rho^{(0)} > 3(n-1)r^{(0)} \quad (4.143)$$

then the method converges with order 3. The same author, with Carstensen (1993) gives a disk version of Nourein's method 4.23, thus:

$$Z_i^{(k+1)} = z_i^{(k)} - \frac{W_i}{1 - \sum_{j=1, \neq i}^n W_j \text{INV}(z_j^{(k)} - Z_i^{(k)} + W_i)} \quad (4.144)$$

where INV may be the usual inverse given by 4.110 (referred to as  $I$ ) or  $I_2$  given by

$$\{c; r\}^{I_2} = \left\{ \frac{1}{c}; \frac{2r}{|c|^2 - r^2} \right\} \quad (4.145)$$

They show that if

$$d^{(0)} \geq 4(n-1)r^{(0)} \quad (4.146)$$

where

$$d^{(0)} = \min_{i,j=1,\dots,n; i \neq j} |z_i^{(0)} - z_j^{(0)}| \quad (4.147)$$

then the  $z_i^{(k)}$  converge to the roots, with order  $\frac{3+\sqrt{17}}{2} = 3.562$  if  $I$  is used, or order 4 if  $I_2$  is used. On the other hand the inclusion disks produced by  $I_2$  are often larger than those from  $I$ , so the authors conclude that  $I$  is best.

Carstensen and Petkovic (1993) describe a hybrid version using  $M$  iterations of the point Nourein method 4.23 followed by one disk iteration using 4.142. It is not clear why they do not use 4.144 in the disk iteration. They claim that the hybrid method is more robust than the pure disk method, and also it is more efficient as it uses the less expensive point iterations most of the time, whereas the final disk iteration gives a good error estimate.

Petkovic and Stefanovic (1986A) use rectangular intervals instead of disks with the WDK method, i.e. 4.128, where the  $Z_j^{(k)}$  are disjoint rectangles each containing one of the roots. That is

$$Z = I_1 + iI_2 \quad (4.148)$$

where  $I_1$  and  $I_2$  are real intervals e.g.

$$I_1 = \{x : a_1 \leq x \leq b_1\} \quad (4.149)$$

with

$$z_j^{(k)} = \text{mid}(I_1) + i \text{mid}(I_2) \quad (4.150)$$

and  $r_j^{(k)} = \text{sd}(Z_j^{(k)})$  (semidiagonal).

They show that if

$$\rho^{(0)} > 4(n-1)r^{(0)} \quad (4.151)$$

(with  $\rho^{(0)}$  and  $r^{(0)}$  more or less as usual, i.e. given by 4.130 and 4.131) then their version of 4.128 converges quadratically.

Finally Markov and Kjurkchiev (1989) describe an interval method for the case where all the roots are real and distinct. They suppose that the disjoint real intervals

$$X_i^{(0)} = [\underline{x}_i^{(0)}, \overline{x}_i^{(0)}] \quad (4.152)$$

each contain a root  $\zeta_i$  (presumably these intervals could be found by Sturm's sequences). Then a variation on the WDK method is used thus:

$$\underline{x}_i^{(k+1)} = \underline{x}_i^{(k)} - \frac{P(\underline{x}_i^{(k)})}{\prod_{j=1}^{i-1} (\underline{x}_i^{(k)} - \underline{x}_j^{(k)}) \prod_{j=i+1}^n (\underline{x}_i^{(k)} - \overline{x}_j^{(k)})} \quad (i = 1, \dots, n; k = 0, 1, \dots) \quad (4.153)$$

with a similar formula for  $\overline{x}_i^{(k+1)}$ . They show that with suitable conditions convergence is quadratic (as are most variations on the WDK method).

## 4.5 Recursive Methods

These are exemplified in Atanassova (1996), where the following method (generalizing the WDK method) is described:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - z_j^{(k)} - \Delta_j^{R,k})} \quad (i = 1, \dots, n) \quad (4.154)$$

$$\Delta_i^{p,k} = - \frac{P(z_i^{(k)})}{\prod_{s=1, \neq i}^n (z_i^{(k)} - z_s^{(k)} - \Delta_s^{p-1,k})} \quad (i = 1, \dots, n; p = 1, \dots, R) \quad (4.155)$$

$$\Delta_i^{0,k} = 0 \tag{4.156}$$

Of course the above 3 formulas are repeated for  $k = 0, 1, \dots$  until convergence. He shows that the rate of convergence is of order  $R+2$ .  $R=0$  gives the WDK method, while  $R=1$  gives Nourein's method 4.25. Since the method requires  $R+2$  function evaluations or equivalent per root, the efficiency is

$$\log((R + 2)^{\frac{1}{R+2}}) \tag{4.157}$$

which is a maximum for integer  $R = 1$  (i.e. 4.25), reaching the value .1590. Note that Atanassova states that the above was first given by Kjurkchiev and Andreev (1985).

Another example, generalizing the Borsch-Supan method 4.22, is given by Carstensen and Petkovic (1993), namely:

$$Z_i^{(k + \frac{\lambda+1}{R})} = z_i^{(k)} - \frac{W_i}{1 + \sum_{j=1, \neq i}^n \left( \frac{W_j}{Z_i^{(k+\lambda/R)} - z_j^{(k)}} \right)}$$

$$(\lambda = 0, \dots, R - 1, i = 1, \dots, n, k = 0, 1, \dots) \tag{4.158}$$

Here the  $Z_i$  are disks as in section 4, but they could equally be points. The order of convergence is  $2R+1$ . For a given  $R$ , the number of equivalent function evaluations per root appears to be  $2+1.5R$ , so the efficiency is  $\log((2R+1)^{\frac{1}{2+1.5R}})$ . These values are tabulated below

R	1	2	3	4
eff	.136	.1398	.1300	.1193

Thus the most efficient method of this class would be for  $R = 2$ , but it is still not as efficient as 4.25.

Kjurkchiev and Andreev (1992) give a similar recursive version of Aberth's method.  $R$  recursions give order  $2R+3$  and require  $2+1.5(R+1)$  function evaluations per root. Again,  $R= 0$  (the ordinary Aberth's method) is most efficient at efficiency =  $\log(\sqrt[3.5]{3}) = .136$

Andreev and Kjurkchiev (1987) give a recursive version of 4.153, for simple real roots. As before it is expected that the interval version of 4.25 will be the most efficient of this class. They also give a recursive interval version of Aberth's method.



### 4.6 Methods Involving Square Roots, Second Order Derivatives, etc

Leuze (1983) derives a modified, simultaneous Laguerre-type method as follows: let

$$S_1(z) = \frac{P'(z)}{P(z)} = \sum_{i=1}^n \frac{1}{z - \zeta_i} \quad (4.159)$$

$$S_2(z) = -\frac{dS_1}{dz} = \frac{P'(z)^2 - P(z)P''(z)}{P(z)^2} = \sum_{i=1}^n \frac{1}{(z - \zeta_i)^2} \quad (4.160)$$

$$\alpha(z) = \frac{1}{z - \zeta_j}, \beta(z) + \gamma_i(z) = \frac{1}{z - \zeta_i} \quad (i = 1, \dots, n; i \neq j) \quad (4.161)$$

where

$$\beta(z) = \frac{1}{n-1} \sum_{i \neq j} \frac{1}{z - \zeta_i} \quad (4.162)$$

Hence

$$\sum_{i=1, \neq j}^n \gamma_i = 0 \quad (4.163)$$

Also define

$$\delta^2 = \sum_{i=1, \neq j}^n \gamma_i^2 \quad (4.164)$$

Hence we have

$$S_1 = \alpha + (n-1)\beta, S_2 = \alpha^2 + (n-1)\beta^2 + \delta^2 \quad (4.165)$$

(For  $S_2 = \frac{1}{(z-\zeta_j)^2} + \sum_{i \neq j} \frac{1}{(z-\zeta_i)^2} = \alpha^2 + \sum_{i \neq j} (\beta + \gamma_i)^2 = \alpha^2 + (n-1)\beta^2 + 2\beta \sum_{i \neq j} \gamma_i + \sum_{i \neq j} \gamma_i^2$ )

Eliminating  $\beta$  and solving the resulting quadratic in  $\alpha$  gives

$$\alpha = \frac{1}{n} [S_1 \pm \sqrt{(n-1)(nS_2 - S_1^2 - n\delta^2)}] \quad (4.166)$$

Since  $\alpha = \frac{1}{z - \zeta_j}$  this gives

$$\zeta_j = z - \frac{n}{S_1 \pm \sqrt{(n-1)(nS_2 - S_1^2 - n\delta^2)}} \quad (4.167)$$

where  $S_1$ ,  $S_2$  and  $\delta$  are evaluated at  $z$ . (Note that dropping  $\delta^2$  gives the classical Laguerre iteration).

Substituting  $\beta$  from 4.162 into  $S_2$  given by 4.165 and solving for  $\delta^2$  gives

$$\delta^2 = \sum_{i=1, \neq j}^n \left[ \frac{1}{z - \zeta_i} - \beta \right]^2 \tag{4.168}$$

$$\begin{aligned} & \text{(For the RHS of the above equation} = \sum_{i \neq j} \left[ \frac{1}{(z - \zeta_i)^2} - 2\beta \frac{1}{z - \zeta_i} + \beta^2 \right] \\ &= \sum_{i \neq j} \frac{1}{(z - \zeta_i)^2} - 2\beta \sum_{i \neq j} \frac{1}{z - \zeta_i} + (n - 1)\beta^2 \\ &= \sum_{i=1}^n \frac{1}{(z - \zeta_i)^2} - \frac{1}{(z - \zeta_j)^2} - 2\beta(n - 1)\beta + (n - 1)\beta^2 \\ &= S_2 - \alpha^2 - (n - 1)\beta^2 = \delta^2 \end{aligned}$$

We approximate  $\delta^2$  by

$$\delta_j^2 = \sum_{i=1, \neq j}^n \left( \frac{1}{z_j^{(k)} - z_i^{(k)}} - \beta_j \right)^2 \tag{4.169}$$

where

$$\beta_j = \frac{1}{n - 1} \sum_{i=1, \neq j}^n \frac{1}{z_j^{(k)} - z_i^{(k)}} \tag{4.170}$$

leading to the iteration

$$z_j^{(k+1)} = z_j^{(k)} - \frac{n}{S_1 \pm \sqrt{(n - 1)(nS_2 - S_1^2 - n\delta_j^2)}} \tag{4.171}$$

where  $S_1$  and  $S_2$  are evaluated at  $z_j^{(k)}$ , e.g.  $S_1 = \frac{P'(z_j^{(k)})}{P(z_j^{(k)})}$ . This is called the modified Laguerre method.

Petkovic, Petkovic and Ilic (2003) derive the same method and show that if

$$w^{(0)} \leq \frac{d^{(0)}}{3n} \tag{4.172}$$

and the roots are simple, then 4.171 converges with order 4. In practise the 3 in 4.172 may be dropped.

Leuze continues by showing that if  $z_l$  and  $z_m$  are two approximations close to each other but far from a root, then convergence may be greatly slowed down. To remedy this, a hybrid method is described as follows: a modified Laguerre step is taken except when it is found that two approximations are much closer to each other than to a root. If that occurs, a classical Laguerre step is taken by one approximation and a modified step by the other. This allows the rapid convergence of the modified method to resume. The criterion for closeness is that  $n\delta_j^2$  be greater

than  $nS_2 - S_1^2$ . For further details see the cited paper by Leuze. Numerous tests show that the hybrid method is much more robust than the modified method by itself.

Hansen, Patrick and Rusnack (1977) point out that the modified method requires nearly twice as many operations per iteration than the classical method. Their tests show that the classical method is more efficient when the starting values are fairly close to the roots. But if this is not the case Laguerre's method often converges to only a few of the roots, i.e. several approximations converge to the same (non-multiple) root. Modified Laguerre worked properly in all cases tested, even when the starting values were clustered about only one of the roots.

Petkovic (2003) gives a circular interval version of 4.171 as follows:

$$Z_i^{(k+1)} = Z_i^{(k)} - \frac{n}{S_1 + \sqrt{(n-1)(nS_2 - S_1^2 - Q_i^{(k)})}} \tag{4.173}$$

where

$$Q_i^{(k)} = n \sum_{j=1, \neq i}^n \left( \frac{1}{z_i^{(k)} - Z_j^{(k)}} \right)^2 - \frac{n}{n-1} \left( \sum_{j=1, \neq i}^n \frac{1}{z_i^{(k)} - Z_j^{(k)}} \right)^2 \tag{4.174}$$

and  $S_1, S_2$  are evaluated at  $z_i^{(k)}$ .

The disk to be chosen in the denominator of 4.173 is that whose center maximizes the modulus of that denominator. He shows that if  $\rho^{(0)} > 3(n-1)r^{(0)}$  (where  $\rho^{(0)}$  and  $r^{(0)}$  are given by 4.130 and 4.131) then for all  $i$  and  $k$

$$\zeta_i \in Z_i^{(k)} \text{ and } r^{(k+1)} < \frac{5(n-1)(r^{(k)})^4}{(\rho^{(0)} - \frac{5}{4}r^{(0)})^3} \tag{4.175}$$

Petkovic (1981) describes a disk interval  $k$ 'th root method as follows: let

$$H_k(z) = \frac{(-1)^{k-1} d^k}{(k-1)! dz^k} [\log P(z)] (k \geq 1) \tag{4.176}$$

$$= \sum_{j=1}^n \frac{1}{(z - \zeta_j)^k} = \frac{1}{(z - \zeta_i)^k} + \sum_{j=1, \neq i}^n \frac{1}{(z - \zeta_j)^k} \tag{4.177}$$

Hence

$$\zeta_i = z - \frac{1}{\left[ H_k(z) - \sum_{j=1, \neq i}^n \frac{1}{(z - \zeta_j)^k} \right]^{\frac{1}{k}}} \tag{4.178}$$

$$= z - \frac{1}{q_i^{\frac{1}{k}}} \text{ (say)} \tag{4.179}$$

The value of the k'th root above should be chosen, for  $z = z_i$ , to satisfy

$$\left| \frac{P'(z_i)}{P(z_i)} - q_i^{\frac{1}{k}} \right| \leq \frac{n-1}{\rho} \tag{4.180}$$

where

$$\rho > k(n-1)r \tag{4.181}$$

4.178 leads us to the disk iteration

$$Z_i^{(m+1)} = z_i^{(m)} - \frac{1}{\left[ H_k(z_i^{(m)}) - \sum_{j=1, \neq i}^n \left( \frac{1}{z_i^{(m)} - z_j^{(m)}} \right)^k \right]^{\frac{1}{k}}} \tag{4.182}$$

(Here the iteration index is taken as m because k is already taken for the k'th root). Petkovic shows that if

$$\rho^{(0)} > \beta(k, n)r^{(0)} \tag{4.183}$$

where

$$\begin{aligned} \beta(k, n) &= 2n & (k = 1) \\ &= k(n-1) & (k > 1) \end{aligned} \tag{4.184}$$

then the order of convergence is k+2, and the disk  $Z_i^{(m+1)}$  always contains  $\zeta_i$ .

Replacing  $Z_i$  by its center  $z_i$  gives a k'th root point iteration, also with order k+2. The case k=1 gives Aberth's method 4.17 having convergence order 3, while k = 2 gives a 4th order modification of Ostrowski's method namely

$$z_i^{(m+1)} = z_i^{(m)} - \frac{1}{\left[ H_2(z_i^{(m)}) - \sum_{j=1, \neq i}^n \frac{1}{(z_i^{(m)} - z_j^{(m)})^2} \right]^{\frac{1}{2}}} \tag{4.185}$$

(“Ostrowski's” (1970) method does not have the last term in the denominator).

Petkovic and Rancic (2004) show that if

$$w^{(0)} < \frac{5}{13n}d^{(0)} \tag{4.186}$$

and the zeros are simple, then 4.185 is guaranteed to converge with order 4, so that efficiency is  $\log(\sqrt[5]{4}) = .1204$ . Petkovic (1982) gives a modification of 4.182 for multiple roots. Petkovic and Stefanovic (1986B) modify 4.185 by replacing  $z_j^{(m)}$  in the denominator by

$$z_j^{(m)} - \frac{P(z_i^{(m)})}{P'(z_i^{(m)})} \tag{4.187}$$

They show that this has convergence order 5, and as it takes about 5 horners, its efficiency index is  $\log(\sqrt[5]{5}) = .139$

They also give another method replacing  $z_j^{(m)}$  by

$$z_j^{(m)} + \frac{2PP'}{PP'' - 2PP'^2} \tag{4.188}$$

where P etc are evaluated at  $z_j^{(m)}$ . This has convergence order 6, and hence efficiency index  $\log(\sqrt[5]{6}) = .156$

Petkovic and Vranic (2000) describe another disk method involving square roots, which they term "Euler-like". It is

$$Z_i^{(k)} = z_i^{(k)} - \frac{2W_i^{(k)}}{1 + g_i^{(k)} + \sqrt{(1 + g_i^{(k)})^2 + 4W_i^{(k)}S_i^{(k)}}} \tag{4.189}$$

where

$$W_i = \frac{P(z_i)}{\prod_{j=1, \neq i}^n (z_i - z_j)} \tag{4.190}$$

(the Weierstrass correction),

$$g_i = \sum_{j \neq i} \frac{W_j}{(z_i - z_j)}, \quad S_i = \sum_{j \neq i} \frac{W_j}{(z_i - z_j)(Z_i - z_j)} \tag{4.191}$$

They show that if

$$\rho^{(0)} > 4(n - 1)r^{(0)} \tag{4.192}$$

then convergence is guaranteed with order 4. Another variation which they describe is to replace  $Z_i$  in 4.191 by  $Z_i - W_i$ , and to use the so-called centered inversion

$$Z_i = \left\{ \frac{1}{c}; \frac{1}{|c|(|c| - r)} \right\} \tag{4.193}$$

for the inversion of  $(Z_i - W_i - z_j)$ . They claim that in this case convergence is of order 5. Indeed if

$$w^{(0)} < \frac{1}{5n}d^{(0)} \tag{4.194}$$

where

$$d^{(0)} = \min_{i,j;i \neq j} |z_i^{(0)} - z_j^{(0)}| \tag{4.195}$$

then convergence is guaranteed provided we start with initial disks

$$\left\{ z_i^{(0)}; \frac{5}{4}W(z_i^{(0)}) \right\} \quad (i = 1, \dots, n) \tag{4.196}$$

which contain the roots  $\zeta_i$  ( $i=1,\dots,n$ ).

Several papers give high-order variations on Halley's method

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P}{P' - \frac{PP''}{2P'}} \tag{4.197}$$

where P etc are evaluated at  $z_i^{(k)}$ . For example, Wang and Wu (1987) give a general formula

$$z_i^{(k+1)} = z_i^{(k)} - \frac{\frac{P}{P'}}{1 - \frac{PP''}{2P'^2} - \frac{1}{2} \left(\frac{P}{P'}\right)^2 (\sigma_1^2 + \sigma_2)} \tag{4.198}$$

where

$$\sigma_l = \sum_{j=1, \neq i}^n \frac{1}{(z_i^{(k)} - w_j^{(k)})^l} \quad (l = 1, 2) \tag{4.199}$$

and  $w_j^{(k)}$  may be replaced by various expressions, such as  $z_j^{(k)}$ , giving a method of convergence order 4. Or we may take

$$w_j^{(k)} = z_j^{(k)} - \frac{P}{P'} \tag{4.200}$$

which is of order 5.

Or again, with  $w_j^{(k)}$  = the Aberth right-hand-side

$$= z_j^{(k)} - \frac{1}{\frac{P'}{P} - \sum_{l=1, \neq i}^n \frac{1}{z_j^{(k)} - z_l^{(k)}}} \tag{4.201}$$

giving order 6. Moreover  $w_j^{(k)}$  given by Halley's formula 4.197 also gives order 6. Of this class of methods, the last is probably the most efficient (index  $\log(\sqrt[5.5]{6}) = .14147$ ).

Finally

$$w_j^{(k)} = z_j^{(k)} - \frac{1}{\frac{P'}{P} - \sum_{l=1, \neq j}^n \frac{1}{z_j^{(k)} - z_l^{(k)} + \frac{P(z_l^{(k)})}{P'(z_l^{(k)})}}} \tag{4.202}$$

gives order 7. Note that in the above  $P$ ,  $P'$  and  $P''$  are to be evaluated at  $z_j^{(k)}$  (unless otherwise indicated).

Petkovic (1989A) gives a disk Halley-like method, suitable for multiple roots  $\zeta_i$  ( $i=1,\dots,m \leq n$ ) of known multiplicity  $\mu_i$ , namely:

$$Z_i^{(k+1)} = z_i^{(k)} - \frac{1}{\frac{P'}{2P} \left(1 + \frac{1}{\mu_i}\right) - \frac{P''}{2P'} - \frac{P}{2P'} \left(\frac{(\sum_{1i})^2}{\mu_i} + \sum_{2i}\right)} \tag{4.203}$$

where

$$\Sigma_{li} = \sum_{j=1, \neq i}^n \frac{\mu_j}{(z_i^{(k)} - z_j^{(k)})^l} \quad (l = 1, 2) \quad (4.204)$$

Provided

$$\rho^{(0)} > 3(n - \min_{i=1}^m \mu_i) r^{(0)} \quad (4.205)$$

this converges with order 4.

Petkovic, Milovanovic and Stefanovic (1986) describe a variation on the square-root iteration 4.185 suitable for multiple roots, thus:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{(\mu_i)^{\frac{1}{2}}}{\left[ H_2(z_i^{(k)}) - \sum_{j=1, \neq i}^m \frac{\mu_j}{(z_i^{(k)} - z_j^{(k)})^2} \right]^{\frac{1}{2}}} \quad (4.206)$$

Let  $Q_i$  be the correction term in 4.206 (before square-rooting) and let  $w_i^{(1)}$  and  $w_i^{(2)}$  be the two values of the  $\sqrt{Q_i}$ . Then we should take that square root which minimizes

$$\left| \frac{P'}{P} - w_i^{(l)} \right| \quad (l = 1, 2) \quad (4.207)$$

They modify 4.206 further by replacing  $z_j^{(k)}$  by

$$z_j^{(k)} - \mu_j \frac{P}{P'} \quad (4.208)$$

or by

$$z_j^{(k)} + \frac{2}{\frac{P''}{P'} - (1 + \frac{1}{\mu_j}) \frac{P'}{P}} \quad (4.209)$$

of orders 5 and 6 respectively. The latter appears to have efficiency index  $\log(\sqrt[5.5]{6}) = .1415$

Petkovic, Stefanovic and Marjanovic (1992) give a series of methods for multiple zeros which do not require square or higher roots. The simplest is

$$z_i^{(k+1)} = z_i^{(k)} - \frac{2 \left[ S_{1i}(z_i^{(k)}) - \frac{P'}{P} \right]}{\frac{P''}{P} - \left( \frac{P'}{P} \right)^2 + S_{2i}(z_i^{(k)}) - \frac{1}{\mu_i} \left[ S_{1i}(z_i^{(k)}) - \frac{P'}{P} \right]^2} \quad (4.210)$$

where

$$S_{li}(z) = \sum_{j=1, \neq i}^m \frac{\mu_j}{(z - z_j^{(k)})^l} \quad (l = 1, 2) \quad (4.211)$$

Note that 4.211 is similar to 4.204.

Another method (of order 5) replaces  $z_j^{(k)}$  in 4.211 by

$$z_j^{(k)} - \frac{\mu_j P(z_j^{(k)})}{P'(z_j^{(k)})} \tag{4.212}$$

while a third (of order 6) replaces  $z_j^{(k)}$  by

$$z_j^{(k)} - \frac{2}{\left(1 + \frac{1}{\mu_j}\right) \frac{P'}{P} - \frac{P''}{P}} \tag{4.213}$$

According to Petkovic et al all these methods require the equivalent of 4.5 complex horners, although this author counts 5. Thus the efficiency index of the last-mentioned is about  $\log({}^4\sqrt[5]{6}) = .1729$ . They also describe Gauss-Seidel variations of slightly greater efficiency, which will be discussed in Sec. 8.

Hansen and Patrick (1977) give a one-parameter family of methods thus:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{(\alpha + 1)P}{\alpha P' \pm \sqrt{(P')^2 - (\alpha + 1)PP''}} \tag{4.214}$$

P etc being evaluated at  $z_i^{(k)}$ . For various values of  $\alpha$ , some well-known methods are obtained. Petkovic, Ilic and Trickovic (1997) give a simultaneous version, namely

$$z_i^{(k+1)} = z_i^{(k)} - \frac{(\alpha + 1)W_i}{\alpha(1 + G_{1i}) \pm \sqrt{(1 + G_{1i})^2 + 2(\alpha + 1)W_i G_{2i}}} \tag{4.215}$$

where  $W_i$  is the usual WDK correction at  $z_i^{(k)}$  and

$$G_{li} = \sum_{j=1, \neq i}^n \frac{W_j}{(z_i^{(k)} - z_j^{(k)})^l} \tag{4.216}$$

Setting  $\alpha = 0, 1, \frac{1}{n-1}, -1$  they obtain respectively what they call the Ostrowski-, Euler-, Laguerre, and Halley-like methods. The last one needs a limiting operation to obtain:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{W_i(1 + G_{1i})}{(1 + G_{1i})^2 + W_i G_{2i}} \tag{4.217}$$

For multiple zeros they obtain:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{\mu_i(\mu_i\alpha + 1)}{\mu_i\alpha\left(\frac{P'}{P} - S_{1i}\right) \pm \sqrt{\mu_i(\mu_i\alpha + 1)\left[\left(\frac{P'}{P}\right)^2 - \frac{P''}{P} - S_{2i}\right] - \mu_i\alpha\left(\frac{P'}{P} - S_{1i}\right)^2}} \tag{4.218}$$



where the  $S_{1i}$  are given by 4.211. Setting  $\alpha = 0$  gives:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{\sqrt{\mu_i}}{\sqrt{\left(\frac{P'}{P}\right)^2 - \frac{P''}{P} - S_{2i}}} \quad (4.219)$$

$\alpha = -\frac{1}{\mu_i}$  gives

$$z_i^{(k+1)} = z_i^{(k)} - \frac{2\mu_i\left(\frac{P'}{P} - S_{1i}\right)}{\left(\frac{P'}{P} - S_{1i}\right)^2 - \mu_i\left(\frac{P''}{P} - \left(\frac{P'}{P}\right)^2 + S_{2i}\right)} \quad (4.220)$$

while  $\alpha = \frac{1}{n-\mu_i}$  gives

$$z_i^{(k+1)} = z_i^{(k)} - \frac{n}{\left(\frac{P'}{P} - S_{1i}\right)\left(1 \pm \sqrt{\frac{n-\mu_i}{\mu_i} - n\left(1 + \frac{S_{2i} - \left(\frac{P'}{P}\right)^2 + S_{2i}}{\left(\frac{P'}{P} - S_{1i}\right)^2}\right)}\right)} \quad (4.221)$$

The authors prove that the above methods are all of order 4, and as they require about 5 horners the efficiency index is  $\log(\sqrt[5]{4})$ . Petkovic, Petkovic and Herceg (1998) give initial conditions which guarantee convergence, namely:

$$w^{(0)} < \frac{d^{(0)}}{3n+3} \quad (4.222)$$

while Petkovic and Herceg (2001) give the less stringent condition

$$w^{(0)} < \frac{2d^{(0)}}{5n-5} \quad (4.223)$$

Petkovic, Sakurai and Rancic (2004) derive a similar set of formulas based on Hansen-Patrick's method. It is not clear that these have any advantage over the previously-mentioned ones.

Other high-order methods are given by Farmer and Loizou (1975), Loizou (1983), Petkovic and Herceg (1998) and Sun and Zhang (2003), to mention a few examples. For further details, see the cited papers.

## 4.7 Effect of Rounding Errors

As in any process which uses floating-point arithmetic, all the above methods are affected by rounding errors, and this is considered by several authors. For example Gargantini (1978) discusses the effect of rounding error upon the disk-Ehrlich-Aberth method 4.137. She points out that iterative methods are usually designed to terminate when the rounding error from the evaluation of  $P$  near a zero is of the same order of magnitude as  $|P|$ . The rounding error ( $\delta P$  for  $P$  and  $\delta P'$  for  $P'$ ) can be evaluated by the methods of Chap. 1 Sec. 2. Moreover the main source of rounding error comes from the evaluation of  $P'/P$ , so we replace  $P'$  and  $P$  by disks

$$E = \{P'; \delta P'\} \text{ and } F = \{P; \delta P\} \quad (4.224)$$

This means that the term  $\frac{P'}{P}$  in 4.137 is replaced by

$$Q_i^{(k)} = E.F^{-1} = \{P'; \delta P'\} \left\{ \frac{\bar{P}}{|P|^2 - (\delta P)^2}; \frac{\delta P}{|P|^2 - (\delta P)^2} \right\} \quad (4.225)$$

$$= \left\{ \frac{P'\bar{P}}{|P|^2 - (\delta P)^2}; \frac{|P'|\delta P + |P|\delta P' + \delta P\delta P'}{|P|^2 - (\delta P)^2} \right\} \quad (4.226)$$

where  $P$  etc are evaluated at  $z_i^{(k)}$ .

Let  $\delta_i^{(k)}$  denote the radius of  $Q_i^{(k)}$ , and let

$$\delta^{(k)} = \text{Max}_{1 \leq i \leq n} \delta_i^{(k)} \quad (4.227)$$

then Gargantini shows that if

$$|\delta P_i^{(k)}| < |P(z_i^{(k)})|, (i = 1, \dots, n) \quad (4.228)$$

$$\delta^{(k)} \leq \frac{r^{(k)}}{(\rho^{(k)})^2} \quad (4.229)$$

and

$$6(n-1)r^{(0)} < \rho^{(0)} \quad (4.230)$$

then convergence is cubic. But if 4.229 is not true and yet

$$\delta^{(k)} \leq \min\left\{1, \frac{1}{(\rho^{(k)})^2}\right\} \quad (4.231)$$

applies (the other conditions being the same), then convergence is only quadratic

Similarly Gargantini (1979) shows that Ostrowski's square-root disk method (similar to 4.185 but with  $z_j$  replaced by the disk  $Z_j$ ) with the conditions

$$\rho^{(0)} > 3(n+1)r^{(0)}; |\delta P_i^{(k)}| < |P(z_i^{(k)})| \text{ and } \delta^{(k)} < \frac{r^{(k)}}{(\rho^{(k)})^3} \quad (4.232)$$

converges with order 4, while if 4.232 is not true, but

$$\delta^{(k)} < \min\left\{1, \frac{1}{(\rho^{(k)})^3}\right\} \quad (4.233)$$

then the order is 3.

Petkovic and Stefanovic (1984) consider the disk-iteration 4.182, which generalizes the last two methods mentioned. They show that if

$$|\delta P_i^{(m)}| < |P(z_i^{(m)})| \quad (4.234)$$

$$\delta_i^{(m)} \leq \frac{r^{(m)}}{(\rho^{(m)})^{k+1}} \quad (4.235)$$

(m is the iteration index here) and

$$\rho^{(0)} > \beta(k, n)r^{(0)} \quad (4.236)$$

where

$$\beta(1, n) = 4n; \beta(k, n) = k(n-1) \quad (k > 1) \quad (4.237)$$

then convergence is of order  $k+2$ . But if the other conditions hold, and 4.235 is not true but still

$$\delta_i^{(m)} \leq \min\left(1, \frac{1}{(\rho^{(m)})^{k+1}}\right) \quad (4.238)$$

then convergence is only of order  $k+1$ .

Petkovic and Stefanovic (1986A) consider the case of the disk-WDK method 4.128, and with  $\delta P_i^{(k)}$  as above and

$$\delta = \max_{i=1}^n |\delta P_i^{(k)}|, \quad r = \max_{i=1}^n sd(Z_i^{(k)}) \quad (4.239)$$

(Note that they use rectangular intervals), they show that

$$\hat{r} \leq \alpha r^2 + \beta \delta \quad (4.240)$$

where  $\alpha$  and  $\beta$  are constants. Note that  $\hat{r}$  refers to the new value.

This means that as long as  $\delta = O(r^2)$  convergence remains quadratic, but if  $\delta = O(r)$  it is linear, while if rounding error exceeds the value of the semi-diagonal (i.e.  $r = o(\delta)$ ) then further convergence is not possible.

Petkovic (1989A) considers the Halley-like disk method 4.203 and shows that with the same meanings for  $r$ ,  $\hat{r}$  and  $\delta$  as in 4.240, then

$$\hat{r} \leq r^2(\gamma_1 \delta + \gamma_2 r^2 + \gamma_3 r \delta^2 + \gamma_4 \delta r^3) \quad (4.241)$$

where the  $\gamma_i$  are constants. Thus if  $\delta = O(r^2)$  the order 4 is preserved, while if  $\delta = O(r)$  convergence is cubic. Also he states that if  $r = o(\delta)$  and  $|\delta P| < |P|$  (still) convergence is quadratic.

## 4.8 Gauss-Seidel and SOR Variations

The Gauss-Seidel method in Linear Algebra consists in using the already obtained values of the next iterate  $\hat{z}_i$  in the correction term whenever possible. Many authors

treat a similar method in reference to simultaneous root-finding. For example, Niell (2001) gives the following modification of the WDK method:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1}^{i-1} (z_i^{(k)} - z_j^{(k+1)}) \prod_{j=i+1}^n (z_i^{(k)} - z_j^{(k)})}$$

$$(i = 1, \dots, n) \tag{4.242}$$

He shows that convergence is of order  $t$ , where  $t$  is the positive root of

$$f(t) \equiv (t - 1)^n - t = 0 \tag{4.243}$$

so that  $2 < t < 3$  (for  $f(2) = -1, f(3) \geq +1$ )

Some of the values of  $t$  for various  $n$  are shown below

$n$	2	3	5	10	15
$t$	2.61	2.32	2.17	2.08	2.05

It is seen that for large  $n$  the Gauss-Seidel variation (also known as serial or single-step) has order not much different from the normal (parallel or total-step) WDK method. So for large  $n$  it may not be worth sacrificing parallelism for the sake of a slightly higher order of convergence. But in cases where a great many low-degree polynomials need to be solved G.-S. versions should be considered. We will mention a few of the many works on this topic. For example Monsi and Wolfe (1988) describe the “symmetric single-step” procedure (PSS)

$$z_i^{(k,1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1}^{i-1} (z_i^{(k)} - z_j^{(k,1)}) \prod_{j=i+1}^n (z_i^{(k)} - z_j^{(k)})}$$

$$(i = 1, \dots, n) \tag{4.244}$$

$$z_i^{(k,2)} = z_i^{(k)}$$

$$- \frac{P(z_i^{(k)})}{\prod_{j=1}^{i-1} (z_i^{(k)} - z_j^{(k,1)}) \prod_{j=i+1}^n (z_i^{(k)} - z_j^{(k,2)})} \quad (i = n, n - 1, \dots, 1) \tag{4.245}$$

$$z_i^{(k+1)} = z_i^{(k,2)} \quad (i = 1, \dots, n) \tag{4.246}$$

They also describe a “repeated symmetric single-step” method (PRSS) whereby 4.244-4.246 are repeated  $r_k$  times with the same value of  $P(z_i^{(k)})$  in the numerator, and in addition they describe interval versions of the above. Numerical experiments show that the mixed interval-point PRSS method with  $r_k = 3$  is more efficient than the WDK, Gauss-Seidel-WDK (4.242) or PSS.

Kanno et al (1996) describe an SOR-GS method

$$z_i^{(k+1)} = z_i^{(k)} - \omega \frac{P(z_i^{(k)})}{\prod_{j=1}^{i-1} (z_i^{(k)} - z_j^{(k+1)}) \prod_{j=i+1}^n (z_i^{(k)} - z_j^{(k)})}$$

$$(i = 1, \dots, n) \tag{4.247}$$

They show that if

$$|\omega - 1| < 1 \tag{4.248}$$

and the zeros are simple, then 4.247 converges locally to these zeros. If  $\omega$  is real 4.248 becomes

$$0 < \omega < 2 \tag{4.249}$$

However if  $\omega$  is close to 2, convergence is slow near the zeros, whereas if  $\omega$  is close to 1 convergence may be fast. Numerical experiments show that the best value of  $\omega$  varies from case to case in the range  $[1, 1.4]$ , sometimes giving an improvement of about 30% compared to the pure G.-S. method ( $\omega = 1$ ). Similarly Petkovic and Kjurkchiev (1997) find that  $\omega = 1$  gives faster solutions than any value  $< 1$ .

Yamamoto (1996) shows that the SOR versions of many well-known methods converge if  $|\omega - 1| < 1$ .

Alefeld and Herzberger (1974) describe and analyse a Gauss-Seidel version of the Ehrlich-Aberth method 4.17 i.e.

$$z_i^{(k+1)} = z_i^{(k)} - \frac{1}{\frac{P'(z_i^{(k)})}{P(z_i^{(k)})} - \sum_{j=1}^{i-1} \frac{1}{z_i^{(k)} - z_j^{(k+1)}} - \sum_{j=i+1}^n \frac{1}{z_i^{(k)} - z_j^{(k)}}}$$

$$(i = 1, \dots, n) \tag{4.250}$$

If

$$h_i^{(k)} = z_i^{(k)} - \zeta_i \tag{4.251}$$

and  $\eta_i^{(k)} = \gamma h_i^{(k)}$ , where  $\gamma$  is a constant, they show that

$$\eta_i^{(k+1)} \leq \frac{1}{n-1} (\eta_i^{(k)})^2 \left[ \sum_{j=1}^{i-1} \eta_j^{(k+1)} + \sum_{j=i+1}^n \eta_j^{(k)} \right] \tag{4.252}$$

Since

$$\lim_{k \rightarrow \infty} |h_i^{(k)}| = 0 \tag{4.253}$$

(i.e. the method converges) we may assume that  $\eta_i^{(0)} \leq \eta < 1$ . Hence

$$\eta_i^{(k+1)} \leq \eta^{m_i^{(k+1)}} \tag{4.254}$$

where  $\mathbf{m}^{(k)} = (m_i^{(k)})$  can be calculated by

$$\mathbf{m}^{(k+1)} = \mathbf{A}\mathbf{m}^{(k)} \tag{4.255}$$

with

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 & \dots & \dots & \dots & \dots \\ 0 & 2 & 1 & 0 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 0 & 2 & 1 \\ 2 & 1 & 0 & \dots & \dots & 0 & 2 \end{bmatrix} \text{ and } \mathbf{m}^{(0)} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ \dots \\ \dots \\ 1 \\ 1 \end{bmatrix} \tag{4.256}$$

**Sketch of Proof.** 4.252 gives

$$\eta_i^{(1)} \leq \frac{1}{n-1} \eta^2 \left[ \sum_{j=1}^{i-1} \eta_j^{(1)} + \sum_{j=i+1}^n \eta \right] \tag{4.257}$$

e.g.  $\eta_1^{(1)} \leq \frac{1}{n-1} \eta^2 [\sum_{j=2}^n \eta] = \eta^3 = \eta^{2+1}$

$\eta_2^{(1)} \leq \frac{1}{n-1} \eta^2 [\eta_1^{(1)} + \sum_{j=3}^n \eta] \leq \eta^3 = \eta^{2+1}$ .

(Note that the dominant term in the square brackets is always the one with the **lowest** power, since  $\eta < 1$ ).

Similarly  $\eta_i^{(1)} \leq \eta^3$  ( $i = 3, \dots, n-1$ ).

However,  $\eta_n^{(1)} \leq \frac{1}{n-1} \eta^2 [\sum_{j=1}^{n-1} \eta_j^{(1)}] = \eta^5 = \eta^{2+1+2}$

Thus  $\mathbf{m}^{(1)} = [3, 3, \dots, 3, 5] = \mathbf{A}\mathbf{m}^{(0)}$

Next for  $k = 1$  we have

$$\eta_1^{(2)} \leq \frac{1}{n-1} (\eta_1^{(1)})^2 \left[ \sum_{j=2}^n \eta_j^{(1)} \right] \leq \eta^9 = \eta^{2 \times m_1^{(1)} + m_2^{(1)}}$$

$$\eta_2^{(2)} \leq \frac{1}{n-1} (\eta_2^{(1)})^2 [\eta_1^{(2)} + \sum_{j=3}^n \eta_j^{(1)}] \leq$$

$$\frac{1}{n-1} \eta^6 [\eta^9 + (n-3)\eta^3 + \eta^5] \leq \eta^9$$

and similarly for  $j=3, \dots, n-2$ .

$$\eta_{n-1}^{(2)} \leq \frac{1}{n-1} (\eta_{n-1}^{(1)})^2 \left[ \sum_{j=1}^{n-2} \eta_j^{(2)} + \eta_n^{(1)} \right]$$

$$\begin{aligned} &\leq \frac{1}{n-1} \eta^6 [(n-2)\eta^9 + \eta^5] \leq \eta^{11} \\ &= \eta^{2 \times m_{n-1}^{(1)} + m_n^{(1)}} \\ \eta_n^{(2)} &\leq \frac{1}{n-1} (\eta_n^{(1)})^2 \left[ \sum_{j=1}^{n-1} \eta_j^{(2)} \right] = \frac{1}{n-1} \eta^{10} [(n-2)\eta^9 + \eta^{11}] \\ &\leq \eta^{19} = \eta^{2 \times m_1^{(1)} + m_2^{(1)} + 2 \times m_n^{(1)}} \end{aligned}$$

i.e.

$$\mathbf{m}^{(2)} = \mathbf{A} \mathbf{m}^{(1)}$$

Thus 4.255 is verified for  $k = 0$  and  $1$ , and the general case follows by induction.  $\mathbf{A}$  is non-negative, irreducible and primitive, so that it has a positive eigenvalue equal to its spectral radius  $\rho(\mathbf{A})$ . The authors show that the order of convergence is  $\geq \rho(\mathbf{A})$ . Now let us consider  $|\mathbf{A} - \lambda \mathbf{I}|_n$  for a few small values of  $n$ , e.g.

$$\begin{aligned} |\mathbf{A} - \lambda \mathbf{I}|_3 &= \begin{vmatrix} 2-\lambda & 1 & 0 \\ 0 & 2-\lambda & 1 \\ 2 & 1 & 2-\lambda \end{vmatrix} = (2-\lambda)[(2-\lambda)^2 - 1] + 2 = \\ & -[(\lambda-2)^3 - (\lambda-2) - 2] \\ |\mathbf{A} - \lambda \mathbf{I}|_4 &= \begin{vmatrix} 2-\lambda & 1 & 0 & 0 \\ 0 & 2-\lambda & 1 & 0 \\ 0 & 0 & 2-\lambda & 1 \\ 2 & 1 & 0 & 2-\lambda \end{vmatrix} = \\ (2-\lambda) &\begin{vmatrix} 2-\lambda & 1 & 0 \\ 0 & 2-\lambda & 1 \\ 1 & 0 & 2-\lambda \end{vmatrix} - 2 \begin{vmatrix} 1 & 0 & 0 \\ 2-\lambda & 1 & 0 \\ 0 & 2-\lambda & 1 \end{vmatrix} = \\ &(2-\lambda)[(2-\lambda)^3 + 1] - 2 = (\lambda-2)^4 - (\lambda-2) - 2 \end{aligned}$$

and it may be proved by induction that, for general  $n$ ,

$$p_n(\lambda) = (-1)^n |\mathbf{A} - \lambda \mathbf{I}|_n = (\lambda-2)^n - (\lambda-2) - 2 \tag{4.258}$$

Setting  $\sigma = \lambda - 2$  we get

$$\hat{p}_n(\sigma) = \sigma^n - \sigma - 2 \tag{4.259}$$

Now  $\hat{p}_n(1) = -2$  and  $\hat{p}_n(2) \geq 0$  for  $n \geq 2$ , so there is a positive root  $\sigma_n$  with  $1 < \sigma_n \leq 2$ , and by Descartes' rule of signs it must be unique. Thus the order of convergence

$$\geq \rho(\mathbf{A}) = 2 + \sigma_n \in [3, 4] \tag{4.260}$$

The authors do not give numerical values of  $\sigma_n$ , but we calculate  $\sigma_{20} = 1.06$ .

Petkovic and Milanovic (1983) give a similar analysis of the Gauss-Seidel version of the “Improved Ehrlich-Aberth Method” 4.26, i.e.

$$z_i^{(k+1)} = z_i^{(k)} - \frac{1}{\frac{P'(z_i^{(k)})}{P(z_i^{(k)})} - \sum_{j=1}^{i-1} \frac{1}{z_i^{(k)} - z_j^{(k+1)}} - \sum_{j=i+1}^n \frac{1}{z_i^{(k)} - z_j^{(k)} + \frac{P(z_j^{(k)})}{P'(z_j^{(k)})}}}$$
(4.261)

They show that the order is  $2(1 + \tau_n) > 4$  where  $\tau_n \in [1, 2]$  is the unique positive root of

$$\tau^n - \tau - 1 = 0$$
(4.262)

Milovanovic and Petkovic (1983) give some typical values of the order as follows:

<i>n</i>	2	3	5	10
<i>order</i>	5.24	4.65	4.34	4.15

Again, it is seen that there is not much advantage in the serial method for large n.

Hansen et al (1977) give a serial version of the modified Laguerre method 4.171 of order  $> 4$ .

Petkovic and Stefanovic (1986B) give a Gauss-Seidel version of their “Square root iteration with Halley correction”, given by 4.185 and 4.188, having convergence order  $\in [6, 7]$  and hence efficiency  $\approx \log(\sqrt[5]{6.5}) = .1626$

Petkovic, Milovanovic and Stefanovic (1986) give a version of the above for multiple roots, again of efficiency .1626

Petkovic, Stefanovic and Marjanovic (1992) and (1993) give several G.-S. methods, the most efficient of which is the G.-S. version of 4.210,, with efficiency  $\log(\sqrt[4.5]{6.5}) = 0.1806$

Petkovic and Stefanovic (1990) show that a forward-backward variation on the G.-S. version of 4.203 has convergence order about 50% higher than the plain forward-forward version (for  $n \geq 10$ ), i.e. the order is at least 6, and so the efficiency is  $\log(\sqrt[5.5]{6}) = .1415$ .

Ellis and Watson (1984) give a method based on divided differences thus: with

$$W_i = \frac{P(z_i)}{\prod_{j \neq i} (z_i - z_j)}$$
(4.263)



and

$$Q_i(s) = \sum_{j \neq i} \frac{W_j}{z_j - s} \quad (4.264)$$

$$r_i^{(k+1)} = r_i^{(k)} - \frac{[(Q_i(r_i^{(k)}) - 1)^2(r_i^{(k)} - z_i^{(0)}) - W_i(Q_i(r_i^{(k)}) - 1)]}{(Q_i(r_i^{(k)}) - 1)^2 + W_i Q_i'(r_i^{(k)})} \quad (4.265)$$

$(k = 0, \dots, \text{to convergence})$

Initially  $r_i^{(0)} = z_i^{(0)}$  ( $i = 0, \dots, n$ ), and after convergence  $z_i^{(K)} = r_i^{(K)}$  ( $i = 1, \dots, n$ ) where  $K$  is the latest value of  $k$  (i.e. value at which convergence occurs). Experiments show that running in parallel, this method is faster than certain standard methods, and always converges.

## 4.9 Real Factorization Methods

Several authors give methods which split a polynomial with real coefficients into two or more **real** factors, such as quadratics (with one real factor as well if the degree is odd). This means that we may work entirely with real numbers, which is often more efficient than finding individual roots by for example the WDK method (which needs complex arithmetic if the roots are complex).

For example, Freeman and Brankin (1990) describe a “divide and conquer” method whereby

$$P_n(x) = Q_N(x)R_N(x) \quad (4.266)$$

The easiest case is where  $n$  is divisible by 4, which we will describe (they also consider  $n$  odd, or even but not a multiple of 4).

Here  $N = \frac{n}{2}$  and

$$Q_N(x) = x^N + b_{N-1}x^{N-1} + \dots + b_1x + b_0 \quad (4.267)$$

$$R_N(x) = x^N + c_{N-1}x^{N-1} + \dots + c_1x + c_0 \quad (4.268)$$

Equating coefficients of  $x^k$  ( $k = n - 1, n - 2, \dots, 1, 0$ ) in 4.266 leads to

$$b_{N-1} + c_{N-1} - a_{n-1} = 0$$

$$b_{N-2} + b_{N-1}c_{N-1} + c_{N-2} - a_{n-2} = 0$$

$$b_{N-3} + b_{N-2}c_{N-1} + b_{N-1}c_{N-2} + c_{N-3} - a_{n-3} = 0$$

.....

.....

$$b_0 + b_1c_{N-1} + b_2c_{N-2} + \dots + b_{N-1}c_1 + c_0 - a_N = 0$$

$$b_0c_{N-1} + b_1c_{N-2} + \dots + b_{N-1}c_0 - a_{N-1} = 0$$

$$b_0c_{N-2} + b_1c_{N-3} + \dots + b_{N-2}c_0 - a_{N-2} = 0$$

.....

.....

$$b_0c_0 - a_0 = 0 \tag{4.269}$$

The above may be written

$$\mathbf{f}(\mathbf{b}, \mathbf{c}) = \mathbf{0} \tag{4.270}$$

where  $\mathbf{f}^T = (f_1, f_2, \dots, f_n)$ ,  $\mathbf{b}^T = (b_{N-1}, \dots, b_0)$ ,  $\mathbf{c}^T = (c_{N-1}, \dots, c_0)$ , and

$$f_k(\mathbf{b}, \mathbf{c}) = b_{N-k} + \sum_{j=1}^{k-1} b_{N-k+j}c_{N-j} + c_{N-k} - a_{n-k}$$

$$(k = 1, 2, \dots, N) \tag{4.271}$$

$$= \sum_{j=k}^n b_{n-j}c_{j-k} - a_{n-k} \quad (k = N + 1, \dots, n) \tag{4.272}$$

The Newton iteration for the solution of 4.270 is given by

$$\begin{pmatrix} \mathbf{b}^{(i+1)} \\ \mathbf{c}^{(i+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{b}^{(i)} \\ \mathbf{c}^{(i)} \end{pmatrix} + \begin{pmatrix} \delta \mathbf{b}^{(i)} \\ \delta \mathbf{c}^{(i)} \end{pmatrix} \tag{4.273}$$

where

$$\mathbf{J}^{(i)} \begin{pmatrix} \delta \mathbf{b}^{(i)} \\ \delta \mathbf{c}^{(i)} \end{pmatrix} = -\mathbf{f}^{(i)} \tag{4.274}$$

and  $\mathbf{J}^{(i)}$  is the Jacobian of  $\mathbf{f}^{(i)}$ . (Of course the superscript i refers to the i-th iteration)

The authors show that

$$\mathbf{J} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \tag{4.275}$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ c_{N-1} & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ c_1 & \dots & \dots & c_{N-1} & 1 \end{pmatrix} \quad (4.276)$$

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ b_{N-1} & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ b_1 & \dots & \dots & b_{N-1} & 1 \end{pmatrix} \quad (4.277)$$

$$\mathbf{C} = \begin{pmatrix} c_0 & c_1 & \dots & \dots & c_{N-1} \\ 0 & c_0 & \dots & \dots & c_{N-2} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & c_0 \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} b_0 & b_1 & \dots & \dots & b_{N-1} \\ 0 & b_0 & \dots & \dots & b_{N-2} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & b_0 \end{pmatrix} \quad (4.278)$$

The blocks  $\mathbf{A}$  etc are Toeplitz. If we partition

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix} \quad (4.279)$$

where  $\mathbf{f}_i$  are  $N$ -vectors, and omit superscript  $i$ , we may write 4.274 as

$$\mathbf{A}\delta\mathbf{b} + \mathbf{B}\delta\mathbf{c} = -\mathbf{f}_1 \quad (4.280)$$

$$\mathbf{C}\delta\mathbf{b} + \mathbf{D}\delta\mathbf{c} = -\mathbf{f}_2 \quad (4.281)$$

Now the inverse of a triangular Toeplitz matrix is also a triangular Toeplitz matrix. Premultiply 4.280 by  $\mathbf{B}^{-1}$  and 4.281 by  $\mathbf{D}^{-1}$  to give

$$\mathbf{B}^{-1}\mathbf{A}\delta\mathbf{b} + \delta\mathbf{c} = -\mathbf{B}^{-1}\mathbf{f}_1; \quad \mathbf{D}^{-1}\mathbf{C}\delta\mathbf{b} + \delta\mathbf{c} = -\mathbf{D}^{-1}\mathbf{f}_2 \quad (4.282)$$

Eliminating  $\delta\mathbf{c}$  gives

$$\mathbf{T}\delta\mathbf{b} = \hat{\mathbf{f}} \quad (4.283)$$

where

$$\mathbf{T} = \mathbf{B}^{-1}\mathbf{A} - \mathbf{D}^{-1}\mathbf{C} \quad (4.284)$$

and

$$\hat{\mathbf{f}} = \mathbf{D}^{-1}\mathbf{f}_2 - \mathbf{B}^{-1}\mathbf{f}_1 \quad (4.285)$$

4.283 may be solved in  $3N^2$  operations, and further calculating  $\delta\mathbf{c}$  requires a total of  $\frac{15}{2}N^2$  operations.  $\mathbf{J}$  is only singular if  $Q_N(x)$  and  $R_N(x)$  have at least one common zero. Convergence (provided good starting values are available) is quadratic.

The authors suggest the following starting values: let  $r$  be a bound on the moduli of the zeros of  $P_n(x)$ , then take

$$b_i = 0 \quad (i = 1, 2, \dots, N - 1), \quad b_0 = r \quad (4.286)$$

while the  $c_i$  starting values can then be obtained from 4.269

Experiments show that this method requires about  $\frac{3}{8}$  the work of a WDK implementation (but it is less robust).

The authors do not state this, but presumably the method could be applied recursively until all the factors are quadratic.

In an earlier article Freeman (1979) derives quadratic factors directly. He assumes that  $n$  is even ( $=2N$ ) and that

$$P_n(x) = (x^2 + \beta_1x + \gamma_1)(x^2 + \beta_2x + \gamma_2)\dots(x^2 + \beta_Nx + \gamma_N) \quad (4.287)$$

Equating coefficients of  $x^{n-1}, x^{n-2}, \dots, x^0$  in the usual representation of  $P_n(x)$  and in the expansion of 4.287 gives a series of non-linear equations such as:

$$f_1(\beta, \gamma) = \bar{F}_1^{(n)}(\beta, \gamma) - c_{n-1} = \sum_{i=1}^N \beta_i - c_{n-1} \quad (4.288)$$

$$f_2(\beta, \gamma) = \bar{F}_2^{(n)}(\beta, \gamma) - c_{n-2} = \sum_{i < j} \beta_i \beta_j + \sum_{k=1}^N \gamma_k - c_{n-2} \quad (4.289)$$

etc, etc. Let

$$\omega = [\beta_1, \gamma_1, \beta_2, \gamma_2, \dots, \beta_N, \gamma_N]^T \quad (4.290)$$

and

$$\mathbf{f}(\omega) = [f_1(\beta, \gamma), f_2(\beta, \gamma), \dots, f_n(\beta, \gamma)]^T = 0 \quad (4.291)$$

The authors use a damped Newton method

$$\omega^{(i+1)} = \omega^{(i)} + \alpha^{(i)} \mathbf{p}^{(i)} \quad (i = 1, 2, \dots) \quad (4.292)$$

where  $\mathbf{p}^{(i)}$  is the solution of

$$\mathbf{J}^{(i)} \mathbf{p}^{(i)} = -\mathbf{f}^{(i)} \quad (4.293)$$

and  $\mathbf{J}^{(i)}$  is the Jacobian evaluated at  $\omega^{(i)}$  i.e.  $i$  refers to the iteration number. Note that

$$J_{k,l} = \frac{\partial f_k}{\partial \omega_l} \quad (4.294)$$

To evaluate the  $f_i$  (now  $i$  refers to the index number in the vector  $\mathbf{f}$ ) we may use certain recurrence relations (proved in a report by Freeman). We have  $f_i = \bar{f}_i^{(n)} - c_{n-i}$

Where

$$\bar{f}_i^{(j)} = \bar{f}_i^{(j-2)} + \beta_{j/2} \bar{f}_{i-1}^{(j-2)} + \gamma_{j/2} \bar{f}_{i-2}^{(j-2)} \quad (i = 1, 2, \dots, j/2) \quad (4.295)$$

$$\bar{f}_{j-i+1}^{(j)} = \bar{f}_{j-i+1}^{(j-2)} + \beta_{j/2} \bar{f}_{j-i}^{(j-2)} + \gamma_{j/2} \bar{f}_{j-i-1}^{(j-2)} \quad (i = 1, 2, \dots, j/2) \quad (4.296)$$

with starting values

$$\bar{f}_0^{(j-2)} = 1; \bar{f}_{-1}^{(j-2)} = \bar{f}_{j-1}^{(j-2)} = \bar{f}_j^{(j-2)} = 0 \quad (4.297)$$

and, as may be seen by inspection of 4.287 with  $N = 2$ ,

$$\bar{\mathbf{f}}^{(4)} = [\beta_1 + \beta_2, \beta_1\beta_2 + \gamma_1 + \gamma_2, \beta_1\gamma_2 + \beta_2\gamma_1, \gamma_1\gamma_2] \quad (4.298)$$

Thus for example if  $n = j = 6$  ( $N = 3$ ) we get

$$\begin{aligned} \bar{f}_1^{(6)} &= \bar{f}_1^{(4)} + \beta_3 \bar{f}_0^{(4)} + \gamma_3 \bar{f}_{-1}^{(4)} \\ &= (\beta_1 + \beta_2) + \beta_3(1) + \gamma_3(0) = \beta_1 + \beta_2 + \beta_3 \end{aligned} \quad (4.299)$$

and

$$\begin{aligned} \bar{f}_2^{(6)} &= \bar{f}_2^{(4)} + \beta_3 \bar{f}_1^{(4)} + \gamma_3 \bar{f}_0^{(4)} \\ &= (\beta_1\beta_2 + \gamma_1 + \gamma_2) + \beta_3(\beta_1 + \beta_2) + \gamma_3(1) \\ &= \sum_{i,j=1, i < j}^3 \beta_i \beta_j + \sum_{k=1}^3 \gamma_k \end{aligned} \quad (4.300)$$

in agreement with 4.288 and 4.289 for  $N = 3$ . The reader may verify that the expressions for  $\bar{f}_3^{(6)} \dots \bar{f}_6^{(6)}$  obtained by the above method agree with what we get by inspection of 4.287 with  $N = 3$ . The author gives explicit expressions for the elements of the Jacobian, but we may obtain them more easily by the following recurrence relations: (but first note that

$$J_{k,l+1} = J_{k-1,l} \quad (k = 2, 3, \dots, n, l \text{ odd}) \quad (4.301)$$

$$J_{1,l+1} = 0; J_{1,l} = 1 \text{ (} l \text{ odd)} \tag{4.302}$$

so that we only need to compute odd-numbered columns).

$$J_{2,l} = \bar{F}_1^{(n)} - \beta_{(l+1)/2} J_{1,l} \tag{4.303}$$

$$J_{k,l} = \bar{F}_{k-1}^{(n)} - \beta_{(l+1)/2} J_{k-1,l} - \gamma_{(l+1)/2} J_{k-2,l} \tag{4.304}$$

(but as we see below these are not needed in practise). 4.293 is best solved by factorizing  $\mathbf{J}$  into  $\bar{\mathbf{L}}$  and  $\bar{\mathbf{U}}$  which are not quite lower and upper triangular matrices. For example, for  $n = 4$  from 4.294 and 4.298 we have

$$\mathbf{J} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ \beta_2 & 1 & \beta_1 & 1 \\ \gamma_2 & \beta_2 & \gamma_1 & \beta_1 \\ 0 & \gamma_2 & 0 & \gamma_1 \end{bmatrix} = \bar{\mathbf{L}}\bar{\mathbf{U}} \tag{4.305}$$

where

$$\bar{\mathbf{L}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \gamma_2 - \beta_2^2 & \beta_2 & 1 & 0 \\ -\beta_2\gamma_2 & \gamma_2 & 0 & 1 \end{bmatrix} \tag{4.306}$$

$$\bar{\mathbf{U}} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ \beta_2 & 1 & \beta_1 & 1 \\ 0 & 0 & (\gamma_1 - \gamma_2 - \beta_2(\beta_1 - \beta_2)) & \beta_1 - \beta_2 \\ 0 & 0 & \gamma_2(\beta_1 - \beta_2) & \gamma_1 - \gamma_2 \end{bmatrix} \tag{4.307}$$

For general  $n$  Freeman gives a lengthy list of expressions for  $\bar{L}_{ij}$  and  $\bar{U}_{ij}$  on p326 of the cited paper. It means that the  $\bar{\mathbf{L}}\bar{\mathbf{U}}$  factors can be found in  $\frac{7}{4}n^2 + O(n)$  multiplications without directly evaluating  $\mathbf{J}$ . Freeman shows that  $\mathbf{J}$  is singular only if two quadratic factors are identical or have one real root in common.

$\alpha^{(i)}$ , the line search parameter, is set to  $2^{-k}$  ( $k = 0, 1, 2, \dots$ ) where  $k$  is the smallest integer such that

$$a) F(\omega^{(i)} + 2^{-k} \mathbf{p}^{(i)}) < \epsilon_1 F(\omega^{(i)}) \text{ or} \tag{4.308}$$

$$b) F(\omega^{(i)} + 2^{-k} \mathbf{p}^{(i)}) < F(\omega^{(i)} + 2^{-(k+1)} \mathbf{p}^{(i)}) \text{ and}$$

$$F(\omega^{(i)} + 2^{-k} \mathbf{p}^{(i)}) < \epsilon_2 F(\omega^{(i)}) \tag{4.309}$$

where  $0 < \epsilon_1 < \epsilon_2 \leq 1$  and

$$F(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \mathbf{D} \mathbf{f}(\mathbf{x}) \tag{4.310}$$

and  $\mathbf{D}$  is a diagonal matrix with

$$D_{kk} = \frac{1}{c_{n-k}^2} \tag{4.311}$$

Experiments show that  $\alpha^{(i)}$  is usually chosen as 1 and the method is quite robust.

### 4.10 Comparison of Efficiencies

A few authors (although in our opinion not enough) compare the efficiencies of various methods, for example Milovanovic and Petkovic (1986). They point out that various measures of efficiency can be found in the literature, such as

$$E(SIP, n) = \frac{r(n)}{\Theta(n)} \quad (4.312)$$

or

$$*E(SIP, n) = r(n)^{\frac{1}{\Theta(n)}} \quad (4.313)$$

where  $r(n)$  is the order of a simultaneous iterative process (SIP) applied to a polynomial of degree  $n$ , and

$\Theta(n)$  is a normalized cost of evaluating the new iterate (including computing the polynomial and some of its derivatives), given by

$$\Theta(n) = \frac{T(n)}{G(n)} = 1 + \frac{w_A A(n) + w_S S(n) + w_M M(n) + w_D D(n)}{G(n)} \quad (4.314)$$

where  $T(n)$  is the total cost of evaluation (for all zeros) per iteration and

$$G(n) = w_A n^2 + w_M n^2 \quad (4.315)$$

is the cost of evaluating the polynomial (of degree  $n$ ) itself. The weights  $w_A$  etc are the times required for the various operations, normalized with respect to addition (i.e.  $w_A = 1$ ), and  $A(n)$  etc are the number of adds etc needed, apart from the evaluation of the polynomial. The cited authors report that 4.312 agrees better with experiment, although theoretically 4.313 or

$$\frac{\log r(n)}{\Theta(n)} \quad (4.316)$$

should be more accurate (see Chapter 1 section 9). They compare 10 different methods, several of which have been mentioned in previous sections of this chapter. Letting

$$W_i = \frac{P(z_i)}{\prod_{j=1, \neq i} (z_i - z_j)} \quad (4.317)$$

(Weierstrass or WDK correction) and

$$N_i = \frac{P(z_i)}{P'(z_i)} \quad (4.318)$$

(Newton's correction), the methods, numbered I through X, are as listed below, with  $z_i^{(k)}$  denoted by  $z_i$  and  $z_i^{(k+1)}$  as  $\hat{z}_i$  ( $i = 1, \dots, n$ )

$$(I) \hat{z}_i = z_i - \frac{P(z_i)}{\prod_{j=1, \neq i}^n (z_i - z_j)} \quad (4.319)$$

(the WDK method, 4.1)

$$(II) \hat{z}_i = z_i - \frac{P(z_i)}{\prod_{j=1}^{i-1} (z_i - \hat{z}_j) \prod_{j=i+1}^n (z_i - z_j)} \quad (4.320)$$

(the Gauss-Seidel WDK-method, 4.242 )

$$(III) \hat{z}_i = z_i - \frac{P(z_i)}{\prod_{j=1, \neq i}^n (z_i - z_j + W_j)} \quad (4.321)$$

(the "Improved Durand-Kerner" method-4.25)

$$(IV) \hat{z}_i = z_i - \frac{P(z_i)}{\prod_{j=1}^{i-1} (z_i - \hat{z}_j) \prod_{j=i+1}^n (z_i - z_j + W_j)} \quad (4.322)$$

(Gauss-Seidel version of 4.321—see Petkovic and Milanovic (1983)).

$$(V) \hat{z}_i = z_i - \frac{W_i}{1 + \sum_{j=1, \neq i}^n \frac{W_j}{z_i - z_j}} \quad (4.323)$$

(Borsch-Supan method-4.22)

$$(VI) \hat{z}_i = z_i - \frac{W_i}{1 + \sum_{j=1, \neq i}^n \frac{W_j}{z_i - W_i - z_j}} \quad (4.324)$$

(Nourein's method-4.23)

$$(VII) \hat{z}_i = z_i - \frac{1}{\frac{P'(z_i)}{P(z_i)} - \sum_{j=1, \neq i}^n \frac{1}{z_i - z_j}} \quad (4.325)$$

(Ehrlich-Aberth method-4.17)

$$(VIII) \hat{z}_i = z_i - \frac{1}{\frac{P'(z_i)}{P(z_i)} - \sum_{j=1}^{i-1} \frac{1}{z_i - \hat{z}_j} - \sum_{j=i+1}^n \frac{1}{z_i - z_j}} \quad (4.326)$$

(Gauss-Seidel version of 4.325 i.e. 4.250)

$$(IX) \hat{z}_i = z_i - \frac{1}{\frac{P'(z_i)}{P(z_i)} - \sum_{j=1, \neq i}^n \frac{1}{z_i - z_j + \frac{P(z_j)}{P'(z_j)}}} \quad (4.327)$$



(“Improved Ehrlich-Aberth method” –4.26)

$$(X) \hat{z}_i = z_i - \frac{1}{\frac{P'(z_i)}{P(z_i)} - \sum_{j=1}^{i-1} \frac{1}{z_i - z_j} - \sum_{j=i+1}^n \frac{1}{z_i - z_j + \frac{P(z_j)}{P'(z_j)}}} \tag{4.328}$$

(G.-S. version of 4.327 i.e. 4.261)

The authors cited give the number of operations (excluding evaluations of the polynomial), and order of convergence of the various methods, as shown in the following table:

METHOD	I	II	III	IV	V	VI	VII	VIII	IX	X
<i>Oper</i> s	$2n^2$	$2n^2$	$5n^2$	$\frac{9}{2}n^2$	$5n^2$	$6n^2$	$5n^2$	$5n^2$	$6n^2$	$\frac{11}{2}n^2$
$r(3)$	2	2.33	3	3.15	3	4	3	3.52	4	4.65
$r_\infty$	2	2	3	3	3	4	3	3	4	4

(In fact the authors distinguish between different types of operation, and give also terms of  $O(n)$ . But we only count terms of  $O(n^2)$ , all together).  $r_\infty$  is the limit of  $r(n)$  as  $n \rightarrow \infty$ . Using the above values and 4.312 and 4.314, and considering several computers, the authors conclude that method II is the most efficient and VII the least. Their theoretical analysis is confirmed by experiments where actual CPU time is measured (except that now V is worst and VII is second worst).

In a slightly earlier paper Petkovic and Milanovic (1985) find X best for large  $n$ , although II is still best for smaller  $n$ .

Petkovic (1990) considers three of the above methods (I, VII, and IX) and Newton’s method

$$\hat{z}_i = z_i - \frac{P(z_i)}{P'(z_i)} \tag{4.329}$$

(He refers to these as P1, P3, P4 and P2 respectively-P for point). Also he considers some disk-methods namely the WDK disk formula 4.128, Ehrlich-Aberth -disk 4.137 and a Halley-like disk method

$$Z_i^{(k+1)} = z_i^{(k)} + \frac{1}{2[\frac{P''}{P'} - 2\frac{P'}{P}] + \frac{P}{2P'}[(\sum_{j=1, \neq i}^n \frac{1}{z_i - Z_j})^2 + \sum_{j=1, \neq i}^n \frac{1}{(z_i - Z_j)^2}]} \tag{4.330}$$

The above 3 disk methods are referred to as I1, I2, and I3 (I for interval). Actually he gives formulas for multiple roots, but only considers the simple-root case, so we do the same. He considers several combined methods, whereby a point method is used for  $M$  iterations and an interval method for one final iteration (to give a bound on the errors). Using an analysis similar to that in his previously-mentioned

paper, Petkovic finds that the most efficient combined method involves P4 and I2, i.e. the improved Ehrlich-Aberth method (IX) for the M point-iterations and the disk Ehrlich-Aberth method at the end. This contrasts with the previous result where Ehrlich-Aberth was worst.

## 4.11 Implementation on Parallel Computers

The methods of this chapter have many advantages (such as good convergence properties) even on a serial computer, but they are especially suited to implementation on parallel computers. Accordingly, several authors have considered this situation. Freeman (1989) considers the general simultaneous iteration

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\phi_i(z_1^{(k)}, z_2^{(k)}, \dots, z_n^{(k)})} \quad (i = 1, \dots, n) \quad (4.331)$$

and gives a parallel algorithm as follows, for p processors in which the l'th processor handles  $j_l$  approximations, and  $i_l = \sum_{m=1}^{l-1} j_m$  ( $l = 1, \dots, p$ ) ( $i_1 = 0$ ):

Step 1 (i)  $k = 1$

(ii) Define initial approximations  $z_i^{(1)}$

Step 2. In parallel, for  $l = 1, 2, \dots, p$  and for  $i = i_l + 1, i_l + 2, \dots, i_l + j_l$

(i) Calculate  $p_i^{(k)} = P(z_i^{(k)})$

(ii) Calculate  $q_i^{(k)} = \phi_i(z_1^{(k)}, z_2^{(k)}, \dots, z_n^{(k)})$

(iii) Set  $z_i^{(k+1)} = z_i^{(k)} - p_i^{(k)} / q_i^{(k)}$

Step 3. for  $i = 1, 2, \dots, n$  communicate  $z_i^{(k+1)}$  to all the processors.

Step 4. (i) Check for convergence

(ii) set  $k = k+1$

(iii) Go to step 2.

Freeman considers the particular cases where 4.331 is replaced by (I) the WDK method, (II) The Ehrlich-Aberth method 4.17 (henceforward referred to as the EA method). (III) A fourth-order formula of Farmer and Loizou (1975).

For step 4(i) he suggests the rounding-error based method of Adams (1967). This step can be performed simultaneously with Step 2(i) for the **next**  $z_i^{(k+1)}$ , and also with step 3. The author describes some experiments on an 8-processor linear chain. The results show that method (III) is often unreliable, but methods I and II show a speed-up of about 5.5 for some of the higher-degree polynomials tested (8 would be the maximum possible). Note that "speed-up" is defined as (Time with one processor)/(Time with p processors) and is  $\leq p$ . Less speed-up is

obtained with lower-degree polynomials.

Freeman and Bane (1991) consider **asynchronous** algorithms, in which each processor continues to update its approximations even although the latest values of the other  $z_i^{(k)}$  have not yet been received from the other processors (in the synchronous version it **would** wait). Thus the WDK method becomes:

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P(z_i^{(k)})}{\prod_{j=1, \neq i}^n (z_i^{(k)} - z_j^{(k_j)})} \quad (i = 1, \dots, n) \tag{4.332}$$

where  $k_j = k - \rho(j, k, h)$  and  $\rho(j, k, h)$  indicates that processor  $P_h$  knows only the value of  $z_j^{(k-\rho(j, k, h))}$ , i.e the value computed at step  $k-\rho(j, k, h)$ . Thus  $z_j$  may have been computed several steps prior to the  $k$ 'th. While saving time on communication, this strategy may lead to more iterations before convergence, and we need to balance these opposing forces.

We let

$$\rho = \text{Max}_{j,k,h} \rho(j, k, h) \tag{4.333}$$

be a measure of the asynchronism. The EA method is similarly modified.

Let

$$\epsilon_i^{(k)} = z_i^{(k)} - \zeta_i \quad (i = 1, \dots, n) \tag{4.334}$$

be small and the  $\zeta_i$  simple zeros. Then the authors prove that

$$\epsilon_i^{(k+1)} = -\epsilon_i^{(k)} \left[ \sum_{j \neq i} \frac{\epsilon_j^{(k_j)}}{(\zeta_i - \zeta_j)} + O(\epsilon^2) \right] \quad (i = 1, \dots, n); \quad k = 1, 2, \dots \tag{4.335}$$

where

$$\epsilon = \text{max}(|\epsilon_i^{(k)}|; |\epsilon_j^{(k_j)}|, j \neq i) \tag{4.336}$$

For the EA method there is a similar result with  $(\epsilon_i^{(k)})^2$  on the right. Note that the order of the WDK method approaches 2 as  $\rho$  approaches 0, but otherwise it is super-linear. The EA method similarly is superquadratic in general and cubic when  $\rho = 0$ .

The authors experiment with the case where approximations are exchanged after every  $m$  iterations. Not surprisingly, the number of iterations required for convergence increases with  $m$ . For small  $p$ , the savings in communication time (which may be an appreciable part of the total) are outweighed by the cost of extra iterations. But for large  $p$ , when communication costs are more significant, the choice of  $m = 2$  or  $3$  in the WDK case leads to a 10-20% net time reduction compared to the synchronous case ( $m = 1$ ). For EA and  $m = 2$  through  $5$  the speed-up is significant

even for small  $p$ , and is again about 20% for large  $p$ . Also EA is more robust.

Petkovic (1996) generalizes 4.335 to

$$\epsilon_i^{(k+1)} = \alpha_i (\epsilon_i^{(k)})^q \sum_{j=1, \neq i}^n \beta_{ij} \epsilon_i^{(k-\rho(j,k,h))} \quad (4.337)$$

(Note that WDK has  $q=1$  and EA has  $q=2$ ). Also if  $\rho=0$ , then by 4.337 the order is  $q+1$ . Then Petkovic shows that the order of the asynchronous method leading to 4.337 is the (only) positive root of

$$\eta_A^{\rho+1} - q\eta_A^\rho - 1 = 0 \quad (4.338)$$

(N.B. if  $\rho=0$  this becomes  $\eta_A = \eta_S = q+1$ , confirming the remark above). He gives a table of orders for several values of  $\rho$  and  $q$ , thus:

	$\rho$				
$q$	0	1	2	3	4
1	2	1.62	1.47	1.38	1.32
2	3	2.41	2.21	2.11	2.06
3	4	3.30	3.10	3.04	3.01

Now let us define  $N_S, T_S$  and  $N_A, T_A$  as the number of iteration steps and time per iteration of the synchronous and asynchronous methods respectively (often  $T_A < T_S$  as there is less communication). Then the asynchronous implementation will be faster overall if

$$\lambda_q = \frac{N_A}{N_S} \approx \frac{\log(q+1)}{\log(\eta_A(q))} < \frac{T_S}{T_A} \quad (4.339)$$

Petkovic shows that for all  $\rho$  considered,  $\lambda_q$  gets smaller (or stays the same) as  $q$  increases, so that 4.339 is more easily satisfied. This gives us a reason, when choosing between methods of the same efficiency (in synchronous case), to choose the one of higher order.

Cosnard and Fraignaud (1990) compare 3 different parallel network topologies (ring, 2-D torus, and hypercube). They conclude that the hypercube is by far the fastest. In experiments they obtain almost perfect speed-up.

Maeder and Wynton (1987) discuss the parallelization of several methods which are not normally considered “simultaneous”, such as the Sturm sequence method (see Chapter 2). They point out that several stages in this method are suitable for parallel computation. Firstly, the process of dividing the polynomials to form the Sturm functions can be performed in parallel. As soon as the first few functions are formed, values at a chosen set of points can be found, again in parallel. Moreover, once all the functions have been obtained and the values at the chosen points

calculated, the counting of the sign changes among the function values at the chosen points can proceed in parallel. Thus we can discover which intervals contain one or more roots, subdivide them further and discard the ones which contain no roots. Experiments using a large number of (simulated) processors gave a speed-up of nearly 40.

## 4.12 Miscellaneous Methods

Patrick (1972) describes a method for polynomials all of whose roots are real (a fairly common case). It takes  $O(n^3)$  operations, but on the other hand it is globally convergent, in contrast to many other methods where the determination of starting points which ensure convergence for those methods is very time-consuming.

It is based on the fact that the zeros of the second derivative of a polynomial with only real zeros can serve as starting values for Newton's method with assured convergence to zeros of the polynomial itself. Thus if the degree  $n$  is even, we start with the  $(n-2)$ 'th derivative, which is a quadratic, find its zeros by the usual formula and hence by Newton's method find 2 zeros of the  $(n-4)$ 'th derivative, which is fourth degree, and so on until we obtain zeros of the original polynomial. Thus in general we find  $2j-2$  zeros of the  $(n-2j)$ 'th derivative, which is of degree  $2j$ . The other two zeros can be found from the fact that the sum and product of all the zeros are equal respectively to  $-(\text{coefficient of } x^{2j-1})$  and the constant term, of the  $(n-2j)$ 'th derivative. Thus we get 2 equations for the missing zeros  $u$  and  $v$ , of the form

$u + v = c$ ,  $uv = d$ , hence  $u + \frac{d}{u} = c$ , hence  $u^2 + du + c = 0$ , a quadratic which can easily be solved.

If the degree  $n$  is odd, the treatment is very similar except that we start with a derivative which is linear.

Patrick proves that Newton's method, when using a zero of the second derivative of  $P_m$  as a starting value, is guaranteed to converge monotonically to a zero of  $P_m$ .

Pasquini and Trigante (1985) also give a method that is globally convergent for the all-real zero case, and uses only  $O(n^2)$  operations. It is derived by considering divided differences, but we will omit the details of the derivation as it is very lengthy. The actual algorithm is as follows:

$$x_i^{(k+1)} = x_i^{(k)} - \Delta_i \quad (i = 1, \dots, n-1) \quad (4.340)$$

$$x_n^{(k+1)} = -c_{n-1} - \sum_{i=1}^{n-1} x_i^{(k+1)} \quad (4.341)$$

where

$$\Delta_i = (P_i - \sum_{l=1}^{i-1} \pi_{li} \Delta_l) / \pi_{ii} \quad (4.342)$$

(with the sum omitted for  $i=1$ ) and

$$P_i = \sum_{j=0}^{n-i+1} c_{n-i+1-j} p_{ji} \quad (4.343)$$

$$\pi_{li} = \sum_{j=0}^{n-i} c_{n-i-j} q_{ji}^{(l)} \quad (4.344)$$

$$p_{hr} = p_{h,r-1} + x_r^{(k-1)} p_{h-1,r}; \quad p_{0r} = 1, \quad (r \geq 1);$$

$$p_{h0} = 0 \quad (h \geq 0) \quad (4.345)$$

$$q_{hr}^{(m)} = p_{hr} + x_m^{(k-1)} q_{h-1,r}^{(m)}; \quad q_{0r}^{(m)} = 1, \quad (r \geq 1, m \geq 1) \quad (4.346)$$

It is claimed that the method converges quadratically even for multiple zeros, provided that we take an average of clusters of zeros converging towards a multiple zero. For more details see the cited paper, theorem 4.

### 4.13 A Robust and Efficient program

Bini (1996) and Bini and Fiorentino (2000) have written a highly efficient and robust program, based on Aberth's method, with cluster analysis to speed convergence of multiple roots, and adaptive multiprecision arithmetic. It never failed on 1000 polynomials of degree up to 25,000.

It can be downloaded from

<http://netlib.bell-labs.com/netlib/numeralgo>

It is contained in item na20, and is called MPsolve (as of Dec. 2004, it is version 2.2). Instructions for running it are contained in the Appendix to this chapter.

#### APPENDIX. RUNNING MPsolve

- IF YOU ARE USING LINUX:

Step 1

Mpsolve uses GMP as a multiprecision arithmetic engine, in most cases it is already available on many linux distributions, if this is not the case, you can get it from:

<http://www.swox.com/gmp/>

the current version is 4.1.4:

<http://ftp.sunet.se/pub/gnu/gmp/gmp-4.1.4.tar.gz>

download it, then you can unpack and install it by issuing:

```
tar xvzf gmp-4.1.4.tar.gz
cd gmp-4.1.4
./configure
make
make install (you may need to be in root to accomplish this step)
```

## Step 2

To install MPSolve, download the package then just type

```
tar xvzf na20.tgz (this will create a directory named MPSolve-2.2)
cd MPSolve-2.2
make
make check (just to check that everything is OK).
```

### - IF YOU ARE USING WINDOWS:

It is slightly more complicated since both GMP and MPSolve use Unix-like features. You need to recreate a Unix-like environment, you can do so by installing the Cygnus package

<http://www.cygwin.com/>

To setup the whole environment simply download and run the installer:

<http://www.cygwin.com/setup.exe>

Upon successful completion, an icon will be available which will launch the new unix environment (it is a bash shell, actually).

GMP will be already available if you checked it from the list of available packages during installation. Otherwise simply run the installer again and check it in the math section.

Copy the package na20.tgz in your cygnus home directory (typically,  $C : \backslash\text{cygwin}\backslash\text{home}\backslash\text{Your\_Name}$ ) then apply Step 2 above.

- IN BOTH CASES: In the package you will find all the instructions about how to write input polynomials and how to feed them to MPSolve. The documentation also describes all the features and runtime options.

The above was written by Dario Bini and Giuseppe Fiorentino

## 4.14 References for Chapter 4

Aberth, O. (1973), Iteration Methods for Finding all Zeros of a Polynomial Simultaneously, *Math. Comp.* **27**, 339-344

Adams, D.A. (1967), A stopping criterion for polynomial root-finding, *Comm. Ass. Comp. Mach.* **10**, 655-658

Alefeld, G. and Herzberger, J. (1974), On the Convergence Speed of Some Algorithms for the Simultaneous Approximation of Polynomial Roots, *SIAM J. Numer. Anal.* **11**, 237-243

——— and ——— (1983), *Introduction to Interval Computations*, Academic Press, New York

Andreev, A.S. and Kjurkchiev, N.V. (1987), Two-Sided Method for Solving the Polynomial Equation, *Math. Balk. (New Series)* **1 (1)**, 72-82

Atanassova, L. (1996), On the R-order of a Generalization of Single-Step Weierstrass Type Methods, *Z. angew. Math. Mech.* **76**, 422-424

Batra, P. (1998), Improvement of a convergence condition for the Durand-Kerner iteration, *J. Comput. Appl. Math.* **96**, 117-125

Bini, D.A. (1996), Numerical computation of polynomial zeros by means of Aberth's method, *Numer. Algorithms* **13**, 179-200

——— and Fiorentino, G. (2000), Design, analysis, and implementation of a multiprecision polynomial rootfinder, *Numer. Algorithms* **23**, 127-173

Borsch-Supan, W. (1963), A Posteriori Error Bounds for the Zeros of Polynomials, *Numer. Math.* **5**, 380-398



————— (1970), Residuenabschätzung für Polynom-Nullstellen mittels Lagrange-Interpolation, *Numer. Math.* **14**, 287-296

Carstensen, C. (1993), On quadratic-like convergence of the means for two methods for simultaneous rootfinding of polynomials, *BIT* **33**, 64-73

————— and Petkovic, M.S. (1993), On iteration methods without derivatives for the simultaneous determination of polynomial zeros, *J. Comput. Appl. Math.* **45**, 251-266

Cosnard, M. and Fraignaud, P. (1990), Finding the roots of a polynomial on an MIMD multicomputer, *Parallel Computing* **15**, 75-85

Docev, K. (1962), An alternative method of Newton for simultaneous calculation of all the roots of a given algebraic equation, *Phys. Math. J., Bulg. Acad. Sci.* **5**, 136-139 (in Bulgarian)

————— (1962), Über Newtonsche Iterationen, *C.R. Acad. Bulg. Sci.* **15**, 699-701

————— and Byrnev, P. (1964), Certain Modifications of Newton's Method for the Approximate Solution of Algebraic Equations, *USSR Comp. Math. Math. Phys.* **4** (5), 174-182

Durand, E. (1960), *Solution Numérique des Équations Algébriques, Vol. 1, Équations du Type  $F(x) = 0$ , Racines d'une Polynôme*. Masson, Paris

Ehrlich, L.W. (1967), A Modified Newton Method for Polynomials, *Comm. Ass. Comput. Mach.* **10**, 107-108

Ellis, G.H. and Watson, L.T. (1984), A parallel algorithm for simple roots of polynomials, *Comput. Math. Appl.* **10**, 107-121

Farmer, M.R. and Loizou, G. (1975), A Class of Iteration Functions for Improving, Simultaneously, Approximations to the Zeros of a Polynomial, *BIT* **15**, 250-258

————— and ————— (1977), An algorithm for the total, or partial, factorization of a polynomial, *Math. Proc. Camb. Phil. Soc.* **82**, 427-437

Fraignaud, P. (1991), The Durand-Kerner polynomials root-finding method in case of multiple roots, *BIT* **31**, 112-123

Freeman, T.L. (1979), A method for computing all the zeros of a polynomial with real coefficients, *BIT* **19**, 321-333

- (1989), Calculating polynomial zeros on a local memory parallel computer, *Parallel Computing* **12**, 351-358
- and Bane, M.K. (1991), Asynchronous polynomial zero-finding algorithms, *Parallel Computing* **17**, 673-681
- and Brankin, R.W. (1990), A divide and conquer method for polynomial zeros, *J. Comput. Appl. Math.* **30**, 71-79
- Gargantini, I. (1978), Further Applications of Circular Arithmetic: Schroeder-like Algorithms with Error Bounds for Finding Zeros of Polynomials, *SIAM J. Numer. Anal.* **15**, 497-510
- (1979), The numerical stability of simultaneous iterations via square-rooting, *Computers Math. Appls.* **5**, 25-31
- (1980), Parallel square-root iterations for multiple roots, *Comput. Math. Appl.* **6**, 279-288
- and Henrici, P. (1971), Circular Arithmetic and the Determination of Polynomial Zeros, *Numer. Math.* **18**, 305-320
- Hansen, E. and Patrick, M. (1977), A family of root-finding methods, *Numer. Math.* **27**, 257-269
- , ——— and Rusnack, J. (1977), Some modifications of Laguerre's method, *BIT* **17**, 409-417
- Hopkins, M. et al (1994), On a Method of Weierstrass for the Simultaneous Calculation of the Roots of a Polynomial, *Z. angew. Math. Mech.* **74**, 295-306
- Hull, T.E. and Mathon, R. (1996), The Mathematical Basis and a Prototype Implementation of a New Polynomial Rootfinder with Quadratic Convergence, *ACM Trans. Math. Softw.* **22**, 261-280
- Ilieff, L. (1948-50), On the approximations of Newton, *Annual Sofia Univ.* **46**, 167-171 (in Bulgarian)
- Iliev, A.I. and Semerdzhiev, Kh.I. (1999), Some Generalizations of the Chebyshev method for Simultaneous Determination of All Roots of Polynomial Equations, *Comp. Math. Math. Phys.* **39**, 1384-1391
- Kanno, S., Kjurkchiev, N.V. and Yamamoto, T. (1996), On Some Methods for the Simultaneous Determination of Polynomial Zeros, *Japan J. Indust. Appl. Math.*

**13**, 267-288

Kerner, I.O. (1966), Ein Gesamtschritteverfahren zur Berechnung der Nullstellen von Polynomen, *Numer. Math.* **8**, 290-294

Kjellberg, G. (1984), Two Observations on Durand-Kerner's Root-Finding Method, *BIT* **24**, 556-559

Kjurkchiev, N.V. (1998), *Initial Approximations and Root-finding Methods*, Wiley-VCH, Weinheim, Germany

————— and Andreev, A. (1985), A modification of Weierstrass-Docev's method with rate of convergence  $R+2$  for the simultaneous determination of zeros of a polynomial, *C.R. Acad. Bulgare Sci.* **38**, 1461-1463 (in Russian)

————— and ————— (1992), On the Generalization of the Alefeld-Herzberger's Method, *Computing* **47**, 355-360

Leuze, M.R. (1983), A hybrid Laguerre method, *BIT* **23**, 132-138

Lo Cascio, M.L., Pasquini, L. and Trigante, D. (1989), Simultaneous Determination of Polynomial Roots and Multiplicities: An Algorithm and Related Problems, *Ricerche Mat.* **38**, 283-305

Loizou, G. (1983), Higher-Order Iteration Functions for Simultaneously Approximating Polynomial Zeros, *Intern. J. Computer Math.* **14**, 45-58

Maeder, A.J. and Wynton, S.A. (1987), Some parallel methods for polynomial root-finding, *J. Comput. Appl. Math.* **18**, 71-81

Markov, S. and Kjurkchiev, N. (1989), A Method for Solving Algebraic Equations, *Z. angew. Math. Mech.* **69**, T106-T107

Maxwell, E.A. (1960), *Advanced Algebra*, Cambridge University Press, Cambridge

Milovanovic, G.V. and Petkovic, M.S. (1983), On the Convergence Order of a Modified Method for Simultaneous Finding Polynomial Zeros, *Computing* **30**, 171-178

————— and ————— (1986), On Computational Efficiency of the Iterative Methods for the Simultaneous Approximation of Polynomial Zeros, *ACM Trans. Math. Softw.* **12**, 295-306

Miyakoda, T. (1989), Iterative methods for multiple zeros of a polynomial by clustering, *J. Comput. Appl. Math.* **28**, 315-326

————— (1992), Balanced convergence of iterative methods to a multiple zero of a complex polynomial, *J. Comput. Appl. Math.* **39**, 201-212

————— (1993), Multiplicity estimating algorithm for zeros of a complex polynomial and its application, *J. Comput. Appl. Math.* **46**, 357-368

Monsi, M. and Wolfe, M.A. (1988), Interval Versions of Some Procedures for the Simultaneous Estimation of Complex Polynomial Zeros, *Appl. Math. Comp.* **28**, 191-209

Niell, A.M. (2001), The Simultaneous Approximation of Polynomial Roots, *Comput. Math. Appl.* **41**, 1-14

Nourein, A.-W. M. (1975), An iterative formula for the simultaneous determination of the zeros of a polynomial, *J. Comput. Appl. Math.* **1(4)**, 251-254

————— (1977a), An improvement on Nourein's method for the simultaneous determination of the zeros of a polynomial (an algorithm), ALGORITHM 007, *J. Comput. Appl. Math.* **3 (2)**, 109-110

————— (1977b), An Improvement on two Iteration Methods for Simultaneous Determination of the Zeros of a Polynomial, *Intern. J. Computer Math.* **6B**, 241-252

Ostrowski, A.M. (1970), *Solution of Equations and Systems of Equations*, Academic Press, New York

Pasquini, L. and Trigante, D. (1985), A Globally Convergent Method for Simultaneously Finding Polynomial Roots, *Math. Comp.* **44**, 135-149

Patrick, M.L. (1972), A Highly Parallel Algorithm for Approximating All Zeros of a Polynomial with Only Real Zeros, *Comm. Ass. Comp. Mach.* **15**, 952-955

Peters, G. and Wilkinson, J.H. (1971), Practical problems arising from the solution of polynomial equations, *J. Inst. Math. Appl.* **8**, 16-35

Petkovic, M.S. (1981), On a Generalization of the Root Iterations for Polynomial Complex Zeros in Circular Arithmetic, *Computing* **27**, 37-55

————— (1982), On an iterative method for simultaneous inclusion of polynomial complex zeros, *J. Comput. Appl. Math.* **8**, 51-56

————— (1989A), On Halley-Like Algorithms for Simultaneous Approximation

of Polynomial Complex Zeros, *SIAM J. Numer. Anal.* **26**, 740-763

————— (1989B) *Iterative Methods for Simultaneous Inclusion of Polynomial Zeros*, Springer-Verlag, Berlin

————— (1990), On the efficiency of some combined methods for polynomial complex zeros, *J. Comput. Appl. Math.* **30**, 99-115

————— (1996), Asynchronous Methods for Simultaneous Inclusion of Polynomial Roots, in *Numerical Methods and Error Bounds*, ed. G. Alefeld and J. Herzberger, Akademie Verlag, Berlin, 178-187

————— (2003), Laguerre-like inclusion methods for polynomial zeros, *J. Comput. Appl. Math.* **152**, 451-465

————— and Carstensen, C. (1993), Some improved inclusion methods for polynomial roots with Weierstrass' correction, *Computers Math. Appl.* **25 (3)**, 59-67

—————, ————— and Trajkovic, M. (1995), Weierstrass' formula and zero-finding methods, *Numer. Math.* **69**, 353-372

————— and Herceg, D.D. (1998), On the convergence of Wang-Zheng's method, *J. Comput. Appl. Math.* **91**, 123-135

————— and ————— (2001), Point estimation of simultaneous methods for solving polynomials: a survey, *J. Comput. Appl. Math.* **136**, 283-307

—————, ————— and Ilic, S. (1998), Safe convergence of simultaneous methods for polynomial zeros, *Numer. Algorithms* **17**, 313-331

—————, Ilic, S. and Trickovic, S. (1997), A Family of Simultaneous Zero-Finding Methods, *Computers Math. Appls.* **34 (10)**, 49-59

————— and Kjurkchiev, N. (1997), A note on the convergence of the Weierstrass SOR method for polynomial roots, *J. Comput. Appl. Math.* **80**, 163-168

————— and Milovanovic, G.V. (1983), A note on some improvements of the simultaneous methods for determination of polynomial zeros, *J. Comput. Appl. Math.* **9(1)**, 65-69

————— and ————— (1985), Computational efficiency of the simultaneous methods for finding polynomial zeros: comparison of various algorithms, in *Numerical Methods and Approximation Theory*, ed. D. Herceg, University of Novi Sad, 89-93

- , ————— and Stefanovic, L.V. (1986), Some higher-order methods for the simultaneous approximation of multiple polynomial zeros, *Computers Math. Appls.* **12A**, 951-962
- , Petkovic, L.D. and Herceg, D.D. (1998), Point Estimation of a Family of Simultaneous Zero-Finding Methods, *Computers Math. Appls.* **36 (2)**, 1-12
- , ————— and Ilic, S. (2003), The Guaranteed Convergence of Laguerre-Like Methods, *Computers Math. Appls.* **46**, 239-251
- and Rancic, L. (2004), On the guaranteed convergence of the square-root iteration method, *J. Comput. Appl. Math.* **170**, 169-179
- , Sakurai, T. and Rancic, L. (2004), Family of simultaneous methods of Hansen-Patrick type, *Appl. Numer. Math.* **50**, 489-510
- and Stefanovic, L.V. (1984), The numerical stability of the generalized root iteration for polynomial zeros, *Computers Math. Appls.* **10 (2)**, 97-106
- and Stefanovic, L.V. (1986A), On a Second Order Method for the Simultaneous Inclusion of Polynomial Complex Zeros in Rectangular Arithmetic, *Computing* **36**, 249-261
- and ————— (1986B), On some improvements of square root iteration for polynomial complex zeros, *J. Comput. Appl. Math.* **15**, 13-25
- and ————— (1987), On some iteration functions for the simultaneous computation of multiple complex polynomial zeros, *BIT* **27**, 111-122
- and ————— (1990), Forward-Backward Serial Iteration Methods for Simultaneously Approximating Polynomial Zeros, *Intern. J. Computer Math.* **37**, 227-238
- , ————— and Marjanovic, Z. M. (1992), A family of simultaneous zero-finding methods, *Intern. J. Computer Math.* **43**, 111-126
- , ————— and ————— (1993), On the R-order of Some Accelerated Methods for the Simultaneous Finding of Polynomial Zeros, *Computing* **49**, 349-361
- and Vranic, D.V. (2000), The Convergence of Euler-Like Method for the Simultaneous Inclusion of Polynomial Zeros, *Computers Math. Appls.* **39(7/8)**, 95-105

Sakurai, T., Torii, T. and Sugiura, H. (1991), A high-order iterative formula for simultaneous determination of zeros of a polynomial, *J. Comput. Appl. Math.* **38**, 387-397

Semerdzhev, K. (1994), Iterative Methods for Simultaneous Finding All Roots of Generalized Polynomial Equations, *Math. Balk. (New Series)* **8**, 311-335

Simeunovic, D.M. (1989), On the Convergence of an Iterative Procedure for the Simultaneous Determination of all Zeros of a Polynomial, *Z. angew. Math. Mech.* **69**, T108-110

Small, R.D. (1976), Problem 75-14, Simultaneous Iteration Towards All Roots of a Complex Polynomial, *SIAM Rev.* **18**, 501-502

Sun, F. and Li, X. (1999), On an accelerating quasi-Newton circular iteration, *Appl. Math. Comp.* **106**, 17-29

——— and Zhang, X. (2003), A new method of increasing the order of convergence step-by-step, *Appl. Math. Comp.* **137**, 15-32

Wang, D. and Wu, Y.-j. (1987), Some Modifications of the Parallel Halley Iterative Method and Their Convergence, *Computing* **38**, 75-87

——— and Zhao, F. (1995), The theory of Smale's point estimation and its application, *J. Comput. Appl. Math.* **60**, 253-269

Weierstrass, K. (1903), Neuer Beweis des Satzes, dass jede ganze rationale Function einer Veränderlichen dargestellt werden kann als ein Product aus linearen Functionen derselben Veränderlichen, *Ges. Werke* **3**, 251-269

Werner, W. (1982), On the simultaneous determination of polynomial zeros, in *Iterative Solution of Nonlinear Systems of Equations*, eds. R. Ansorge et al, Springer-Verlag, Berlin, 188-202

Yamamoto, T. (1996), SOR-like Methods for the Simultaneous Determination of Polynomial Zeros, in *Numerical Methods and Error Bounds*, G. Alefeld and J. Herzberger (eds), Akademie Verlag, Berlin, 287-296

Zhao, F. and Wang, D. (1993), The theory of Smale's point estimation and the convergence of Durand-Kerner program, *Math. Numer. Sinica* **15**, 196-206 (in Chinese)

## Chapter 5

# Newton's and Related Methods

### 5.1 Definitions and Derivations

Newton's method in its modern form is given by the iteration

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (i = 0, 1, \dots, m) \quad (5.1)$$

starting with an initial guess  $x_0$ , and with  $m$  determined by some stopping criterion, i.e. we terminate the iteration process when accuracy is considered good enough.  $f(x)$  could be any function, although in our context it is a polynomial.

This is probably the most widely known method for solving equations, although it is not the most efficient, nor the most robust. That is, it is not guaranteed to converge from an arbitrary starting point  $x_0$ , so it is usually used in conjunction with some other method that is globally convergent (e.g. bisection). This should give an  $x_0$  close enough to the root  $\zeta$  for Newton's method to converge.

Many different derivations are given in the literature, but the most straightforward is based on Taylor's theorem (see e.g. Stoer and Bulirsch (1986)), i.e.

$$\begin{aligned} f(\zeta) = 0 &= f(x_0 + [\zeta - x_0]) = f(x_0) + (\zeta - x_0)f'(x_0) + \\ &\frac{(\zeta - x_0)^2}{2}f^{(2)}(x_0) + \dots \end{aligned} \quad (5.2)$$

If the powers after the first are ignored, we get

$$0 = f(x_0) + (\bar{\zeta} - x_0)f'(x_0) \quad (5.3)$$

where  $\bar{\zeta}$  is a new approximation to  $\zeta$ , hopefully better than  $x_0$ . Thus

$$\bar{\zeta} = x_0 - \frac{f(x_0)}{f'(x_0)} \quad (5.4)$$



We set  $x_1 = \bar{\zeta}$  and repeat the process to give  $x_2, x_3, \dots$ , i.e. we have established 5.1. We continue until we believe the desired accuracy has been attained (see section 6 for stopping criteria).

Several authors, such as Stoer and Bulirsch, derive a similar method of higher order by keeping the second degree term in 5.2, i.e.

$$\bar{\zeta} = x_0 - \frac{f'(x_0) \pm \sqrt{(f'(x_0))^2 - 2f(x_0)f''(x_0)}}{f''(x_0)} \quad (5.5)$$

In general an iterative method may be written as

$$x_{i+1} = \phi(x_i) \quad (i = 0, 1, 2, \dots) \quad (5.6)$$

The solutions are "fixed points" i.e. given by

$$\zeta = \phi(\zeta) \quad (5.7)$$

Now if the first  $p-1$  derivatives  $\phi'(\zeta) = \phi''(\zeta) = \dots = \phi^{(p-1)}(\zeta)$  are zero, but

$$\phi^{(p)}(\zeta) \neq 0 \quad (5.8)$$

then Taylor's theorem gives

$$x_{i+1} = \phi(x_i) = \phi(\zeta) + \frac{(x_i - \zeta)^p}{p!} \phi^{(p)}(\zeta) + O(|x_i - \zeta|^{p+1}) \quad (5.9)$$

But  $\phi(\zeta) = \zeta$  by 5.7 so we have

$$\lim_{i \rightarrow \infty} \frac{x_{i+1} - \zeta}{(x_i - \zeta)^p} = \frac{\phi^{(p)}(\zeta)}{p!} \quad (5.10)$$

i.e. convergence is of order  $p$ , with "asymptotic error constant"

$$C \equiv \frac{\phi^{(p)}(\zeta)}{p!}, \quad (\text{i.e. } x_{i+1} - \zeta = C(x_i - \zeta)^p) \quad (5.11)$$

With

$$\phi = x - \frac{f}{f'} \quad (5.12)$$

as for Newton's method we get

$$\phi' = 1 - \frac{(f')^2 - ff''}{(f')^2} = \frac{f(x)f''(x)}{(f'(x))^2} \quad (5.13)$$

so  $\phi'(\zeta) = 0$  and Newton's method is of order at least 2. But

$$\phi'' = \frac{(f'f'' + ff''')(f')^2 - ff''(2f'f'')}{(f')^4} \quad (5.14)$$

so

$$\phi''(\zeta) = \frac{f''(\zeta)}{f'(\zeta)} \text{ (since } f(\zeta) = 0) \quad (5.15)$$

and the asymptotic error constant is

$$\frac{f''(\zeta)}{2f'(\zeta)} \quad (5.16)$$

It can be proved (see e.g. Matthews (1992) theorem 2.2) that if  $y = \phi(x)$  satisfies  $a \leq y \leq b$  for all  $a \leq x \leq b$ , then  $\phi$  has a fixed point in  $[a, b]$ . Further, if

$$|\phi'(x)| \leq K < 1 \text{ for } x \in [a, b] \quad (5.17)$$

then  $x_{i+1} = \phi(x_i)$  ( $i = 0, 1, 2, \dots$ ) (5.6) converges to  $\zeta$ . For Newton's method 5.17 becomes: "Newton converges if

$$\left| \frac{f(x)f''(x)}{(f'(x))^2} \right| < 1 \text{ for } x \in [a, b]'' \quad (5.18)$$

This is mostly of theoretical interest, as it may be hard to estimate the values of  $f$  and its derivatives over a range of  $x$ . More practical conditions are given in section 3.

Another condition for convergence is given by Ostrowski (1973) and quoted by several other authors such as Presic (1978). It follows: Let  $f(x)$  be a real function of a real variable  $x$ ,  $f(x_0)f'(x_0) \neq 0$ , and put

$$h_0 = -\frac{f(x_0)}{f'(x_0)}, \quad x_1 = x_0 + h_0 \quad (5.19)$$

Consider the interval  $J_0 = \langle x_0, x_0 + 2h_0 \rangle$  and assume that  $f''(x)$  exists in  $J_0$ , that

$$\text{Sup}_{J_0} |f''(x)| = M \quad (5.20)$$

and

$$2h_0M \leq |f'(x_0)| \quad (5.21)$$

Let

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (i = 0, 1, \dots) \quad (5.22)$$

Then all  $x_i$  lie in  $J_0$  and  $x_i \rightarrow \zeta$  as  $i \rightarrow \infty$  where  $\zeta$  is the only zero in  $J_0$ . Unless  $\zeta = x_0 + 2h_0$ ,  $\zeta$  is simple. Further

$$(a) \left| \frac{x_{i+1} - x_i}{(x_i - x_{i-1})^2} \right| \leq \frac{M}{2|f'(x_i)|} \quad (i = 1, 2, \dots) \quad (5.23)$$

$$(b) |\zeta - x_{i+1}| \leq \frac{M}{2|f'(x_i)|} |x_i - x_{i-1}|^2 \quad (5.24)$$

For proof see Ostrowski (1973) pp57-59. There is a similar result for a complex function of a complex variable (see Ostrowski pp59-60). Again this is of mostly theoretical interest, as  $M$  is hard to evaluate except for very low-degree polynomials.

A further condition (again only theoretically useful) is given by Franklin (1881), quoting Fourier, as follows:

If  $f(x)$  has only one root between  $a$  and  $b$ , and if  $f''(x)$  does not change sign between those limits, then Newton's method converges provided we start at that one of the points  $a$  or  $b$  for which  $f$  has the same sign as  $f''$ . For proof see quoted paper.

A useful variation on Newton's method, which originated with Fourier, is described by Householder (1970) among others. Let the interval  $I = [x_0, t_0]$  contain a simple zero, assume  $f'(x)f''(x) \neq 0$  on  $I$ , and

$$f(x_0)f''(x_0) > 0 \quad (5.25)$$

Form the sequence  $x_0, x_1, x_2, \dots$  by Newton's method, and also form

$$t_{i+1} = t_i - \frac{f(t_i)}{f'(t_i)} \quad (5.26)$$

Then both sequences converge monotonically to the root  $\zeta$  from opposite directions, and

$$\lim_{i \rightarrow \infty} \frac{(t_{i+1} - x_{i+1})}{(t_i - x_i)^2} = \frac{f''(\zeta)}{2f'(\zeta)} \quad (5.27)$$

We may also show that

$$|\zeta - x_i| \leq |t_i - x_i| \quad (5.28)$$

For proof see Householder pp156-157.

Another variation often quoted is to let

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_0)} \quad (i = 0, 1, \dots, k) \quad (5.29)$$

i.e. we need only **one** evaluation per iteration. When  $i$  reaches  $k$  we restart with  $f'(x_k)$  in the denominator, and so on. The order of a set of  $k$  iterations is  $k+1$ , according to Ortega and Rheinboldt (1970). Thus the efficiency is  $\log(\sqrt[k+1]{k+1})$ , which is maximum for  $k=2$ .

Kung and Traub (1976) show that, among all rational iterations using two function evaluations, the most efficient is either Newton's or

$$x_{i+1} = x_i - \frac{f^2(x_i)}{f(x_i + f(x_i)) - f(x_i)} \quad (5.30)$$

Atkinson (1989) shows that provided  $f'(x)$  does not change rapidly between  $x_i$  and  $\zeta$  (as is usually the case), then

$$\zeta - x_i \approx x_{i+1} - x_i \quad (5.31)$$

which gives a practical stopping criterion.

## 5.2 Early History of Newton's Method

Strictly speaking, the method commonly known as "Newton's" or "Newton-Raphson's" is not really due to either of these gentlemen, but rather to Thomas Simpson (1740).

On the other hand Newton and Raphson did publish methods which are equivalent to the modern formulation

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (5.32)$$

The problem is that (unlike Simpson) they did not use calculus to compute  $f'(x_i)$ , but rather some strictly algebraic methods based on the binomial theorem (which are comparatively laborious).

Newton's version of the method was first written down in a tract "De analysi..." in 1669, although not published in its own right until 1711 (it was published as part of a book by Wallis (1685)). An English translation appears in Whiteside (1967-1976). It may be described in modern notation as follows: let  $x_0$  be a first guess at the solution  $\zeta$  of  $f(x) = 0$ . Write

$$g_0(x) = f(x) = \sum_{i=0}^n c_i x^i \quad (5.33)$$

Writing  $e_0 = \zeta - x_0$  and using the binomial expansion we get

$$0 = g_0(\zeta) = g_0(x_0 + e_0) = \sum_{i=0}^n c_i (x_0 + e_0)^i = \sum_{i=0}^n c_i \left[ \sum_{j=0}^i C_j^i x_0^j e_0^{i-j} \right] \quad (5.34)$$

$$= g_1(e_0) \quad (5.35)$$

N.B.  $x_0$  is a constant here, while the variable is  $e_0$ .

Neglecting terms involving powers of  $e_0$  higher than the first gives

$$0 = g_1(e_0) \approx \sum_{i=0}^n c_i [x_0^i + i x_0^{i-1} e_0] = \sum_{i=0}^n c_i x_0^i + e_0 \sum_{i=1}^n i c_i x_0^{i-1} \quad (5.36)$$

from which we deduce

$$e_0 \approx b_0 = -\frac{\sum_{i=0}^n c_i x_0^i}{\sum_{i=1}^n i c_i x_0^{i-1}} \quad (5.37)$$

and set  $x_1 = x_0 + b_0$ . Now repeat the process, but instead of expanding the original equation  $g_0$  about  $x_1$  expand the **new** polynomial  $g_1$  of the RHS of 5.34 about  $b_0$ , i.e. write  $g_1(e_0) = g_1(b_0 + e_1) = g_2(e_1)$

In the modern notation of the calculus 5.37 could be written

$$c_0 = -\frac{g_0(x_0)}{g_0'(x_0)} = -\frac{f(x_0)}{f'(x_0)} \quad (5.38)$$

but Newton (or Raphson) did not seem to realize the possibilities of 5.38, which is much easier to calculate than the process 5.34-5.37. In fact Newton did not even describe the above general formulation, but restricted himself to one single (now famous) example  $x^3 - 2x - 5 = 0$ .

Raphson's version was first published in 1690 in a tract (Raphson 1690). Raphson's treatment was similar to Newton's, inasmuch as he used the binomial theorem, but was more general. He treats the equation  $a^3 - ba - c = 0$  in the unknown  $a$ , and states that if  $g$  is an estimate of the solution  $\zeta$ , a better estimate is given by  $g+x$  where

$$x = \frac{c + bg - g^3}{3g^2 - b} \quad (5.39)$$

Successive corrections are obtained by substituting in the **original** equation, rather than using a new equation each time as in Newton's version. Raphson gave explicit formulas similar to 5.39 for polynomials of all degrees up to the tenth.

As mentioned, the first formulation in terms of calculus was given by Simpson in 1740. Even he did not give it in its modern form, but in terms of "fluxions" of the form  $\dot{y}$  or  $\frac{dy}{dt}$ , to be divided by  $\dot{x}$  to give  $\frac{dy}{dx}$  or  $f'(x)$ .

Lagrange (1798) gave the modern formula, mentioning Newton and Raphson but not Simpson. Fourier in 1831 ascribed the method to Newton, with no mention of Raphson or Simpson. The great popularity of Fourier may account for the use of the term "Newton's method", with no mention of Simpson until perhaps Kollerstrom (1992) and Ypma (1995).

### 5.3 Computable Conditions for Convergence

Conditions for safe (i.e. guaranteed) convergence of Newton’s method from a starting point  $x_0$ , which were given in the literature prior to about 1978, were difficult or impossible to evaluate, as they involved knowledge of  $f(x)$  and some of its derivatives over a range of  $x$  (e.g.  $M$  in Ostrowski’s condition 5.3 in Section 1 =

$$\text{Max}_{x_0 \leq x \leq x_0 + 2h_0} |f''(x)| \tag{5.40}$$

Or, even worse, they often involved knowledge of the roots.

As far as we know, the first paper giving a computable condition was by Presic (1978). Theorem 2 of that paper is the same as Ostrowski’s condition quoted in Section 1 of this chapter, except that instead of

$$2M|h_0| \leq |f'(x_0)| \text{ where } M = \text{Max}_{J_0} |f''(x)| \tag{5.41}$$

we now have

$$M(x_0)|h_0| \leq q|f'(x_0)| \tag{5.42}$$

where

$$M(x) = \text{Max}\{|f'(x)|, |f''(x)|, \dots, f^{(n)}(x)\} \tag{5.43}$$

(for a fixed  $x$ ) is relatively easy to compute. Here  $q \approx .3466$  Also, the conclusions (a) (5.23) and (b) (5.24) are replaced by

$$(a') \frac{|x_{i+1} - x_i|}{|x_i - x_{i-1}|^2} \leq r \frac{M(x_i)}{2|f'(x_i)|} \quad (i = 1, 2, \dots) \tag{5.44}$$

$$(b') |x_{i+1} - \zeta| \leq r \frac{M(x_i)}{2|f'(x_i)|} |x_i - x_{i-1}|^2 \quad (i = 1, 2, \dots) \tag{5.45}$$

Here  $r = .5631$ . For proof see the quoted paper.

The above is for real  $x_i$ , but in a second paper Presic (1979) proves a very similar result for complex  $x_i$ , except that now  $q = .2314$  and  $r = .5409$ .

To use these conditions (and others to be described later), we may use some globally convergent but slow method (e.g. bisection) until the condition in use is satisfied, and then switch to Newton’s method.

Later Smale (1986) gave a more elaborate criterion for safe convergence from  $z_0$ . He defines an “approximate zero” (complex or real)  $z_0$  as satisfying

$$|z_i - z_{i-1}| \leq \left(\frac{1}{2}\right)^{2^{i-1}-1} |z_1 - z_0| \quad (i = 1, 2, \dots) \tag{5.46}$$

Of course  $z_i \rightarrow \zeta$  as  $i \rightarrow \infty$  in this case. Also he defines

$$\alpha(z, f) = \left| \frac{f(z)}{f'(z)} \right| \sup_{k>1} \left| \frac{f^{(k)}(z)}{k!f'(z)} \right|^{\frac{1}{k-1}} \tag{5.47}$$

Then his theorem A states: "There is a number  $\alpha_0 \approx .1307$  such that if  $\alpha(z_0, f) < \alpha_0$ , then  $z_0$  is an approximate zero of  $f$ ".

Note that  $f(z_0), f'(z_0), \dots, f^{(k)}(z_0) \dots$  are relatively easy to evaluate in principle, but may take a large amount of computing power for high degree polynomials. So Smale's theorem B is useful, as it gives a much easier-to-evaluate criterion, i.e.

$$\alpha(z, f) \leq |f|_{max} \frac{|f(z)| \phi'_d(|z|)^2}{|f'(z)|^2 \phi_d(|z|)} \tag{5.48}$$

where

$$|f|_{max} = \sup_i |c_i| \tag{5.49}$$

$$\phi_d(r) = \sum_{i=0}^d r^i = \frac{1 - r^{d+1}}{1 - r} \tag{5.50}$$

Then if the R.H.S. of 5.48 (for  $z = z_0$ ) is  $\leq \alpha_0$ , it follows that  $\alpha(z, f) < \alpha_0$  and theorem A tells us that  $z_0$  is an approximate zero, i.e. Newton's method starting from  $z_0$  converges to a root. It is not clear whether Smale's criteria are easier or harder to satisfy than Presic's— this would be an interesting research topic.

Petkovic et al (1977) quote Wang and Han (1989) as follows: " If  $\alpha = \alpha(z, f) \leq 3 - 2\sqrt{2}$ , then

$$|z^{(k+1)} - z^{(k)}| \leq \frac{(1 - q^{2^k})K}{2\alpha(1 - \eta q^{2^k - 1})(1 - \eta q^{2^{k+1} - 1})} \eta q^{2^k - 1} |z^{(1)} - z^{(0)}| \tag{5.51}$$

where

$$K = \sqrt{(1 + \alpha)^2 - 8\alpha} \tag{5.52}$$

$$\eta = \frac{1 + \alpha - K}{1 + \alpha + K}, \quad q = \frac{1 - \alpha - K}{1 - \alpha + K} \tag{5.53}$$

Since  $3 - 2\sqrt{2} \approx .1716 > .1307$ , this is easier to fulfill than Smale's criterion.

Typical values of q are :

- 1 for  $\alpha = .17$
- .17 for  $\alpha = .1$
- 0 for  $\alpha = 0$ .

Another approach is taken by Hubbard et al (2001); they construct a finite set of points such that for every root of a polynomial of degree  $d$ , at least one of these points will provide a starting point for convergence to this root under Newton's iteration. They assume that all the roots are in the unit disk  $|z| < 1$ ; if instead the roots are in the disk  $|z| < r$ , we may scale all the starting points by a factor of  $r$ . We select our set of points so that there is at least one in the "immediate basin" of every root (i.e. the Newton iterations are guaranteed to converge from that point). To do this we take

$$s = \lceil .26632 \log d \rceil + 1 \tag{5.54}$$

circles centered at the origin with

$$N = \lceil 8.32547 d \log d \rceil + 1 \tag{5.55}$$

points on each circle. (Here  $\lceil x \rceil$  means 'integer part of  $x$ '). Let

$$r_\nu = (1 + \sqrt{2}) \left( \frac{d-1}{d} \right)^{\frac{2\nu-1}{4s}} \text{ and } \theta_j = \frac{2\pi j}{N} \text{ for } 1 \leq \nu \leq s \text{ and } 0 \leq j-1 \leq N-1 \tag{5.56}$$

i.e. we have a collection of  $Ns$  points

$$r_\nu \exp(i\theta_j) \tag{5.57}$$

The number of circles is usually quite small, e.g. for degree  $\leq 42$ ,  $s$  is  $= 1$ ; for  $d > 42$  and  $\leq 1825$ ,  $s = 2$ , and so on.

We are trying to find points  $\hat{\zeta}_1, \hat{\zeta}_2, \dots, \hat{\zeta}_d$  which approximate the  $d$  roots of  $f$  so that  $|\hat{\zeta}_j - \zeta_j| < \epsilon$ . First make a guess  $K$  at the number of iterations required, such as

$$K = \lceil d \log \left( \frac{1 + \sqrt{2}}{\epsilon} \right) \rceil + 1 \tag{5.58}$$

This is approximately the number of iterations required when the root is multiple, and should be close to the worst case. Then, for each point  $z_0$  in our set of starting points  $S_d$ , apply Newton's method at most  $K$  times, stopping when

$$|z_n - z_{n-1}| < \frac{\epsilon}{d} \tag{5.59}$$

According to Henrici (1974) Cor. 6.4g, this guarantees that there is a root with

$$|z_n - \zeta_j| < \epsilon \tag{5.60}$$

for some  $j$ . The authors recommend that each successive point  $z_0$  should have an argument differing by  $\frac{2\pi}{d}$  from the previous one.



If the root  $\zeta_j$  approximated by  $z_n$  is different from any previously found root, set  $\hat{\zeta}_j = z_n$ . Otherwise discard this set of iterations entirely. If Newton's method has been applied  $K$  times from  $z_0$  without locating a root, save the value  $z_K$  in a new set  $S_d^1$  for possible future use. Also, if  $z_k > 1 + \sqrt{2}$  for any  $k > 1$ , store  $z_k$  in  $S_d^1$ . If, after trying all the points in our initial set the number of roots found is  $< d$ , then begin again, starting from points in  $S_d^1$  and saving non-convergent points in a new set  $S_d^2$ . Continue thus until all  $d$  roots are found. When we have an approximation to a root  $\zeta_j$ , we must decide whether it approximates a previously found root or a new one. Kim and Sutherland (1994), lemma 2.7, describe how to do this. In fact these authors describe an asymptotically very efficient root-finding method. See the cited paper for details. Schleicher (2002) gives an upper bound on the numbers of iterations required to find each root with an accuracy  $\epsilon$ . It is

$$\frac{9\pi d^4 f_d^2}{\epsilon^2 \log 2} + 1 + \frac{|\log \epsilon| + \log 13}{\log 2} \quad (5.61)$$

where

$$f_d = \frac{d^2(d-1)}{2(2d-1)} C_d^{2d} \quad (5.62)$$

He conjectures that this is a gross over-estimate, and that a more realistic bound is

$$d \log\left(\frac{2}{\epsilon}\right) \quad (5.63)$$

Carniel (1994) gives a method based on finite-sized cells. If the region where we suspect roots and cycles is bounded by

$$x_i^L \leq x_i \leq x_i^U \quad (i = 1, 2) \quad (5.64)$$

then we divide each range into  $N_i$  equal subintervals of size  $\frac{x_i^U - x_i^L}{N_i}$  ( $i = 1, 2$ ). Now the iterations are given by a "cell-to-cell" mapping whereby the image of the cell  $\mathbf{z}$  is the center of the cell to which the image  $C(z)$  of the center of the cell  $\mathbf{z}$  belongs. Newton's method may converge not only to a fixed point (root), but also to a cycle. Cycles of period  $k$  are approximately detected by the condition

$$C^k(z^*(1)) = z^*(k+1) = z^*(1) \text{ where } z^*(m+1) = C^m(z^*(1)) \quad (5.65)$$

Fixed points are regarded as cycles of period 1. The author shows how to detect cycles and fixed points, but there are some problems if the iterations go outside the region of interest. Of course we may enlarge the region of interest, but this may result in excessive use of space and time resources. He suggests as an alternative the use of polar and especially spherical coordinates. The latter sends  $\infty$  to a finite point if we use stereographic projection. Polar or spherical coordinates have the advantage of generating relatively small-sized cells near the origin, where the roots are usually found, and bigger cells far from the origin, where great detail is not

required. He suggests the following algorithm:

a) Transform the point  $\mathbf{y}_n$  (given in spherical coordinates) into cartesian coordinates  $\mathbf{x}_n$ .

b) Apply Newton's method to  $\mathbf{x}_n$  to give  $\mathbf{x}_{n+1}$ .

c) Transform  $\mathbf{x}_{n+1}$  back to spherical coordinates  $\mathbf{y}_{n+1}$

If the "latitude" and "longitude" of a point on a sphere of radius 1 are  $\phi$  and  $\lambda$ , then the cartesian coordinates in the central plane of the sphere, projected from the "North Pole", are given by

$$x = \frac{\cos\phi \cos\lambda}{1 - \sin\phi}, \quad y = \frac{\cos\phi \sin\lambda}{1 - \sin\phi} \quad (5.66)$$

while the reverse is given by

$$\sin\phi = \frac{(x^2 + y^2) - 1}{(x^2 + y^2) + 1}, \quad \tan\lambda = \frac{y}{x} \quad (5.67)$$

(the last relation has to be modified if  $x$  and/or  $y$  are negative).

For the important case of a real polynomial with all real roots, Cosnard and Masse (1983) prove that Newton's method converges from almost any starting point. Apparently a similar result was proved by Barna (1951).

## 5.4 Generalizations of Newton's Method

This section describes a considerable number of methods, usually more efficient than Newton's itself (which has efficiency .15). They are based on the evaluation of  $f$  and/or  $f'$  at some point(s) other than  $x_i$  (as well as that point). Also included are some that involve multiplying  $f$  and/or  $f'$  by a power of  $x_i$  or a constant, as well as some further miscellaneous ones. The first set of methods will be listed more or less in order of increasing efficiency. It is worth noting that **none** of this first set have efficiencies as high as Muller's, which requires only **one** new function evaluation per step.

We will give in detail the derivation of a third-order method due to Jarratt (1966), requiring one function and two derivative evaluations per step. (normally,

because of lack of space, we will not give details of derivations). We start by supposing that

$$x_{i+1} = x_i - \frac{f(x_i)}{a_1 w_1(x_i) + a_2 w_2(x_i)} \quad (5.68)$$

where

$$w_1(x_i) = f'(x_i), \quad w_2(x_i) = f'(x_i + \alpha u(x_i)) \quad \text{and} \quad u = \frac{f}{f'} \quad (5.69)$$

Assume a simple root at  $\zeta$  and let

$$\epsilon = \epsilon_i = x_i - \zeta \quad (5.70)$$

By Taylor's series

$$f(x_i) = c_0 + c_1 \epsilon + c_2 \epsilon^2 + O(\epsilon^3) \quad (5.71)$$

Where

$$c_r = \frac{f^{(r)}(\zeta)}{r!}, \quad c_0 = f(\zeta) = 0 \quad (5.72)$$

Similarly

$$f'(x_i) = c_1 + 2c_2 \epsilon + 3c_3 \epsilon^2 + O(\epsilon^3) \quad (5.73)$$

And

$$f''(x_i) = 2c_2 + 6c_3 \epsilon + 12c_4 \epsilon^2 + O(\epsilon^3) \quad (5.74)$$

$$f^{(3)}(x_i) = 6c_3 + 24c_4 \epsilon + \dots \quad (5.75)$$

Then

$$\begin{aligned} u(x_i) &= \frac{f(x_i)}{f'(x_i)} = (c_1 \epsilon + c_2 \epsilon^2 + \dots) c_1^{-1} (1 + 2 \frac{c_2}{c_1} \epsilon + 3 \frac{c_3}{c_1} \epsilon^2 + \dots)^{-1} = \\ &\epsilon - \frac{c_2}{c_1} \epsilon^2 + O(\epsilon^3) \end{aligned} \quad (5.76)$$

So

$$\begin{aligned} w_2(x_i) &= f'(x_i + \alpha u(x_i)) = \\ &f'(x_i) + f''(x_i) \alpha u(x_i) + \frac{f^{(3)}(x_i)}{2} \alpha^2 u^2(x_i) = \\ &c_1 + 2c_2 \epsilon + 3c_3 \epsilon^2 + (2c_2 + 6c_3 \epsilon) \alpha (\epsilon - \frac{c_2}{c_1} \epsilon^2) \end{aligned} \quad (5.77)$$

$$+\frac{1}{2}(6c_3 + 24c_4\epsilon)\alpha^2\left(\epsilon - \frac{c_2}{c_1}\epsilon^2\right)^2 + O(\epsilon^3) = \quad (5.78)$$

$$c_1 + 2c_2(1 + \alpha)\epsilon + [3c_3(1 + 2\alpha + \alpha^2) - 2\frac{c_2^2}{c_1}\alpha]\epsilon^2 + O(\epsilon^3) \quad (5.79)$$

Consequently

$$\begin{aligned} a_1 w_1(x_i) + a_2 w_2(x_i) &= \\ a_1(c_1 + 2c_2\epsilon + 3c_3\epsilon^2) + a_2\{c_1 + 2c_2(1 + \alpha)\epsilon + \\ [3c_3(1 + \alpha)^2 - 2\frac{c_2^2}{c_1}\alpha]\epsilon^2\} &= c_1(a_1 + a_2) \\ + 2c_2[a_1 + (1 + \alpha)a_2]\epsilon + [3c_3a_1 + a_2\{3c_3(1 + \alpha)^2 - 2\frac{c_2^2}{c_1}\alpha\}]\epsilon^2 + O(\epsilon^3) \end{aligned} \quad (5.80)$$

We write this as

$$p_1 + p_2\epsilon + p_3\epsilon^2 + O(\epsilon^3) \quad (5.81)$$

with the obvious meanings for  $p_1, p_2, p_3$ . Substituting in 5.68 gives

$$\begin{aligned} \epsilon_{i+1} = x_{i+1} - \zeta &= \epsilon - (c_1\epsilon + c_2\epsilon^2 + c_3\epsilon^3)(p_1 + p_2\epsilon + p_3\epsilon^2)^{-1} = \quad (5.82) \\ \left(1 - \frac{c_1}{p_1}\right)\epsilon_i + \frac{1}{p_1}\left(\frac{p_2}{p_1}c_1 - c_2\right)\epsilon_i^2 + \frac{1}{p_1}\left[\frac{p_2}{p_1}c_2 + \left(\frac{p_3}{p_1} - \frac{p_2^2}{p_1^2}\right)c_1 - c_3\right]\epsilon_i^3 + O(\epsilon_i^4) \end{aligned} \quad (5.83)$$

We will choose our parameters  $a_1, a_2, \alpha$  to ensure that the order is 3, i.e. we require

$$1 - \frac{c_1}{p_1} = \frac{1}{p_1}\left(\frac{p_2}{p_1}c_1 - c_2\right) = 0 \quad (5.84)$$

i.e.  $1 - \frac{1}{a_1 + a_2} = 0$ , so that

$$a_1 + a_2 = 1 \text{ and } p_1 = c_1 \quad (5.85)$$

Then the second part of 5.84 gives  $p_2 = c_2$  i.e.

$$2(a_1 + a_2 + \alpha a_2)c_2 = c_2 \text{ i.e. } 2(1 + \alpha a_2) = 1, \text{ i.e. } 2\alpha a_2 = -1 \quad (5.86)$$

Or, expressing  $a_1$  and  $a_2$  in terms of  $\alpha$ ,

$$a_1 = \frac{1 + 2\alpha}{2\alpha}, \quad a_2 = -\frac{1}{2\alpha} \quad (5.87)$$

Choosing  $\alpha = -\frac{1}{2}$  gives the simple formula

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i - \frac{1}{2}\frac{f(x_i)}{f'(x_i)})} \quad (5.88)$$

This was also derived by Kogan (1967), who generalized it. We have ensured that this method has order 3, so the efficiency =  $\log \sqrt[3]{3} = .159$ ; this is somewhat better than Newton.

Homeier (2005) uses Newton's integral theorem for the inverse function  $x(y)$ :

$$x(y) = x(y_i) + \int_{y_i}^y x'(\eta) d\eta \quad (5.89)$$

and replaces the integral by an interpolating quadrature rule, such as the Trapezoidal rule, giving

$$x_{i+1} = x_i - \frac{f(x_i)}{2} \left( \frac{1}{f'(x_i)} + \frac{1}{f'(x_i - \frac{f(x_i)}{f'(x_i)})} \right) \quad (5.90)$$

Again, this is third order and uses 3 evaluations, so the efficiency is .159. Frontini and Somani (2003) and Weerakom and Fernando (2000) give similar formulas based on the same ideas. In one test case the last-mentioned method was **much** more efficient than Newton's.

Kizner (1964) uses the relation

$$\zeta = \int_{f(z_1)}^0 \frac{dz}{df} df + z_1 \quad (5.91)$$

Approximating the integral by the rectangular rule gives Newton's method, but Kizner applies the classical Runge-Kutta method, which requires 5 function evaluations and gives order 5. Thus the efficiency is  $\log(\sqrt[5]{5}) = .140$ , which is less than Newton. On the other hand, the author claims that this method often converges when Newton's does not.

Jarratt (1966) also gives a method of order 5, namely:

$$x_{i+1} = x_i - \frac{f(x_i)}{\frac{1}{6}[f'(x_i) + f'(x_i - \frac{f(x_i)}{f'(x_i)})] + \frac{2}{3}f'(x_i - \frac{1}{8}\frac{f(x_i)}{f'(x_i)}) - \frac{3}{8}\frac{f(x_i)}{f'(x_i - \frac{f(x_i)}{f'(x_i)})}} \quad (5.92)$$

It is seen that this uses one function and three derivative evaluations, so the efficiency is  $\log(\sqrt[4]{5}) = .175$ . The derivation is a generalization of that for the third-order method mentioned earlier.

King (1971) gives another 5th order method:

$$w_i = x_i - \frac{f(x_i)}{f'(x_i)} \quad (5.93)$$

$$x_{i+1} = w_i - \frac{f(w_i)}{f'(w_i)} - \frac{f(x_i)}{f'(x_i)} \left[ \frac{f(w_i)}{f(x_i)} \right]^3 \quad (5.94)$$

This uses 2 function and 2 derivative evaluations, so the efficiency is again  $\log(\sqrt[4]{5}) = .175$

Murakami (1978) also gives a 5th order method with 4 evaluations.

Werner (1982) gives a generalized method:

$$x_{i+1}^{(0)} = x_i^{(m-1)} \quad (5.95)$$

$$x_{i+1}^{(k)} = x_{i+1}^{(k-1)} - \frac{1}{f'(\frac{1}{2}(x_i^{(m-1)} + x_i^m))} f(x_{i+1}^{(k-1)}) \quad (5.96)$$

( $i = 0, 1, 2, \dots$ ;  $k = 1, 2, \dots, m$ ;  $m \geq 2$ ) with given starting points  $x_0^{(m-1)}, x_0^m$ . The order is shown to be

$$\frac{m}{2} + \sqrt{\frac{m^2}{4} + 1} \quad (5.97)$$

e.g.  $m = 3$  gives order 3.303 and efficiency .173. The case  $m = 2$  gives

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(\frac{x_i + y_i}{2})} \quad (5.98)$$

$$y_{i+1} = x_{i+1} - \frac{f(x_{i+1})}{f'(\frac{x_i + y_i}{2})} \quad (5.99)$$

This is of order  $1 + \sqrt{2}$  and efficiency .191. It is probably the most efficient of the class 5.95-5.96. Werner also gives another variation in which 5.98 is followed by:

$$x_{i+1}^* = x_{i+1} - 2 \frac{f(x_{i+1})}{f'(\frac{x_i + y_i}{2})} \quad (5.100)$$

$$y_{i+1} = \frac{1}{2}(x_{i+1} + x_{i+1}^*) \quad (5.101)$$

and shows that (if we start close enough to a root  $\zeta$ ) that  $\min(x_i, x_i^*)$  and  $\max(x_i, x_i^*)$  converge from below and above to  $\zeta$  with order  $1 + \sqrt{2}$ , thus giving a good error estimate.

Neta (1979) gives a sixth-order method using 3 function and 1 derivative evaluations. It is

$$w_i = x_i - \frac{f(x_i)}{f'(x_i)} \quad (5.102)$$

$$z_i = w_i - \frac{f(w_i) f(x_i) - \frac{1}{2}f(w_i)}{f'(x_i) f(x_i) - \frac{5}{2}f(w_i)} \quad (5.103)$$

$$x_{i+1} = z_i - \frac{f(z_i) f(x_i) - f(w_i)}{f'(x_i) f(x_i) - 3f(w_i)} \quad (5.104)$$

and it has efficiency  $\log(\sqrt[4]{6}) = .195$ .

Also King (1973) has a fourth order family of methods with 2 function and 1 derivative evaluation per step, including:

$$w_i = x_i - \frac{f(x_i)}{f'(x_i)} \quad (5.105)$$

$$x_{i+1} = w_i - \frac{f(w_i) f(x_i)}{f'(x_i) f(x_i) - 2f(w_i)} \quad (5.106)$$

The efficiency is  $\log(\sqrt[3]{4}) = .2007$

Jarratt (1970), p12 eq (16), gives a similar formula with the same efficiency, while Jarratt (1966B) gives another formula with that same order and efficiency, namely:

$$x_{i+1} = x_i - \frac{1}{2} \frac{f(x_i)}{f'(x_i)} + \frac{f(x_i)}{f'(x_i) - 3f'(x_i - \frac{2}{3} \frac{f(x_i)}{f'(x_i)})} \quad (5.107)$$

Earlier in the 1970 paper that author describes two methods of order 2.732 with 2 evaluations, i.e. efficiency .218 (p9 eq 10 and p10 eq 12).

Kung and Traub (1974) give a family of inverse Hermite interpolatory formulas of which the first 3 members are:

$$w_1 = x; \quad w_2 = x - \frac{f(x)}{f'(x)} \quad (5.108)$$

$$w_3 = w_2 - \frac{f(x)f(w_2)}{[f(x) - f(w_2)]^2} \frac{f(x)}{f'(x)} \quad (5.109)$$

An Algol program is given to construct higher-order methods.  $w_n$  requires  $n-1$  function evaluations and 1 derivative and is of order  $2^{n-1}$ ; thus its efficiency is  $\log(2^{1-\frac{1}{n}})$ . For example,  $n=4$  gives  $\log(2^{\frac{3}{4}}) = .226$ . They conjecture that the order of any iteration with  $n$  evaluations and no memory is at most  $2^{n-1}$ .

King (1972) describes what he calls the "tangent- parabola method". It is rather complicated but has high efficiency. Let

$$a_0 = f_0 - f_2 + f'_1(x_2 - x_0) \quad (5.110)$$

$$b_0 = 2x_1(f_2 - f_0) + f'_1(x_0^2 - x_2^2) \quad (5.111)$$

$$c_0 = x_2(2x_1 - x_2)f_0 + x_0(x_0 - 2x_1)f_2 + x_0x_2(x_2 - x_0)f'_1 \quad (5.112)$$

where

$$f_i = f(x_i) \text{ and } f'_i = f'(x_i) \quad (i = 0, 1, 2, \dots) \quad (5.113)$$

Also let

$$A_1 = f'_1 - f'_3, \quad B_1 = 2(x_1f'_3 - x_3f'_1), \quad (5.114)$$

$$C_1 = 2(x_1 - x_3)f_2 - x_2^2A_1 - x_2B_1 \quad (5.115)$$

Then we compute

$$x_3 = \frac{1}{2} \left( x_2 + \frac{-b_0 \pm \sqrt{b_0^2 - 4a_0c_0}}{2a_0} \right) \quad (5.116)$$

and

$$x_4 = \left( \frac{-B_1 \pm \sqrt{B_1^2 - 4A_1C_1}}{2A_1} \right) \quad (5.117)$$

These 2 substeps may be repeated until convergence as usual. The order is 3 and 2 new evaluations are required per full step, so that the efficiency =  $\log(\sqrt[3]{3}) = .238$ .

Neta (1981) gives an even more efficient method of order 16 with 5 evaluations. It is given by: Let

$$w_i = x_i - \frac{f(x_i)}{f'(x_i)} \quad (5.118)$$

$$z_i = w_i - \frac{f(w_i)}{f'(x_i)} \frac{f(x_i) + 2f(w_i)}{f(x_i)} \quad (5.119)$$

Now let

$$F_\delta = f(\delta_i) - f(x_i) \quad (5.120)$$

$$\phi_\delta = \frac{\delta_i - x_i}{F_\delta^2} - \frac{1}{F_\delta f'(x_i)} \quad (5.121)$$

where  $\delta = w$  or  $z$ , (e.g. if  $\delta = w$ ,  $\delta_i = w_i$ ,  $F_\delta = f(w_i) - f(x_i)$ ).

Next compute

$$D = \frac{\phi_w - \phi_z}{F_w - F_z}, \quad \gamma = \phi_w - DF_w \quad (5.122)$$

$$t_i = x_i - \frac{f(x_i)}{f'(x_i)} + \gamma f^2(x_i) - Df^3(x_i) \quad (5.123)$$



$$F_t = f(t_i) - f(x_i), \phi_t = \frac{t_i - x_i}{F_t^2} - \frac{1}{F_t f'(x_i)} \quad (5.124)$$

$$e = \frac{\frac{\phi_t - \phi_z}{F_t - F_z} - \frac{\phi_w - \phi_z}{F_w - F_z}}{F_t - F_w} \quad (5.125)$$

$$d = \frac{\phi_t - \phi_z}{F_t - F_z} - e(F_t + F_z), c = \phi_t - dF_t - eF_t^2 \quad (5.126)$$

and finally

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} + cf^2(x_i) - df^3(x_i) + ef^4(x_i) \quad (5.127)$$

The efficiency is  $\log(\sqrt[5]{16}) = .241$ . The last two methods have much "overhead", i.e. calculations other than evaluations, and so would not be suitable for low-degree polynomials. Muller's method, which is even more efficient (.265), also has a fairly high overhead.

We turn now to some different types of methods, for example Dawson (1982) generates 2 quadratics  $y = g_i(x)$  ( $i = 1, 2$ ) defined by

$$y_i = g_i(x_i) \quad (5.128)$$

$$y'_i = g'_i(x_i) \quad (5.129)$$

and  $g_i(r) = 0$  such that

$$g'_1(r) = g'_2(r) \quad (5.130)$$

Assuming that  $y_1 y_2 < 0$  he shows that  $r$  is unique, and it is given by

$$r = \frac{x_1 + x_2}{2} - \frac{y_1 - y_2}{y'_1 - y'_2} \pm \left[ \left( \frac{x_1 - x_2}{2} \right)^2 + \left( \frac{y_1 - y_2}{y'_1 - y'_2} \right)^2 - \frac{(x_1 - x_2)(y_1 + y_2)}{y'_1 - y'_2} \right]^{\frac{1}{2}} \quad (5.131)$$

the + or - sign being taken according to the sign of  $\frac{y_1 - y_2}{y'_1 - y'_2}$ . Here we use  $y_i$  for  $f(x_i)$ .

$$\text{(When } y'_1 = y'_2, r = \frac{y_1 x_2 - y_2 x_1}{y_1 - y_2} \text{)} \quad (5.132)$$

$r$  is taken as the next approximation to the root. The efficiency is the same as Newton, but the method compares favourably to several methods in the earlier literature.

Costabile et al (2001) give another method based on quadratic interpolation. Let

$$P_2[f, a, b] = (f_b - f_a - f'_a(b-a))\left(\frac{x-a}{b-a}\right)^2 + f'_a(x-a) + f_a \quad (5.133)$$

which satisfies

$$P_2(a) = f_a, P_2(b) = f_b, P'_2(a) = f'_a \quad (5.134)$$

Here of course  $f_a = f(a)$  etc. If we take  $x_0 = b$  and  $x_1 = a$  as starting points, then  $x_{i+1}$  is taken as the root of  $P_2[f, x_i, x_{i+1}] = 0$ , i.e.

$$x_{i+1} = x_i - \frac{2f_i}{f'_i \pm \sqrt{f'^2_i - 4\sigma_i f_i}} \quad (5.135)$$

where  $f_i = f(x_i)$  etc, and

$$\sigma_i = \frac{f_{i-1} - f_i - f'_i(x_{i-1} - x_i)}{(x_{i-1} - x_i)^2} \quad (5.136)$$

We select  $x_0$  and  $x_1$  so that  $f(x_0)f(x_1) < 0$  and calculate  $x_2$  as the unique root of  $P_2$  inside  $[x_0, x_1]$ . It is shown that this solution exists and is unique. Then we define the new  $x_1$  as  $x_0$  or (the old)  $x_1$  so that  $f$  takes opposite signs at the edges of the new interval  $[x_1, x_2]$ . Finally we iterate this process to convergence (which is **guaranteed**). The author shows that the order is  $1 + \sqrt{2}$ , and as 2 evaluations are required the efficiency is  $\log(\sqrt[2]{1 + \sqrt{2}}) = .190$ . This is not as efficient as some of the methods mentioned above, but as convergence (to a real root) is guaranteed, the method may be quite useful.

Alefeld and Potra (1995) describe several rather complicated bracketing methods for real roots based on inverse cubic interpolation. The best has order  $2 + \sqrt{7}$  for 3 evaluations per step (asymptotically- i.e. near the root), thus efficiency .22. Again, as it uses bracketing, convergence is guaranteed.

Clegg (1981) suggests the formula

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i) - \frac{r}{x_i}f(x_i)}, \quad (x_i \neq 0) \quad (5.137)$$

He describes several ways of choosing  $r$ , but we suspect that they may be too expensive in practise. One of them is more robust than Newton.

Hines (1951) gives a formula equivalent to 5.137, and claims that for a range of  $r$  it is faster than Newton, but does not show how to choose  $r$ .

He (2004) gives another variation:

$$x_{i+1}^r = x_i^r - r x_i^{r-1} \frac{f(x_i)}{f'(x_i)} \quad (5.138)$$

( $r = 1$  gives Newton).

The author reports experiments in which a higher value of  $r$  gives a greatly increased rate of convergence compared to Newton, or gives convergence where Newton diverges. He suggests solving the equation

$$\frac{\partial x_{i+1}}{\partial r} = 0 \quad (5.139)$$

to find the optimum  $r$ . However calculations by this author on one of He's examples did not agree with his results.

Wu (2000) gives the formula

$$x_{i+1} = x_i - \frac{f(x_i)}{q_i f(x_i) + f'(x_i)} \quad (i = 0, 1, 2, \dots) \quad (5.140)$$

where  $q_i$  is chosen so that  $q_i f(x_i)$  and  $f'(x_i)$  have the same sign. In experiments with 5 simple functions, 5.140 converged much faster than Newton in one case, and converged quite fast in the others, although Newton failed in those cases. In fact the method is almost globally convergent for real roots. Apart from the condition on  $q_i f(x_i)$  and  $f'(x_i)$  given above, the author does not explain how  $q_i$  is chosen.

Tikhonov (1976) offers the following generalization: let

$$u = \frac{f(x_i)}{f'(x_i)} \quad (5.141)$$

then

$$x_{i+1} = x_i \left[ 1 - \frac{1}{m + \frac{x_i}{u} - (s_0 + \frac{s_1}{x_i} + \dots + \frac{s_{m-1}}{x_i^{m-1}})} \right] \quad (5.142)$$

( $m=0,1,2,\dots$ ). N.B.  $m = 0$  gives Newton. Here the  $s_i$  are the sums of the  $i$ 'th powers of all the roots, and are given by Newton's identities:

$$s_0 = n, \quad s_1 = -c_{n-1}, \quad s_2 = -c_{n-1}s_1 + 2c_{n-2}, \quad \text{etc} \quad (5.143)$$

It is claimed that this converges faster than Newton, especially if the largest magnitude root is computed first. In an example, starting from -47, Newton reached the value -7.01 after 15 iterations, whereas 5.142 (with  $m = 1$ ) reached -7.0001 after 4 (the true root being exactly -7).

Finally Burgstahler (1986) describes an interesting method which replaces each power of  $x$ , other than the leading power and constant term, by a multiple of the leading power and a new constant term, using

$$\left(\frac{x}{R}\right)^p \approx \frac{p}{n}\left(\frac{x}{R}\right)^n + \frac{(n-p)}{n} \quad (p = n-1, n-2, \dots, 1) \quad (5.144)$$

which he proves true if

$$|x - R| \ll |R| \quad (5.145)$$

The result is

$$\frac{f'(R)}{nR^{n-1}}x_1^n = \frac{Rf'(R)}{n} - f(R) \quad (5.146)$$

$$\text{or } x_1 = R\left[1 - \frac{nf(R)}{Rf'(R)}\right]^{\frac{1}{n}} \quad (5.147)$$

taking the complex root closest to  $(1,0)$ . Experiments show that the new method is often, but by no means always, faster than Newton. The author suggests running both algorithms in parallel and choosing the result of whichever converges faster (or sometimes one may converge and the other not).

## 5.5 Methods for Multiple Roots

Rall (1966) shows that the unmodified Newton method converges linearly to a multiple root. Details of his proof follow: Assume the method is converging towards a root  $\zeta$  of multiplicity  $m$ . Let

$$\epsilon_i = x_i - \zeta \quad (5.148)$$

$$\text{and } \eta_i = -\frac{f(x_i)}{f'(x_i)} \quad (5.149)$$

then since

$$x_{i+1} = x_i + \eta_i \quad (5.150)$$

we have

$$\epsilon_{i+1} = \epsilon_i + \eta_i \quad (5.151)$$

Expanding  $f(x_i)$  and  $f'(x_i)$  about  $\zeta$  by Taylor's theorem gives

$$\eta_i = -\frac{\frac{1}{m!}f^{(m)}(\zeta + \theta\epsilon_i)\epsilon_i^m}{\frac{1}{(m-1)!}f^{(m)}(\zeta + \theta'\epsilon_i)\epsilon_i^{m-1}} \quad (5.152)$$

with  $0 < \theta, \theta' < 1$ . (Note that  $f(\zeta) = f'(\zeta) = \dots = f^{(m-1)}(\zeta) = 0$ ). Or, since

$$f^{(m)}(\zeta + \theta\epsilon_i) = f^{(m)}(\zeta + \theta'\epsilon_i) + f^{(m+1)}(\zeta + \theta''(\theta - \theta')\epsilon_i)(\theta - \theta')\epsilon_i \quad (5.153)$$

where  $0 < \theta, \theta'' < 1$ , it follows that

$$\eta_i = -\frac{1}{m}\epsilon_i + O(\epsilon_i^2) \quad (5.154)$$

Substituting in 5.151 gives

$$\epsilon_{i+1} = \frac{m-1}{m}\epsilon_i + O(\epsilon_i^2) \quad (5.155)$$

Now he defines

$$\rho_i = \frac{\eta_{i+1}}{\eta_i} = \frac{\epsilon_{i+2} - \epsilon_{i+1}}{\epsilon_{i+1} - \epsilon_i} \quad (5.156)$$

and shows that

$$\rho_i = \frac{m-1}{m} + O(\epsilon_i) \quad (5.157)$$

$$\text{so } \text{Lim}_{i \rightarrow \infty} \rho_i = \frac{m-1}{m} \quad (5.158)$$

Hence we can find  $m$ , for large enough  $i$ , that is when  $\rho_i$  stabilizes.

Rall also describes the “corrected Newton method”

$$\tilde{x}_{i+1} = \tilde{x}_i - m \frac{f(\tilde{x}_i)}{f'(\tilde{x}_i)} \quad (5.159)$$

(originally suggested by Schroder (1870)).

He shows that

$$\tilde{\eta}_i = -\tilde{\epsilon}_i + O(\tilde{\epsilon}_i^2) \quad (5.160)$$

so that

$$\tilde{\epsilon}_{i+1} = \tilde{\epsilon}_i + \tilde{\eta}_i = O(\tilde{\epsilon}_i^2) \quad (5.161)$$

Traub (1967) shows that

$$\text{Lim}_{i \rightarrow \infty} \frac{\tilde{\epsilon}_{i+1}}{\tilde{\epsilon}_i^2} = \frac{f^{(m+1)}(\zeta)}{m(m+1)f^{(m)}(\zeta)} \quad (5.162)$$

The problem of course is to find  $m$ . In fact McNamee (1998) has compared (for speed) 7 methods for finding  $m$ , as follows:

(A) Schroder (1870) uses

$$m = \frac{1}{u'} \text{ where } u = \frac{f}{f'} \quad (5.163)$$

$$\text{so } u' = \frac{(f')^2 - ff''}{(f')^2} \quad (5.164)$$

and 5.159 becomes

$$x_{i+1} = x_i - \frac{f(x_i)f'(x_i)}{[f'(x_i)]^2 - f(x_i)f''(x_i)} \quad (5.165)$$

Note that if we use 5.165 we do not need to evaluate  $m$  explicitly.

(B) Ostrowski (1973) uses

$$\tilde{m} = \frac{x_{2i} - x_{2i+1}}{x_{2i} - 2x_{2i+1} + x_{2i+2}} \quad (5.166)$$

where  $x_{2i+1}$  and  $x_{2i+2}$  are obtained from  $x_{2i}$  by two pure Newton steps. Then we round  $\tilde{m}$  to the nearest integer  $m$  and apply 5.159.

(C) Madsen (1973) forms  $x_i + pdx_i$  for  $p = 1, 2, \dots$  where

$$dx_i = -\frac{f(x_i)}{f'(x_i)} \quad (5.167)$$

terminating when  $|f(x_i + pdx_i)|$  starts to increase. He takes  $m =$  that  $p$  which gives the minimum  $|f(x_i + pdx_i)|$ .

(D) Hansen and Patrick (1976) give a rather complicated procedure. See their paper or the one of McNamee for details.

(E) Chanabasappa (1979) takes

$$S_a = |af_i f_i'' - (a-1)(f_i')^2| \quad (5.168)$$

for  $a = 1, 2, \dots, n$  and takes  $m =$  that value of  $a$  which gives minimum  $S_a$ .

(F) Van der Straeten and Van de Vel (1992) set

$$m_0 = 1, \quad x_1 = x_0 - m_0 \frac{f(x_0)}{f'(x_0)} \quad (5.169)$$

and for  $i = 0, 1, \dots$

$$m_{i+1} = \frac{m_i}{1 - \frac{f(x_{i+1})f'(x_i)}{f'(x_{i+1})f(x_i)}} \quad (5.170)$$

$$x_{i+2} = x_{i+1} - m_{i+1} \frac{f(x_{i+1})}{f'(x_{i+1})} \quad (5.171)$$

(G) Traub (1964) uses

$$m = \frac{\ln(f(x_i))}{\ln\left(\frac{f(x_i)}{f'(x_i)}\right)} \quad (5.172)$$

rounded to the nearest integer.

McNamee applied all the above methods to over 500 polynomials, mostly random. He found that Madsen's method was fastest with Aitken (see below), Ostrowski and Schroder close. McNamee also tested two methods which do not compute  $m$ , but apply acceleration to the linearly converging sequence produced by pure Newton iteration. One is due to Aitken (1926): after each pair of Newton steps giving  $x_{2i+1}$ ,  $x_{2i+2}$  we compute

$$\tilde{x}_{2i+2} = x_{2i} - \frac{(x_{2i+1} - x_{2i})^2}{x_{2i+2} - 2x_{2i+1} + x_{2i}} \quad (5.173)$$

The other is due to Levin (1973). It is quite complicated and the tests showed that it is relatively slow, so it will not be explained here. As mentioned above, Aitken's is among the better methods of those tested.

Not only is Schroder's method 5.165 among the fastest of those tested, but according to Gilbert (1994) it is much more reliable than those that compute  $m$  explicitly and use 5.159, at least in the presence of rounding error.

Lagouanelle (1966) describes a variation on Schroder's method 5.163 of finding  $m$ , namely

$$m = l - 1 + \text{Lim}_{x \rightarrow \zeta} \left[ \frac{[f^{(l)}(x)]^2}{[f^{(l)}(x)]^2 - f^{(l-1)}(x)f^{(l+1)}(x)} \right] \quad (5.174)$$

Then we may apply Newton's or some other method with  $f^{(m-1)}(x)$  in place of  $f(x)$ . He does not state how to choose  $l$ , but Derr (1959) gives a similar method in which  $l_i$  is chosen to be the smallest non-negative integer such that

$$|f^{(l_i+1)}(x_i)| \geq \eta = (\text{say}) \sqrt{\epsilon} \quad (5.175)$$

where  $\epsilon$  = machine precision (e.g.  $10^{-8}$ ). Then take

$$k_i = l_i + \hat{\theta} - 1 \quad (5.176)$$

where  $\hat{\theta}$  is the nearest integer to

$$\theta = \frac{\frac{f^{(l_i)}}{f^{(l_i+1)}}}{\frac{f^{(l_i)}}{f^{(l_i+1)}} - \frac{f^{(l_i-1)}}{f^{(l_i)}}} \quad (5.177)$$

where all the derivatives of  $f$  are evaluated at  $x_i$ . Our next iteration is then given by

$$x_{i+1} = x_i - (k_i - l_i) \frac{f^{(l_i)}(x_i)}{f^{(l_i+1)}(x_i)} \quad (5.178)$$

Derr shows that as  $i \rightarrow \infty$ ,  $l_i = k_i - 1 = m - 1$  and 5.178 becomes

$$x_{i+1} = x_i - \frac{f^{(m-1)}(x_i)}{f^{(m)}(x_i)} \quad (5.179)$$

This process is claimed to be of second order, and Lagouanelle claims that the multiple roots are obtained as accurately as simple roots with the normal Newton iteration.

Ypma (1983) compares a number of methods in which  $f(x)$  is replaced by a function  $T(x)$  such that

$$\text{Lim}_{x \rightarrow \zeta} T(x) = 0, \text{Lim}_{x \rightarrow \zeta} T'(x) \neq 0 \quad (5.180)$$

i.e.  $T(x)$  has a simple root identical to a (perhaps) multiple root of  $F(x)$ . Most of the  $T(x)$  are of the form

$$\frac{f^2(x)(\alpha + \beta)}{f(x + \alpha f(x)) - f(x - \beta f(x))} \quad (5.181)$$

He then applies Newton's method to  $T(x)$ . He concludes that the most reliable and efficient case for polynomials is given by

$$T(x) = \frac{f(x)}{f'(x)} \quad (5.182)$$

just as  $u(x)$  in 5.163..leading to Schroder's method 5.165.

King (1979), (1980), (1983A), (1983B) describes a series of extrapolation methods, each one more efficient than the previous. The best (1983B) is described below. From  $x_0$  we compute  $x_1$  and  $x_2$  by two Newton steps, letting  $g_0 = x_0 - x_1$ ,  $g_1 = x_1 - x_2$ . Take  $\tilde{x}_2$  as the Aitken-extrapolation of  $(x_0, x_1, x_2)$  (i.e. apply 5.173) and set

$$g_2 = \frac{f(\tilde{x}_2)}{f'(\tilde{x}_2)} \quad (5.183)$$

Now we fit a parabola  $\tilde{g}_2$  through  $(x_0, g_0)$ ,  $(x_1, g_1)$  and  $(\tilde{x}_2, g_2)$ , compute the derivative  $\tilde{g}'_2$  at  $\tilde{x}_2$  and take a Newton-like step

$$x_3 = \tilde{x}_2 - \frac{\tilde{g}_2}{\tilde{g}'_2}, g_3 = \frac{f(x_3)}{f'(x_3)} \quad (5.184)$$



Note that

$$\tilde{g}'_2 = g[\tilde{x}_2, x_1] + g[\tilde{x}_2, x_0] - g[x_0, x_1] \quad (5.185)$$

and that  $g_2 = \tilde{g}_2$ . Steps 5.185 and 5.184 are repeated as needed. King shows that the order is 1.839, but as 2 evaluations are needed per step, the efficiency is  $\log(\sqrt[2]{1.839}) = .132$ . This compares favourably with Schroder's method 5.165 which has efficiency  $\log(\sqrt[3]{2}) = .1003$ .

Dong (1987) gives a method of order 3 requiring 3 evaluations per step, thus efficiency  $\log(\sqrt[3]{3}) = .159$ . It is

$$x_{i+1} = x_i - u_i - \frac{f(x_i)}{\left(\frac{m}{m-1}\right)^{m+1} f'(x_i - u_i) + \frac{m-m^2-1}{(m-1)^2} f'(x_i)} \quad (5.186)$$

where (presumably)

$$u_i = \frac{f(x_i)}{f'(x_i)} \quad (5.187)$$

The above assumes  $m$  is known, and may be subject to some of the unreliability reported by Gilbert. Victory and Neta (1983) give a similar method of the same order and efficiency as Dong's, but it is a little more complicated. See their paper for details.

Forsythe (1958) points out, by means of an example, that the "plain" Newton's method will converge linearly, if started from a large distance from several simple roots, until  $x_i$  is close enough to a root to "see" it as a separate entity. In that case it may be better to start by using one of the methods designed for multiple roots, such as Schroder's.

The above gives a similar result to several papers which observe that, because of rounding errors, a root which is mathematically multiple may be replaced computationally by a **cluster** of close, but not equal, roots. For example Yakoubsohn (2000) describes algorithms which detect such a cluster. We need some definitions: an  $m$ -cluster of  $f$  is defined as an open disk

$$D(z, r) = \{x : |x - z| < r\} \quad (5.188)$$

which contains  $m$  zeros of  $f$ , counting multiplicities. A full  $m$ -cluster is an  $m$ -cluster which contains  $m-1$  zeros of  $f'$ , counting multiplicities. Let

$$N_f = x - \frac{f(x)}{f'(x)} \quad (5.189)$$

$$\beta_m(f, z) = \text{Max}_{0 \leq k \leq m-1} \left| \frac{m! f^{(k)}(z)}{k! f^{(m)}(z)} \right|^{\frac{1}{m-k}} \quad (5.190)$$

$$\gamma_m(f, z) = \text{Max}_{m+1 \leq k \leq n} \left| \frac{m! f^{(k)}(z)}{k! f^{(m)}(z)} \right|^{\frac{1}{k-m}} \tag{5.191}$$

$$R_m(f, z, r) = \frac{|f^{(m)}(z)|}{m!} r^m - \sum_{k=0}^{m-1} \frac{|f^{(k)}(z)|}{k!} r^k - \sum_{k=m+1}^n \frac{|f^{(k)}(z)|}{k!} r^k \tag{5.192}$$

The first algorithm, called “m-cluster”, detects a **probable** m-cluster. From  $x_0$ , we take two Newton steps, to  $x_1$  and  $x_2$ . Then we find the integer m which minimizes

$$\left| \frac{|x_2 - x_1|}{|x_1 - x_0|} - \frac{m - 1}{m} \right| \tag{5.193}$$

Now compute

$$z = mx_2 - (m - 1)x_1 \tag{5.194}$$

and

$$r = \frac{1}{(2\gamma_m(f, z))} \tag{5.195}$$

The disk  $D(z, r)$  is a probable m-cluster. Finally, check whether  $R_m(z, r) > 0$ . If so,  $D(z, r)$  is definitely an m-cluster, otherwise not.

An arbitrary  $x_0$  will not generally lead to an m-cluster as above. Yakoubsohn describes a global Newton homotopy method which usually **does** obtain an m-cluster. Let

$$f_t(x) = f(x) - tf(x_0) \tag{5.196}$$

with  $x_0$  a given complex number. If  $z_{k-1}$  is the point at step k-1 corresponding to  $t_{k-1}$ , set

$$z_0 = x_0, y_0 = z_{k-1}; y_i = y_{i-1} - \frac{f_{t_k}(y_{i-1})}{f'_{t_k}(y_{i-1})} \tag{5.197}$$

$$(i = 1, 2, \dots, n_{it}). \text{ (N.B. } f'_{t_k} = f') \tag{5.197}$$

$$z_k = y_{n_{it}} \text{ (} k = 1, 2, \dots) \tag{5.198}$$

Let

$$\beta_k = \left| \frac{f_{t_k}(z_k)}{f'(z_k)} \right| \tag{5.199}$$

Now if  $\beta_k > (\text{some small number}) \epsilon$  perform the previously- described algorithm “m-cluster”. If  $z_k$  is an m-cluster we are done. Otherwise set

$$t_{k+1} = \frac{(t_k + t_{k-1})}{2} \tag{5.200}$$

and continue to the next  $k$ . (N.B. We let  $t_0 = 1$ ,  $t_1 = 1 - \epsilon$ ,  $\beta_0 = 2\epsilon$ ).  
On the other hand if  $\beta_k \leq \epsilon$  and  $t_k > 0$  set

$$t_{k+1} = \max(t_k - 2(t_{k-1} - t_k), 0) \quad (5.201)$$

and continue to the next  $k$ .

Finally if  $\beta_k \leq \epsilon$  and  $t_k = 0$  and  $R_1(f, z_k, \frac{1}{2\gamma_1(f, z_k)}) > 0$  then the disk  $D(z_k, \frac{1}{2\gamma_1(f, z_k)})$  contains only one root and we stop. Otherwise apply 5.200 and continue to the next  $k$ . Yakoubsohn proves that this algorithm always terminates.

Kirrinis (1997) also gives a method of detecting clusters, but it is too complicated to describe here. See the cited article for details.

## 5.6 Termination Criteria

Since Newton's method (like most methods for roots) is iterative, we need a criterion to decide when to terminate the iteration. Note that much of the material of this section applies to other methods besides Newton's.

A popular method is to stop when

$$|x_{i+1} - x_i| < \epsilon \quad (5.202)$$

or

$$\left| \frac{x_{i+1} - x_i}{x_i} \right| < \epsilon \quad (5.203)$$

There are several problems with this. First, if  $\epsilon$  is chosen too small, 5.202 or 5.203 may never be satisfied, because rounding error will cause the LHS to increase and oscillate before they are satisfied.

At the other extreme, for some functions 5.202 or 5.203 may be satisfied although  $x_i$  is not close to a root. Donovan et al (1993) give an example where  $x_i = \sqrt{u_i}$ ,  $u_{i+1} = u_i + 1$ . The function producing this behaviour, by Newton iterations, is shown to be

$$f(x) = C \frac{\exp[-\frac{1}{2}(x^2 + x\sqrt{x^2 + 1})]}{\sqrt{x + \sqrt{x^2 + 1}}} \quad (5.204)$$

This has no real roots, but 5.202 is satisfied for any  $\epsilon$  provided  $i$  is large enough. They also show another function

$$h(x) = \sqrt[3]{x} \exp(-x^2) \quad (5.205)$$

which has a root at  $x = 0$ . Newton's iteration does not converge to this root (unless  $x_0 = 0$ ), but the  $x_i$  satisfy 5.202 for large  $i$  (e.g.  $i = 250,000$  for  $\epsilon = 10^{-3}$ ).

Garwick (1961) suggests stopping when the LHS of 5.203 starts to increase (provided it is less than .01, for in the first few iterations it may increase for reasons not to do with rounding error).

Igarashi (1985) gives an alternative stopping criterion, which also tells how accurate  $x_i$  is. We calculate  $f(x)$  by Horner's method, calling the result  $A(x)$ . Then compute  $G(x) = xf'(x) - f(x)$  by

$$G(x) = (n-1)c_n x^n + (n-2)c_{n-1} x^{n-1} + \dots + c_2 x^2 - c_0 \quad (5.206)$$

and finally

$$xf'(x) - G(x) \quad (5.207)$$

Call the latter  $B(x)$  ( $= f(x)$ ).

If we are far from a root, or using infinite precision,  $A(x)$  should equal  $B(x)$ . But near a root, if

$$f(x) = (x - \zeta)^m g(x) \quad (g(\zeta) \neq 0) \quad (5.208)$$

we have (usually)

$$|x_i f'(x_i)| = |x_i [(x_i - \zeta)^{(m-1)} g(x_i) + (x_i - \zeta)^m g'(x_i)]| > \quad (5.209)$$

$$|f(x_i)| = |(x_i - \zeta)^m g(x_i)|$$

provided  $\zeta$ , and hence  $x_i$ ,  $\neq 0$ .

Consequently  $A(x_i)$  will have more correct digits than  $B(x_i)$ . When  $x_i$  is very close to a root, both  $A(x_i)$  and  $B(x_i)$  cease to have any correct digits and the two values are completely different. The following criterion can be used to detect this situation: if

$$R(f, x_i) = \frac{|A(x_i) - B(x_i)|}{\min(|A(x_i)|, |B(x_i)|)} > 1 \quad (5.210)$$

then  $f(x_i)$  has no correct digits and we are as close to a root as we will ever get. Before this situation is reached, we may estimate the number of correct digits in  $f(x_i)$  as

$$-\log_{10} R(f, x_i) \quad (5.211)$$

We can also estimate the number of correct digits in  $x_i$ , as the number of leading digits in agreement between

$$x_i = x_{i-1} - \frac{A(x_{i-1})}{f'(x_{i-1})} \quad \text{and} \quad \hat{x}_i = x_{i-1} - \frac{B(x_{i-1})}{f'(x_{i-1})} \quad (5.212)$$

Igarashi (1982) gives another criterion: let the calculation errors in  $f(x)$  be  $\delta f(x)$ ; then we stop iterating if

$$|f(x)| \leq |\delta f(x)| \leq \sum_{i=0}^n |c_i x^i| \frac{b^{-t}}{2} \quad (5.213)$$

where  $t$  is the number of places in the mantissa in base  $b$  (usually 2).

Adams (1967) lets

$$e_0 = \frac{1}{2}|c_n|, e_k = |x|e_{k-1} + |b_{n-k}| \quad (5.214)$$

where the  $b_i$  are the coefficients in the deflated polynomial, found by Horner's method. Then he estimates the rounding error as

$$E = (e_n - \frac{1}{2}|b_0|)b^{1-t} \quad (5.215)$$

and stops when the computed  $|f(x)| < 2E$ .

Mcnamee (1988) compares the above three methods and concludes that Garwick's method is the best among them (but see later).

Vignes (1978) gives a different approach: let an exact number  $x = mb^e$  ( $m$  unlimited) be represented in the computer by  $X = Mb^E$  where  $M$  is limited to  $t$  places in base  $b$ , and usually  $e = E$ . The relative error in  $X$  is

$$\alpha = \frac{X - x}{X} = -\frac{r}{M} \text{ where } r = m - M \quad (5.216)$$

Statistically, with rounding  $\alpha$  has a mean of 0 and a standard deviation of  $.4 \times 2^{-t}$ . Let the mathematical operation  $z = x\omega y$  where  $\omega \in \{+, -, \times, /\}$  be performed on the computer as  $Z = X \Omega Y$  where  $\Omega$  is the computer equivalent of  $\omega$ . We let  $\Omega \in \{\oplus, \text{etc}\}$  (i.e. including rounding). Then the error

$$\epsilon_z = Z - z = (x + \epsilon_x)\omega(y + \epsilon_y) - x\omega y + \alpha(X\Omega Y) \quad (5.217)$$

where  $\epsilon_x$  and  $\epsilon_y$  are errors in  $X$  and  $Y$ .

Suppose some exact mathematical procedure

$$\text{proc}(d, r, +, -, \times, /, \text{funct}) \quad (5.218)$$

(where  $d$  and  $r$  are respectively data and results) is replaced on the computer by

$$\text{PROC}(D, R, \oplus, \text{etc}, \text{FUNCT}) \quad (5.219)$$

where again  $D$  and  $R$  are data and results. Each computer procedure corresponding to a different permutation of the operands in 5.219 is equally representative of 5.218. Suppose there are  $C_{op}$  of them. The generation of these is called the "permutation method". Moreover, each operator is subject to rounding error, which may go up or down. Thus we have 2 results for each operator, and if there are  $k$  elementary operations, there will be  $2^k$  results. This is called the perturbation method. If this is applied for each permutation of the operators, then there are  $2^k C_{op}$  results

in total. This total population of results will be called  $\langle R \rangle$ . Let  $R_0$  be the result corresponding to the mathematical result  $r$  without any permutations or perturbations, and  $\bar{R}$  and  $\delta$  be the mean and standard deviation of the elements of  $\langle R \rangle$ . Then it may be shown that the number  $C$  of significant digits in the result is given by

$$10^{-C} = \frac{\sqrt{(R_0 - \bar{R})^2 + \delta^2}}{|R_0|} \quad (5.220)$$

It is not possible or necessary to generate all the elements of  $\langle R \rangle$ . Instead in practise we may generate successive elements of  $\langle R \rangle$  until successive values of  $C$  agree. Usually a small number, about 3, are sufficient for this purpose. Vignes give a subroutine PEPER which performs the necessary permutations and perturbations. Since it is subtraction of nearly equal numbers (or additions of positive and negative numbers) which causes by far the most serious errors, the permutations are restricted to adds and subtracts.

For an iterative method such as Newton's, the above method gives a useful stopping criterion, i.e. stop when  $C$  given by 5.220 as applied to  $f(x)$  is  $< 1$ . This would mean that  $f(x)$  has no significant digits, as it is dominated by rounding errors. In an experiment with Newton's method on 4 related examples the conventional criterion  $|x_i - x_{i-1}| < \epsilon$  required 143 iterations while the  $C < 1$  criterion obtained the same accuracy with only 31 iterations. It is true that the permutation-perturbation method requires more work per iteration, but on the other hand applying it to  $x_i$  gives us reliably the number of correct digits in  $x_i$ , unlike the conventional criteria.

## 5.7 Interval Methods

Interval arithmetic has been described in Chapter 4 (Section 4) in connection with simultaneous methods. It can also be applied very usefully to Newton's and related methods. It has at least two advantages: firstly, as before, it provides guaranteed error bounds, and secondly it often converts methods which are only locally convergent in point-form into globally convergent ones.

We will divide this section into three parts:

- 1) methods for real roots,
- 2) methods for complex roots based on rectangular intervals,
- 3) methods for complex roots based on circular intervals (disks).

Moore (1966) appears to have pioneered the treatment of real roots, with the often-quoted equation:

$$N(X) = m(X) - \frac{f(m(X))}{F'(X)} \quad (5.221)$$

Here  $X$  is a real interval, initially  $X_0 = [a, b]$  say,  $m(X) =$  the midpoint of  $X = \frac{1}{2}(a + b)$ , and  $F'(X)$  is an interval extension of  $f'(x)$ , i.e. the range

$$\overline{f'}(X) = \{f'(w) : w \in X\} \subset F'(X) \quad (5.222)$$

and

$$f'(x) = F'([x, x]) \quad (5.223)$$

We then define a series of intervals by:

$$x_i = m(X_i); \quad N(x_i, X_i) = x_i - \frac{f(x_i)}{F'(X_i)}; \quad X_{i+1} = N(X_i) \cap X_i \quad (i = 0, 1, 2, \dots) \quad (5.224)$$

If the coefficients of  $f(x)$  are only known to lie in certain intervals, we replace  $f(m(X))$  in 5.221 by  $F(m(X))$  where  $F$  evaluates the range of  $f$  over the intervals of the coefficients. Moore shows that a necessary condition for  $N(x, X)$  to be defined is that  $X$  contains at most one zero and that such a zero must be simple (then  $F'(X) \not\ni 0$ ). Also he shows that if  $X$  in 5.221 contains a simple root  $\zeta$ , then  $\zeta \in N(x, X)$ . Consequently, either  $N(x, X) \cap X$  is empty, in which case  $X$  does not contain a zero, or else  $N(x, X) \cap X$  contains a zero if  $X$  does.

Moore shows how to use the above to find a partition of  $[a, b]$  into a set of adjacent intervals which alternately **may** contain a zero and definitely do **not**:

1. Evaluate  $F([a, b])$  and  $F'([a, b])$ .

If  $F([a, b])$  does not contain 0, the process is complete for  $[a, b]$ . Otherwise,  $F([a, b])$  contains 0 and **may** contain a zero of  $f$ . In the latter case,  $F'([a, b])$  may contain 0. If it does perform step 2.

2. Put

$$[a, b] = [a, \frac{a+b}{2}] \cup [\frac{a+b}{2}, b] \quad (5.225)$$

and begin again at step 1 for each subinterval.

If  $F'([a, b])$  does not contain 0:

3. Evaluate

$$N([a, b]) = \frac{a+b}{2} - \frac{F(\frac{a+b}{2})}{F'([a, b])} \quad (5.226)$$

Now either i)  $N([a, b]) \cap [a, b]$  is empty and  $[a, b]$  does not contain a zero of  $f$  and we are done with  $[a, b]$ , or ii)  $N([a, b]) \cap [a, b]$  is a non-trivial interval which may contain a zero of  $f$ . In the last case:

4. Put

$$[a, b] = X_1 \cup X_2 \text{ where } X_1 = N([a, b]) \cap [a, b] \quad (5.227)$$

Now since  $F'([a, b]) \not\supseteq 0$ , there can be at most one zero in  $[a, b]$ , and this (if it exists) must be in  $X_1$ . Hence  $X_2$  does not contain a zero and we are done with it. We add it to the list of subintervals into which we are decomposing the original interval. We repeat the process for  $X_1$ , from step 1.

5. When an interval is found which does not contain a zero of  $f$ , and which intersects with another such interval already found, the two are combined into a single one. The process is continued until, for the precision being used, no further splitting can be done.

Moore also shows that for a small enough interval  $X_i$  containing a simple root, and for which  $F'(X_i)$  does not contain zero, then there exists a positive  $K$  such that

$$w(X_{i+1}) \leq K(w(X_i))^2 \tag{5.228}$$

We can also determine if certain intervals are guaranteed to contain a zero; we do this by testing the sign of  $f$  in two intervals which are known **not** to contain zeros, and which are separated by a single interval which **may** contain a zero. If the two bordering intervals have **opposite sign**, then the one in between does contain a simple zero.

Dargel et al describe a detailed algorithm based on Moore's method, which obtains a decomposition of  $[a, b]$  into adjacent intervals  $[a, a_1], [a_1, b_1], \dots, [b_m, b]$  which alternately do **not** contain a root and **may** contain a root.

Moore and Dargel et al consider only the case  $f'(X) \not\supseteq 0$  (i.e. a simple root), but Hansen (1992) extends the method to the case  $0 \in f'(X)$ , in which case evaluation of  $N(x_i, X_i)$  requires the use of extended interval arithmetic, as first discussed by Hanson in an unpublished report (1968). It was later described by Hansen (1978A) as follows: if  $[c, d]$  is an interval containing 0, we let

$$\frac{1}{[c, d]} = \left[\frac{1}{d}, \infty\right] \text{ if } c = 0 \tag{5.229}$$

$$= \left[-\infty, \frac{1}{c}\right] \text{ if } d = 0 \tag{5.230}$$

$$= \left[-\infty, \frac{1}{c}\right] \cup \left[\frac{1}{d}, \infty\right] \text{ otherwise} \tag{5.231}$$

The application to  $N(x_i, X_i)$  is given below. Even though  $N(x_i, X_i)$  is not finite, the intersection  $X_{i+1} = X_i \cap N(x_i, X_i)$  is finite. We use interval arithmetic to bound rounding errors in the evaluation of  $f(x)$ , giving say  $f^I(x_i) = [a_i, b_i]$ . If  $0 \in f^I(x_i)$ , then  $x_i$  is a zero of  $f$  or is near to one. Now suppose  $0 \notin f^I(x_i)$ .



Let  $f'(X_i) = [c_i, d_i] \ni 0$ . We will be using extended interval arithmetic. Since  $0 \notin f^I(x_i)$ , either  $a_i > 0$  or  $b_i < 0$ . In the first case

$$N(x_i, X_i) = [-\infty, q_i] \text{ if } c_i = 0 \quad (5.232)$$

$$= [p_i, \infty] \text{ if } d_i = 0 \quad (5.233)$$

$$[-\infty, q_i] \cup [p_i, \infty] \text{ if } c_i < 0 < d_i \quad (5.234)$$

where

$$p_i = x_i - \frac{a_i}{c_i} \quad (5.235)$$

$$q_i = x_i - \frac{a_i}{d_i} \quad (5.236)$$

The results for  $b_i < 0$  are similar. The intersection

$$X_{i+1} = X_i \cap N(x_i, X_i) \quad (5.237)$$

may be a single interval, the union of two intervals, or empty.

Returning to the case where  $0 \in f^I(x)$  and  $0 \in f'(X)$ : this leads to  $N(x, X) = [-\infty, \infty]$  and hence  $X_{i+1} = X_i$ . Usually this means that  $X_i$  is small and contains a multiple zero of  $f(x)$ . But it can also sometimes occur if  $X_i$  is large and  $x_i$  is a zero or near one. Hansen (1992) goes on to consider termination criteria. He suggests

$$A) w(X_i) < \epsilon_X \text{ for some } \epsilon_X \quad (5.238)$$

$$B) |f(x_i)| < \epsilon_F \text{ for some } \epsilon_F \quad (5.239)$$

but points out that the choice of  $\epsilon_X$  and  $\epsilon_F$  is not easy. He then suggests

$$C) 0 \in f^I(x_i), 0 \notin f'(X_i), \text{ and } N(x_i, X_i) \supset X_i \quad (5.240)$$

(so that  $X_{i+1} = X_i$ ). This is satisfied if rounding error prevents further accuracy. Hansen suggests stopping if either (i) A and B are both satisfied, or (ii) C is satisfied.

The above criteria do not work very well in the case of multiple roots. Usually this coincides with  $f'(X) \ni 0$  (where  $X$  is small), but the latter may also be true if  $X$  is large and contains more than one simple root. Hansen includes a new criterion for this case:

$$R(X) = \frac{w(f'(X))}{w(f^I(x))} < 1024 \quad (5.241)$$

If  $f'(X) \ni 0$  and  $R > 1024$ , we split  $X$  in half and apply 5.224 to both intervals.

The algorithm possesses the following properties:

- 1) Every zero of  $f$  in  $X_0$  will be found and correctly bounded. No deflation is needed.
- 2) If there is no zero in  $X_0$ , this will be proven in a finite number of iterations.
- 3) If  $0 \notin f'(X_i)$ , convergence is reasonably rapid at the start, and asymptotically quadratic. For detailed proofs, see the quoted book.

Dimitrova (1994) gives a slightly different version of the algorithm described by Moore and Hansen: let

$$X_0 = [x_0^-, x_0^+] \quad (5.242)$$

contain several zeros of  $f(x)$  and

$$F'(X_0) = [F'^-, F'^+] \quad (5.243)$$

Then let  $x$  be an interior point of  $X_0$ , say  $m(X_0)$ . Consider the subintervals

$$X_{1,0} = [x_0^-, x], X_{2,0} = [x, x_0^+] \quad (5.244)$$

Now let

$$x_{1,1}^- = x_0^- + \frac{|f(x_0^-)|}{|F'^+|} \quad (5.245)$$

$$x_{1,1}^+ = x - \frac{|f(x)|}{|F'^+|} \quad (5.246)$$

with similar definitions for  $x_{2,1}^-$  and  $x_{2,1}^+$ . Let

$$X_{1,1} = [x_{1,1}^-, x_{1,1}^+] \quad (5.247)$$

$$X_{2,1} = [x_{2,1}^-, x_{2,1}^+] \quad (5.248)$$

If both intervals are empty there are no roots in  $X_0$ . If one is empty we disregard it and apply the above procedure to the other one. If neither is empty we apply it to both in turn, and so on. Thus we get a list  $L$  of subintervals. When we process  $X \in L$  we first compute  $F'(X)$ . If this does not contain 0, then  $f(x)$  has **at most** one zero in  $X$ . We test this by computing  $f(X)$ ; if this does not contain 0 we delete  $X$  from the list. If  $f(X)$  does contain 0 then  $f(x)$  has a unique zero in  $X$ , which we can estimate accurately by iterating the process above (5.244-5.248). For details of the case where  $F'(X)$  contains 0, see the cited paper (section 3).

It was pointed out in section 1 of this Chapter that Ostrowski's condition for convergence (equations 5.21-5.25) is of little practical value, as it is very difficult to evaluate  $M = \text{Sup}_{x \in J_0} |f''(x)|$ . However, if we use interval arithmetic, we automatically obtain a range  $[m, M]$  for  $f''(x)$  and the condition becomes computable.

See Rokne and Lancaster (1969) for an interval version of Ostrowski's condition.

Hansen (1978B) gives a method of reducing the size of successive intervals compared with the straightforward Moore's method: suppose  $x$  occurs more than once in  $f(x)$  (as is usual). Replace  $x$  by  $x_1$  in one or more places and by  $x_2$  in the remaining places. Call the result  $g(x_1, x_2)$  (note that  $g(x, x) \equiv f(x)$ ). If  $x \in X$ , the root  $\zeta \in N_2(X) =$

$$x - \frac{f(x)}{\frac{\partial}{\partial x_1}g(X_{11}, x) + \frac{\partial}{\partial x_2}g(X_{21}, X_{22})} \quad (5.249)$$

Actually  $X_{11}$ ,  $X_{21}$  and  $X_{22}$  all =  $X$ , but we wish to emphasize that they are independent. The author extends the above to the case where  $x$  occurs  $m$  times and shows that

$$\zeta \in N_m(x) = x - \frac{f(x)}{g'(X)} \quad (5.250)$$

where

$$\begin{aligned} g'(X) &= \frac{\partial}{\partial x_1}g(X_{11}, x, x, \dots, x) + \frac{\partial}{\partial x_2}g(X_{21}, X_{22}, x, \dots, x) + \dots \\ &+ \frac{\partial}{\partial x_m}g(X_{m1}, \dots, X_{mm}) \end{aligned} \quad (5.251)$$

Here many arguments are real instead of intervals, usually leading to a smaller interval result. The iterations follow as in Moore's method.

If  $0 \in X$ , we should choose  $x = 0$  to give a narrow interval  $g'(X)$ . Writing a polynomial  $p(x) = \sum_{i=0}^n c_i x^i$  as

$$g(x_1, x_2) = c_0 + \sum_{i=1}^n c_i x_1^{i-1} x_2 \quad (5.252)$$

Then

$$g'(X) = \sum_{i=1}^n c_i [(i-1)X_{11}^{i-2} x + X_{21}^{i-1}] \quad (5.253)$$

and letting  $x = 0$ ,

$$g'(X) = \sum_{i=1}^n c_i X_{21}^{i-1} \quad (5.254)$$

whereas

$$p'(X) = \sum_{i=1}^n i c_i X^{i-1} \quad (5.255)$$

so that the  $i$ 'th term is  $i$  times as wide as in  $g'(X)$ .

Since  $F'(X) \ni \{f'(x)|x \in X\}$ ,  $F'(X)$  may contain 0 even if  $f'(x)$  has constant sign on  $X$ . To circumvent this problem, Petkovic (1981) interpolates a monotonic function  $f$  on  $X = [a, b]$  containing a simple zero  $\zeta$  by

$$q(x) = A + Be^{kx} \tag{5.256}$$

at  $a$ ,  $c = \frac{a+b}{2}$ ,  $b$ .

He solves for  $A$  and  $B$  and gives an approximation to  $\zeta$  as

$$\alpha = \frac{1}{k} \log\left(-\frac{A}{B}\right) \in [a, b] \tag{5.257}$$

Then in Moore's method 5.221-5.224 we replace  $x$  by  $\alpha$  and  $F'(X)$  by the interval extension  $Q'(X)$  of  $q'(x)$  i.e.

$$Bk[e_1, e_2] = [E_1, E_2] \tag{5.258}$$

where

$$e_1 = \min(e^{ka}, e^{kb}), e_2 = \max(e^{ka}, e^{kb}) \tag{5.259}$$

Since  $e_1$  and  $e_2 > 0$ ,  $0 \notin Q'(X)$ . It is still possible that  $\alpha - \frac{f(\alpha)}{Q'(X)}$  does not contain  $\zeta$ , so Petkovic replaces  $N(x, X)$  by

$$N_\epsilon(\alpha, X) = \alpha - \frac{f(\alpha)}{[E_1, E_2] + I_\epsilon} \tag{5.260}$$

where  $I_\epsilon$  is an interval chosen so that  $N_\epsilon$  contains  $\zeta$  and  $0 \notin [E_1, E_2] + I_\epsilon$ . He shows how to do this (see the cited paper). He shows that the order of the method is almost 3.

Herzberger (1986) describes a recursive version of Moore's method thus:

For a fixed  $p$

$$X^{(0,p)} = X^{(0)} \tag{5.261}$$

$$X^{(i+1,0)} = \left\{ m(X^{(i)}) - \frac{f(m(X^{(i)}))}{f'(X^{(i,p)})} \right\} \cap X^{(i)} \tag{5.262}$$

$$X^{(i+1,k)} = \left\{ m(X^{(i+1,k-1)} - \frac{f(m(X^{(i+1,k-1)})))}{f'(X^{(i,p)})} \right\} \cap X^{(i+1,k-1)} \tag{5.263}$$

$(k = 1, \dots, p) \text{ (if } p > 0)$

$$X^{(i+1)} = \left\{ m(X^{(i+1,p)} - \frac{m(f(X^{(i+1,p)}))}{f'(X^{(i+1,p)})} \right\} \cap X^{(i+1,p)} \tag{5.264}$$

The above set of 3 equations are repeated for  $i = 0, 1, \dots$  until convergence. He shows that the order is  $p+3$ , and as  $p+3$  evaluations are needed, the efficiency is  $\log(p+3)^{\frac{1}{p+3}}$ . This is a maximum for  $p = 0$  (and then it is  $\log\sqrt[3]{3}$ )

Alefeld and Potra (1988) give a bracketing method similar to the Newton - Fourier method: suppose the interval  $[a, b]$  contains a zero of  $f$ . Let  $y_0 = a, z_0 = b$  and for  $i=0, 1, 2, \dots$  compute

$$y_{i+1} = y_i - \Delta f(y_i, z_i)^{-1} f(y_i) \tag{5.265}$$

$$\bar{z}_{i+1} = y_{i+1} - \frac{f(y_i)}{f'(y_{i+1})} \tag{5.266}$$

$$z_{i+1} = \min\{\bar{z}_{i+1}, z_i\} \tag{5.267}$$

where  $\Delta f(s, t)$  is the divided difference of  $f$  at the points  $s, t$ . A second method is given in which  $f'(y_{i+1})$  in 5.266 is replaced by  $\Delta f(y_i, y_{i+1})$ . They show that for both these methods  $y_i$  and  $z_i$  tend to the root  $\zeta$  from below and above respectively. Also the order of the first method is 3, and that of the second is  $1 + \sqrt{2}$ . As the first method takes 3 evaluations, the efficiencies are respectively  $\log\sqrt[3]{3} = .159$  and  $\log\sqrt{1 + \sqrt{2}} = .191$ . Strictly speaking these are not interval methods, but like the latter these methods provide error bounds.

Lin and Rokne (1995) give a variation on Hansen's method 5.232- 5.237 suitable for multiple roots (i.e. such that  $f'(X) \ni 0$ ). Let  $\phi$  be a point iterative method such as Schroeder's 5.167. At the general ( $i$ 'th) step  $X^{(i)}$  consists of several non-intersecting subintervals,  $X_j^{(i)}$ , ( $j = 1, \dots, l_i$ ). If  $x_i \in X_k^{(i)}$ , compute  $N(x_i, X_k^{(i)}) \cap X_k^{(i)}$ , and combine it with the other subintervals to form  $X^{(i+1)}$ . Then compute  $z_{i+1} = \phi(x_i)$ . If  $z_{i+1} \in X^{(i+1)}$  take  $x_{i+1} = z_{i+1}$ , otherwise let  $x_{i+1}$  be such that

$$|x_{i+1} - z_{i+1}| = \min_{x \in X^{(i+1)}} |x - z_{i+1}| \tag{5.268}$$

They show that the order is the same as that of  $\phi$ .

Revol (2003) describes a multiple precision version of the Moore-Hansen method.

Alefeld (1981), for a polynomial  $p(x)$ , replaces  $f'(X)$  in Moore's method by the difference quotient

$$\Delta(x, y) = \frac{p(x) - p(y)}{x - y} \in \left(\sum_{i=1}^n a_{i-1} X^{i-1}\right)_H \equiv J_1 \tag{5.269}$$

where

$$a_{i-1} = \sum_{j=i}^n c_j y^{j-1} \quad (i = 1, \dots, n) \tag{5.270}$$

and the subscript H means evaluating by Horner's method. He also gives several alternate interval expressions which contain  $\Delta(x, y)$ . He shows theoretically that  $J_1$  is narrower than the other expressions, or  $p'(X)$  (thus leading to faster convergence). This is confirmed by an example.

We turn now to methods for complex roots by rectangular intervals, starting with a method of Hansen (1968). He expresses  $p(z)$  where  $z = x_1 + ix_2$  in terms of real and imaginary parts  $f_1(x_1, x_2)$  and  $f_2(x_1, x_2)$ . Let

$$A = \frac{\partial f_1}{\partial x_1}, B = \frac{\partial f_1}{\partial x_2}, C = \frac{\partial f_2}{\partial x_1}, D = \frac{\partial f_2}{\partial x_2} \tag{5.271}$$

Suppose initially  $\zeta \in X_1^{(0)} + iX_2^{(0)}$ , then we get a new containing rectangle by

$$Y_1^{(0)} = x_1 - \frac{D(X_1^{(0)}, X_2^{(0)})f_1(x_1, x_2) - B(X_1^{(0)}, X_2^{(0)})f_2(x_1, x_2)}{DENOM} \tag{5.272}$$

$$Y_2^{(0)} = x_2 - \frac{A(X_1^{(0)}, X_2^{(0)})f_2(x_1, x_2) - C(X_1^{(0)}, x_2)f_1(x_1, x_2)}{DENOM} \tag{5.273}$$

where  $DENOM = A(X_1^{(0)}, X_2^{(0)})D(X_1^{(0)}, X_2^{(0)}) - B(X_1^{(0)}, X_2^{(0)})C(X_1^{(0)}, x_2)$  (Grant and Hitchins (1973) show that  $DENOM \neq 0$  if  $X_1^{(0)} + iX_2^{(0)}$  contains only one zero). Then we take

$$X_1^{(1)} = X_1^{(0)} \cap Y_1^{(0)} \tag{5.274}$$

$$X_2^{(1)} = X_2^{(0)} \cap Y_2^{(0)} \tag{5.275}$$

Arthur (1972) describes an interval-Bairstow method for complex roots. Bairstow's method in ordinary (real) arithmetic starts with an approximate quadratic factor

$$x^2 - px - q \tag{5.276}$$

corresponding to a root and its complex conjugate. It then computes

$$b_i = c_{n-i} + pb_{i-1} + qb_{i-2}; b_{-1} = b_{-2} = 0 \ (i = 0, 1, \dots, n) \tag{5.277}$$

$$e_i = b_i + pe_{i-1} + qe_{i-2}; e_{-1} = e_{-2} = 0 \ (i = 0, 1, \dots, n - 1) \tag{5.278}$$

In the interval version we start with an approximate factor

$$x^2 - Px - Q \tag{5.279}$$

where P and Q are intervals containing  $\tilde{p}$  and  $\tilde{q}$  respectively, where  $x^2 - \tilde{p}x - \tilde{q}$  is an exact factor. Then find  $b_{n-1}$  and  $b_n$  by 5.277 using  $p = m(P)$ ,  $q = m(Q)$ . The  $b_i$  will be intervals as rounded-interval computation is used, i.e.  $b_i$  is actually

an interval  $\hat{B}_i$ . Then use 5.277 and 5.278 with P and Q replacing p and q to find intervals  $B_i$  and  $E_i$ . Next compute

$$\delta P = \frac{\hat{B}_n E_{n-3} - \hat{B}_{n-1} E_{n-2}}{DENOM} \quad (5.280)$$

$$\delta Q = \frac{\hat{B}_{n-1} E_{n-1} - \hat{B}_n E_{n-2}}{DENOM} \quad (5.281)$$

where

$$DENOM = E_{n-2}^2 - E_{n-1} E_{n-3} \quad (5.282)$$

(if  $DENOM \ni 0$  the method breaks down). Finally set

$$\hat{P} = m(P) + \delta P, \hat{Q} = m(Q) + \delta Q \quad (5.283)$$

and

$$P^* = P \cap \hat{P}, Q^* = Q \cap \hat{Q} \quad (5.284)$$

Repeat the process; usually the iteration converges to narrow intervals  $\tilde{P}$  and  $\tilde{Q}$  containing the exact p and q. Let

$$-\tilde{P}^2 - 4\tilde{Q} = [s, t] \quad (5.285)$$

Then if the zeros are  $a \pm ib$ , we have

$$a \in \frac{1}{2}\tilde{P}, b \in \left[\frac{1}{2}\sqrt{s}, \frac{1}{2}\sqrt{t}\right] \quad (5.286)$$

unless  $[s, t]$  contains 0. For that case see the cited paper.

Arthur also suggests an interval method for multiple roots (of multiplicity m) based on Derr's method (equations 5.175- 5.179). It is:

$$Y_{i+1} = m(X_i) - \frac{f^{(m-1)}(m(X_i))}{F^{(m)}(X_i)} \quad (5.287)$$

$$X_{i+1} = X_i \cap Y_{i+1} \quad (5.288)$$

Rokne (1973) gives a rectangular interval version of Ostrowski's condition for complex roots.

Petkovic and Herzberger (1991) describe a combined interval method for multiple complex roots in rectangular arithmetic: let  $\Phi$  be a point-iterative function in "normal" (not interval) complex arithmetic, of order r. Also let

$$R^{(i+1)} = \Psi(z^{(i)}, R^{(i)}) \quad (i = 0, 1, \dots) \quad (5.289)$$

be an interval method of order  $q$  which starts from  $R^{(0)} \ni \zeta$ . They define the combined method

$$z^{(i,0)} = \text{mid}(R^{(i)}) \tag{5.290}$$

$$z^{(i,j+1)} = \Phi(z^{(i,j)}) \quad (j = 0, 1, \dots, k - 1) \tag{5.291}$$

$$z^{(i)} = z^{(i,k)} \text{ if } z^{(i,k)} \in R^{(i)} \tag{5.292}$$

$$\text{mid}(R^{(i)}) \text{ otherwise} \tag{5.293}$$

$$R^{(i+1)} = \Psi(z^{(i)}, R^{(i)}) \quad (i = 0, 1, \dots) \tag{5.294}$$

The authors show that the order of the above combined method is  $r^k \cdot q$ , and hence the efficiency is

$$\log(r^k q)^{\frac{1}{k\theta_\Phi + \theta_\Psi}} \tag{5.295}$$

$$\rightarrow r^{\frac{1}{\theta_\Phi}} \tag{5.296}$$

for large  $k$ , where  $\theta_\Phi, \theta_\Psi$  are the amounts of work involved in  $\Phi$  and  $\Psi$ .

For a root of multiplicity  $m$ , the authors suggest for  $\Psi$ :

$$R^{(i+1)} = z^{(i)} - \frac{m}{\frac{P'(z^{(i)})}{P(z^{(i)})} - (n - m)[(z^{(i)} - \text{ext}R^{(i)})^{-1}]} \tag{5.297}$$

where  $[[A]]$  means the smallest rectangle enclosing a complex set  $A$ . If a rectangle is given by  $[a,b]+i[c,d]$ , the enclosing rectangle for the inverse of its exterior is given by

$$[\underline{\alpha}, \bar{\alpha}] + i[\underline{\beta}, \bar{\beta}] \tag{5.298}$$

where

$$\underline{\alpha} = \min\{\frac{1}{a}, \frac{1}{2c}, -\frac{1}{2d}\}, \quad \bar{\alpha} = \max\{\frac{1}{b}, -\frac{1}{2c}, \frac{1}{2d}\} \tag{5.299}$$

$$\underline{\beta} = \min\{-\frac{1}{d}, \frac{1}{2a}, -\frac{1}{2b}\}, \quad \bar{\beta} = \max\{-\frac{1}{c}, -\frac{1}{2a}, -\frac{1}{2b}\} \tag{5.300}$$

To find  $m$  Petkovic and Herzberger suggest Lagouanelle's method, but this author prefers Derr's. Petkovic and Herzberger show that, if  $R^{(0)}$  is an initial rectangle (with  $z^{(0)} = \text{mid}(R^{(0)})$  and  $d^{(0)} = \text{sd}(R^{(0)})$ ) containing only one zero  $\zeta$  of multiplicity  $m$ , and if

$$\left| \frac{p(z^{(0)})}{p'(z^{(0)})} \right| < \frac{d^{(0)}}{2(m + 1)(n - m)} \tag{5.301}$$



then  $\zeta \in R^{(i)}$  ( $i = 0, 1, \dots$ ) and  $d^{(i)} \rightarrow 0$  quadratically.

For  $\Phi$  the authors suggest, among other methods, Schroeder's:

$$z^{(i+1)} = z^{(i)} - m \frac{p(z^{(i)})}{p'(z^{(i)})} \quad (5.302)$$

of order 2, or the Halley-like method of order 3:

$$z^{(i+1)} = z^{(i)} - \frac{2}{\left(1 + \frac{1}{m}\right) \frac{p'(z^{(i)})}{p(z^{(i)})} - \frac{p''(z^{(i)})}{p'(z^{(i)})}} \quad (5.303)$$

Calculating the efficiency by 5.296 shows that 5.302 is generally the most efficient of 5.302, 5.303 and three other methods considered by the authors.

Henrici (1971) uses circular intervals to refine complex zeros. His theorem 1 states: "Let  $z_0$  be a complex number, and let  $C_1; W_2, W_3, \dots, W_n$  be disks. Let

$$\frac{p'(z_0)}{p(z_0)} \in C_1 \text{ and } w_k \in W_k \text{ (} k = 2, \dots, n \text{)} \quad (5.304)$$

where  $w_1, w_2, \dots, w_n$  are the zeros of  $p(z)$ , a polynomial of degree  $n$ . Then

$$w_1 \in W_1 = z_0 - \frac{1}{C_1 - \sum_{k=2}^n \frac{1}{z_0 - W_k}} \quad (5.305)$$

He describes a Newton-like algorithm based on the above, and shows that under certain initial conditions convergence is quadratic (see the cited work for details).

Petkovic (1987) also gives a circular disk iteration: Let  $\{c, r\}$  be a disk of center  $c$ , radius  $r$ , and suppose we have found a disk  $\{z^{(0)}, r^{(0)}\} \equiv \{a, R\}$  containing exactly one zero  $\zeta$ . Let

$$h^{(i)} = \frac{\bar{a} - \overline{z^{(i)}}}{R^2 - |z^{(i)} - a|^2} \quad (5.306)$$

and

$$d^{(i)} = \frac{R}{R^2 - |z^{(i)} - a|^2} \quad (5.307)$$

then we define

$$Z^{(i+1)} = z^{(i)} - \frac{1}{\frac{p'(z^{(i)})}{p(z^{(i)})} - (n-1)\{h^{(i)}; d^{(i)}\}} \quad (5.308)$$

and

$$z^{(i+1)} = \text{mid}(Z^{(i+1)}) \quad (5.309)$$

For multiple roots (n-1) above is replaced by n-m. Under certain conditions convergence is quadratic.

Gargantini (1976) gives a kind of simultaneous Newton-like method:

$$Z_k^{(i+1)} = z_k^{(i)} - \frac{1}{\frac{p'(z_k^{(i)})}{p(z_k^{(i)})} - \sum_{j=1, \neq k}^n \frac{1}{z_k^{(i)} - Z_j^{(i)}}} \tag{5.310}$$

and shows that it has convergence order 3. A similar Laguerre-type method has order 4, but the Newton-type method is considerably more efficient.

L.D. Petkovic et al (1997) give a slope method using complex intervals. Let

$$g(z, y) = \frac{p(y) - p(z)}{y - z} \tag{5.311}$$

(note that since  $y - z$  is a factor of the numerator,  $g(y, y)$  is defined). and use

$$Z^{(i+1)} = z^{(i)} - \frac{p(z^{(i)})}{g(z^{(i)}, Z^{(i)})} \tag{5.312}$$

where  $z^{(i)} = \text{mid}(Z^{(i)})$ , or better still

$$Z^{(i+1)} = z^{(i)} - \frac{1}{\frac{p'(z^{(i)})}{p(z^{(i)})} - \frac{g'(z^{(i)}, Z^{(i)})}{g(z^{(i)}, Z^{(i)})}} \tag{5.313}$$

Usually the interval  $g(z^{(i)}, Z^{(i)})$  is narrower than  $f'(Z^{(i)})$ , so the slope method converges faster than the Newton-Moore-like methods. We may combine 5.312 with several prior iterations of the point-slope method

$$z^{(i+1)} = z^{(i)} - \frac{1}{\frac{p'(z^{(i)})}{p(z^{(i)})} - \frac{g'(z^{(i)}, z^{(i)})}{g(z^{(i)}, z^{(i)})}} \tag{5.314}$$

which has order 3, as does 5.313.

## 5.8 Parallel Methods

Akl (1989) describes a parallel implementation of Newton's method which usually overcomes the lack of global convergence (for real roots): suppose the interval [a,b] is known to contain exactly one zero of  $f(x)$ . The interval is divided into N+1 subintervals of equal size ( $N \geq 2$ ), and the division points are taken as initial approximations for Newton's method, one on each processor. As soon as the method converges on one processor, the result is written to a shared memory location ROOT (initially set to  $\infty$ ). As soon as that value is changed, all the processors stop working. In a set of iterations does not converge after (say) I iterations,

its processor stops working.

Shedler (1967) gives a slightly different parallel method: the interval  $[a, b]$  is divided into  $N+1$  subintervals as in Akl's method and  $f$  evaluated at the points of division. If  $|f(x)|$  is not  $< \epsilon_1$  at any of these points, we obtain a new approximation by linear interpolation between  $a$  and  $b$ , and  $N$  others by applying Newton's method at each of the division points. If none of the new points satisfy  $|f(x)| < \epsilon_1$ , we choose a new interval as the smallest one having a sign change between adjacent points contained in the set  $\{a, b, \text{the } N \text{ section points, and the } N+1 \text{ new approximations}\}$ . If the length of the new interval is not  $< \epsilon_2$ , we perform a further iteration. Obviously, the new points at each iteration can be found in parallel.

Wiethoff (1996) gives a parallel extended interval Newton method, which uses extended interval operations and bisection. The master (M) keeps a list of remaining intervals to be examined; it also stores whether a slave processor ( $P_1, P_2, \dots, P_n$ ) is busy or idle. The starting interval  $[x]$  is divided equally into  $n$  subintervals  $[x_1], [x_2], \dots, [x_n]$ , and each subinterval is sent to the corresponding slave. All slaves start working on their subintervals. In the case of bisection, the second interval is returned to the master to be redistributed when possible. Either it is added to the waiting list if no idle slave is available or it is sent to an idle slave. If a slave has computed a result (an interval containing a zero), it is returned to the master and the slave marked as idle (if the waiting list is empty), or it gets the next interval from the waiting list.

The algorithm for the slave  $P_i$  follows: do

1. Receive  $[y]$  from M.
2.  $[Zero] = null$  (result interval = null).
3. If  $0 \notin f([y])$  then go to 10. (null will be added to result-list)
4.  $c = \text{mid}([y])$ .
5.  $[z] = c - f(c)_{\diamond} / f'([y])$  (extended interval Newton step;  $f_{\diamond}$  is an interval bounded by the rounded up (and down) values of  $f(x)$ ).
6.  $[y_p] = [y] \cap [z]$  (intersection may contain 2 disjoint intervals  $[y_p]_1 \cup [y_p]_2$ ).
7. if  $[y_p]_1 = [y]$  (only one interval) then  
 $[y_p]_1 = [y, c], [y_p]_2 = [c, \bar{y}]$  where  $[y] = [\underline{y}, \bar{y}]$  (bisection)
8. If  $[y_p]_1 \neq null$  and  $[y_p]_2 \neq null$  then  
 Send BISECTION-SIGNAL to M; send  $[y_p]_2$  to M.
9. If  $[y_p]_1 \neq null$  then  
 if  $\frac{\text{width}([y_p]_1)}{\text{mid}([y_p]_1)} < \epsilon$  and  $0 \in f([y_p]_1)$  then  $[zero] = [y_p]_1$  (zero found)  
 else  $[y] = [y_p]_1$ ; go to 3.
- 10 Send RESULT-SIGNAL to M; send  $[zero]$  to M  
 while (true) (Slave does not terminate).

At the end we have a list of intervals of width  $< \epsilon$  each containing a root.

Patrick (1972) gives a method of finding real roots which lends itself to parallel operations. This was discussed in Chapter 4, Section 12 of the present work, so will not be discussed further here, except to point out that the various roots of the derivatives can be found simultaneously on a parallel processor.

### 5.9 Hybrid Methods Involving Newton's Method

Nesdore (1970) describes a program which potentially uses 14 different methods, including Newton's. For the function being solved, the "computational efficiency" (C.E.) is used to rate and order all the methods in consideration (the C.E. is defined as  $p^{\frac{1}{\theta}}$  where  $p$  = order and  $\theta$  = work per iteration). A few iterations are executed with the most efficient method; if divergence is detected the iterations are repeated with a different starting point. If divergence still occurs the next most efficient method is tried, and so on until the list is exhausted. If convergence is detected but appears linear, a switch is made to the most efficient multiple zero method (there are 3 included in the program, such as Schroeder's method). Some tests show that a random choice of method is 35% slower than the selection method described above.

Bini and Pan (1998) give a method for computing the eigenvalues of a real symmetric tridiagonal (rst) matrix. This problem is related to solving a polynomial with only real roots, for given the coefficients of an  $n$ 'th degree polynomial  $p(z)$  having only real zeros  $\zeta_1, \zeta_2, \dots, \zeta_n$ , we may compute an  $n \times n$  rst matrix  $T_n$  that has characteristic polynomial  $p(z)$  and eigenvalues  $\zeta_1, \zeta_2, \dots, \zeta_n$  (see 5.348 below). Their method approximates the eigenvalues of  $T_n$  (with integer entries at most  $2^m$ ), within error bounds  $2^{-h}$ , at cost bounded by

$$O(n \log^2 n (\log^2 b + \log n)) \tag{5.315}$$

where  $b = m+h$ . The same bounds apply to finding the roots of  $p(z)$ .

$$\text{Let } \mathbf{T}_n = \begin{bmatrix} a_1 & b_1 & 0 & \dots & 0 \\ b_1 & a_2 & b_2 & \dots & 0 \\ 0 & b_2 & a_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & b_{n-1} & a_n \end{bmatrix} \tag{5.316}$$

where

$$|a_i|, |b_i| \leq 2^m \tag{5.317}$$

so that by Gershgorin's theorem

$$-3(2^m) \leq \zeta_i \leq 3(2^m) \tag{5.318}$$

We say that  $R = \{r_0, \dots, r_k\}$  **interleaves** the set  $Q = \{q_1, \dots, q_k\}$ , or that “ $R$  is an interlacing set for  $Q$ ” if

$$r_0 \leq q_1 \leq r_1 \leq q_2 \leq \dots q_{k-1} \leq r_{k-1} \leq q_k \leq r_k \tag{5.319}$$

(We allow  $r_0 = -\infty$  and  $r_k = +\infty$ ). We say  $s$  is a splitting point of the level  $(g,h)$  for the set  $Q$  if

$$q_g < s < q_h \tag{5.320}$$

Let  $Diag(\mathbf{B}_1, \dots, \mathbf{B}_s)$  denote the block diagonal matrix having the blocks  $\mathbf{B}_1, \dots, \mathbf{B}_s$ .

Cauchy's interlace theorem states:

**Theorem 5.9.1** If  $A_r$  is an  $r \times r$  submatrix of an  $n \times n$  real tridiagonal matrix  $\mathbf{A}$ , then the eigenvalues  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$  of  $\mathbf{A}$  and the eigenvalues  $\mu_1 \leq \mu_2 \leq \dots \leq \mu_r$  of  $\mathbf{A}_r$  obey

$$\lambda_i \leq \mu_i \leq \lambda_{i+n-r} \quad (i = 1, \dots, r) \tag{5.321}$$

Then the eigenvalues of  $\mathbf{T}_n$  satisfy:

**Theorem 5.9.2:** (a) If  $nk$  is a multiple of  $2(k+1)$  and if  $\{\mu_1 < \mu_2 < \dots < \mu_{\frac{k}{k+1}n}\}$  is the set of all the eigenvalues of the  $k \times k$  principal submatrices  $\tilde{\mathbf{T}}_i$  of  $\mathbf{T}_n$ , ( $i=1, \dots, \frac{n}{k+1}$ ) containing the entries  $a_s$  of  $\mathbf{T}_n$  for  $s = (i-1)(k+1)+j$  for  $j = 1, \dots, k$ . Then

$$\lambda_{\frac{nk}{2(k+1)}} \leq \mu_{\frac{nk}{2(k+1)}} \leq \lambda_{\frac{n(k+2)}{2(k+1)}} \tag{5.322}$$

(b) If  $1 \leq j \leq n - 2$  and if  $\{\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{n-1}\}$  is the set of all the eigenvalues of the following two principal submatrices of  $\mathbf{T}_n$ :

$$\mathbf{T}_j = \begin{bmatrix} a_1 & b_1 & 0 & \dots & 0 \\ b_1 & a_2 & b_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & a_{j-1} & b_{j-1} \\ 0 & \dots & 0 & b_{j-1} & a_j \end{bmatrix},$$

$$\tilde{\mathbf{T}}_{n-j-1} = \begin{bmatrix} a_{j+2} & b_{j+2} & 0 & \dots & 0 \\ b_{j+2} & a_{j+3} & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & a_{n-1} & b_{n-1} \\ \dots & \dots & 0 & b_{n-1} & a_n \end{bmatrix} \tag{5.323}$$

then

$$\lambda_i \leq \gamma_i \leq \lambda_{i+1} \quad (i = 1, \dots, n - 1) \tag{5.324}$$

Next **Corollary 5.9.2.1** states: if  $n$  is a multiple of 4,  $\{\sigma_1 \leq \sigma_2 \leq \dots \leq \sigma_{\frac{n}{2}}\} = \{a_{2i-1}, i = 1, \dots, \frac{n}{2}\}$ , and  $\{\theta_1 \leq \theta_2 \leq \dots \leq \theta_{\frac{n}{2}}\} = \{a_{2i}, i = 1, \dots, \frac{n}{2}\}$ , then

$$\lambda_{\frac{n}{4}} \leq \sigma_{\frac{n}{4}} \leq \lambda_{\frac{3}{4}n}; \lambda_{\frac{n}{4}} \leq \theta_{\frac{n}{4}} \leq \lambda_{\frac{3}{4}n} \tag{5.325}$$

while **Theorem 5.9.3** states: Let  $\{\phi_1 \leq \phi_2 \leq \dots \leq \phi_n\}$  be the set of all eigenvalues of

$$\mathbf{S}_k = \mathbf{T}_k - \text{Diag}(0, 0, \dots, 0, b_k) \tag{5.326}$$

$$\mathbf{R}_{n-k} = \tilde{\mathbf{T}}_{n-k} - \text{Diag}(b_k, 0, \dots, 0) \tag{5.327}$$

where  $\mathbf{T}_k$  and  $\tilde{\mathbf{T}}_{n-k}$  are defined in Theorem 5.9.2b. Set  $\phi_{n+1} = \phi_n + 2b_k$ ,  $\phi_0 = \phi_1 + 2b_k$

If  $b_k > 0$ , then

$$\phi_i \leq \lambda_i \leq \phi_{i+1} \quad (i = 1, \dots, n) \tag{5.328}$$

while if  $b_k < 0$ , then

$$\phi_{i-1} \leq \lambda_i \leq \phi_i \quad (i = 1, \dots, n) \tag{5.329}$$

Next the authors show how to compute the number of eigenvalues in the intervals of nearly interlacing sets. At the first stage we approximate some eigenvalues of  $\mathbf{T}_n$  within a required error bound and cover each remaining eigenvalue by an interval containing no other eigenvalues. Let the set  $\{d_0, \dots, d_n\}$  interleave the set  $\Lambda$  of eigenvalues of  $\mathbf{T}_n$  (we will show later how to find the  $d_i$  or approximations to them). Thus

$$d_0 < \lambda_1 < d_1 \leq \dots \leq d_{n-1} \leq \lambda_n < d_n \tag{5.330}$$

Let  $d_i^-, d_i^+$  be approximations to  $d_i$ , i.e. for a fixed  $\Delta$ :

$$d_i^+ = d_i^- + 2\Delta, \quad d_i^- \leq d_i \leq d_i^+ \quad (i = 0, \dots, n) \tag{5.331}$$

Suppose we know how to determine

$$p(\lambda) = \det(\mathbf{T}_n - \lambda \mathbf{I}) \tag{5.332}$$

Then for every  $\lambda_j$  we will either compute its approximation within the error  $2\Delta$ , or determine that the interval

$$K_j \equiv \{\lambda : d_{j-1}^+ \leq \lambda \leq d_j^-\} \tag{5.333}$$

contains  $\lambda_j$  and no other eigenvalue. This is done by Bini and Pan's Algorithm 3.1 which inputs  $\Delta, n, D = \{d_i^-, d_i^+, i = 1, \dots, n-1\} \cup \{d_0^+ = -3(2^m), d_n^- = 3(2^m)\}$  such that 5.330 and 5.331 and

$$p(d_i^-)p(d_i^+) \neq 0 \quad (i = 0, \dots, n) \tag{5.334}$$

hold. It also inputs  $d_0^+$  and  $d_n^-$  where

$$d_0^+ \leq \lambda_1, \lambda_n \leq d_n^- \tag{5.335}$$

and outputs  $K_j$  or an approximation  $\bar{\lambda}_j$  such that

$$|\bar{\lambda}_j - \lambda_j| < 2\Delta \tag{5.336}$$

For details of how the algorithm works see the cited paper. Their Algorithm 3.1 is complemented by their Algorithm 4.1, which uses Newton iteration and the bisection method. It is based on the following theorem, proved by Renegar (1987): Let  $x^{(0)}$  be such that

$$|x^{(0)} - \zeta_1| \leq |x^{(0)} - \zeta_2| \leq \dots \leq |x^{(0)} - \zeta_n| \tag{5.337}$$

If

$$|x^{(0)} - \zeta_1| < \frac{1}{5n^2}|x^{(0)} - \zeta_2| \tag{5.338}$$

then the Newton iteration converges to  $\zeta_1$  so that

$$|x^{(i)} - \zeta_1| < 2^{3-2^i}|x^{(0)} - \zeta_1| \tag{5.339}$$

In Bini and Pan's "Algorithm 4.1", starting with  $c < \lambda < d$ , we apply  $\log(5n^2)$  bisection steps until we obtain  $x^{(0)}$  satisfying 5.338 with  $c_0 < x^{(0)} < d_0$ . Note that the bisection steps are preceded by a more complicated process-see the cited paper. After the bisection steps we apply  $\log\log(.8\frac{d_0-c_0}{\Delta n^2})$  Newton steps to find  $\bar{\lambda}$  such that  $|\bar{\lambda} - \lambda| < \Delta$ .

We will apply Algorithm 4.1 concurrently to all the intervals which contain a single eigenvalue. We will thus need to compute  $p(\lambda)$  and  $p'(\lambda)$  at a set of up to  $n$  points. We will use the following recurrence relation for

$$p_i(\lambda) = \det \left( \begin{bmatrix} a_1 & b_1 & 0 & \dots \\ b_1 & a_2 & b_2 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & b_{i-1} & a_i \end{bmatrix} - \lambda \mathbf{I} \right) \tag{5.340}$$

namely

$$p_0(\lambda) = 1, p_1(\lambda) = a_1 - \lambda \tag{5.341}$$

$$p_{i+1}(\lambda) = (a_{i+1} - \lambda)p_i(\lambda) - b_i^2 p_{i-1}(\lambda) \quad (i = 1, 2, \dots, n - 1) \tag{5.342}$$

or equivalently

$$\begin{bmatrix} p_{i+1}(\lambda) \\ p_i(\lambda) \end{bmatrix} = \begin{bmatrix} a_{i+1} - \lambda & -b_i^2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p_i(\lambda) \\ p_{i-1}(\lambda) \end{bmatrix} \tag{5.343}$$

$$= F_i F_{i-1} \dots F_0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{5.344}$$

where

$$F_j = \begin{bmatrix} a_{j+1} - \lambda & -b_j^2 \\ 1 & 0 \end{bmatrix} \quad (j = 0, 1, \dots, i), \quad b_0 = 0 \tag{5.345}$$

The above leads us to Bini and Pan's Algorithm 5.1 which inputs  $n = 2^h$  (we pad our matrix with zeros if necessary to bring  $n$  up to such a value),  $a_1, \dots, a_n, b_1, \dots, b_{n-1}$  and outputs the coefficients of  $p(\lambda) = \det(\mathbf{T}_n - \lambda \mathbf{I})$ . It works by initially setting  $H_j^{(0)} = F_j$  ( $j = 0, \dots, n - 1$ ) and then, for  $i = 1, 2, \dots, \log n$  compute

$$H_j^{(i)} = H_{2j+1}^{(i-1)} H_{2j}^{(i-1)} \quad (j = 0, \dots, 2^{-i}n - 1) \tag{5.346}$$

Then

$$p(\lambda) = [10] H_0^{(h)} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{5.347}$$

Given the coefficients of  $p(\lambda)$  (and  $p'(\lambda)$ ) we may compute the values of  $p(\lambda)$  and  $p'(\lambda)$  at a set of  $n$  points at a parallel cost of  $O(\log^2 n \log \log n)$  using  $\frac{n}{\log \log n}$  processors (see Aho, Hopcroft and Ullman (1976)).

The main Algorithm (6.1) recursively reduces the original problem to two problems of half-size. It inputs integers  $m, n, u, a_1, \dots, a_n, b_1, \dots, b_{n-1}$  where  $n$  is a power of 2;  $u$  such that the output errors are  $< 2^{-u}$ ; and  $m$  such that  $|a_i|, |b_i| \leq 2^m$ . It outputs  $\gamma_1, \dots, \gamma_n$  such that  $|\lambda_i - \gamma_i| < 2^{-u}$  where the  $\lambda_i$  are the eigenvalues of  $\mathbf{T}_n$ . It works as follows: we compute the coefficients of  $p(\lambda) = \det(\mathbf{T}_n - \lambda \mathbf{I})$  by using Algorithm 5.1; then we apply Algorithm 6.1 to the set

$$m, \frac{n}{2}, u + 1, a_1, \dots, a_{\frac{n}{2}-1}, a_{\frac{n}{2}} - b_{\frac{n}{2}}, b_1, \dots, b_{\frac{n}{2}-1}$$

which defines an  $\text{rst}$  matrix  $\mathbf{S}_{\frac{n}{2}}$  and the set

$$m, \frac{n}{2}, u + 1, a_{\frac{n}{2}+1} - b_{\frac{n}{2}}, a_{\frac{n}{2}+2}, \dots, a_n, b_{\frac{n}{2}}, \dots, b_n$$

which defines a matrix  $\mathbf{R}_{\frac{n}{2}}$ , thus obtaining approximations  $\delta_1 \leq \delta_2 \leq \dots \leq \delta_n$  to the eigenvalues of  $\mathbf{S}_{\frac{n}{2}}$  and  $\mathbf{R}_{\frac{n}{2}}$  within the absolute error  $\Delta = 2^{-u-1}$  (see Theorem 5.9.3). Now recall that the set of all eigenvalues of  $\mathbf{S}_{\frac{n}{2}}$  and  $\mathbf{R}_{\frac{n}{2}}$  interleaves the set  $\{\lambda_i\}$  (see Theorem 5.9.3, i.e. 5.328 and 5.329). Set  $d_i^+ = \delta_i + \Delta, d_i^- = \delta_i - \Delta$  and apply Algorithms 3.1 and 4.1 to obtain  $\gamma_i, \dots, \gamma_n$  such that  $|\gamma_i - \lambda_i| < 2^{-u}$ .

Suppose  $p(x)$  is a polynomial of degree  $n$  with coefficients in the range  $-2^m$  to  $+2^m$ . We set  $p_n(x) \equiv p(x), p_{n-1} = -p'(x)$  and apply the extended Euclidean algorithm to  $p_n$  and  $p_{n-1}$ . i.e.

$$p_{i+1}(x) = q(x)p_i(x) - rp_{i-1}(x) \quad (i = n - 1, \dots, 1) \tag{5.348}$$



Then we may obtain the entries  $a_{i+1}$  and  $b_i$  of  $\mathbf{T}_n$  according to 5.342 by setting  $q(x) = a_{i+1} - x$ ,  $b_i = \sqrt{r}$  (the sign of  $b_i$  is arbitrary, but if the roots of  $p(x)$  are all real  $r \geq 0$ ). Note: it may be thought that Algorithm 5.1 is not necessary, since we start knowing the coefficients of  $p(x)$ , but in fact “ $p(x)$ ” most often refers to the characteristic equation of  $\mathbf{T}_n$  of size  $\frac{n}{2}$ ,  $\frac{n}{4}$  etc.

Locher and Skrzipek (1995) describe a globally convergent method for calculating zeros (including complex ones) of a real polynomial using Chebyshev polynomials. Let the disk  $K_r(0) = \{r \exp(it) : 0 \leq t \leq 2\pi\}$ . Consider  $p_n = \sum_{i=0}^n c_i z^i$  on  $\Gamma \equiv K_1(0)$ . We can write

$$p_n(\exp(it)) = u_n(t) + iv_n(t) \tag{5.349}$$

where

$$u_n(t) = \sum_{\nu=0}^n c_\nu \cos(\nu t), \quad v_n(t) = \sum_{\nu=1}^n c_\nu \sin(\nu t) \quad (0 \leq t \leq 2\pi) \tag{5.350}$$

$$\text{Let } \bar{v}_{n-1}(t) = \frac{v_n(t)}{\sin(t)} = \sum_{\nu=0}^{n-1} c_{\nu+1} \frac{\sin(\nu+1)t}{\sin(t)}.$$

A zero of  $p_n$  on  $\Gamma$  is a common zero of  $u_n$  and  $v_n$ , i.e. each zero of  $p_n$  on  $\Gamma/\{\pm 1\}$  is a common zero of  $u_n$  and  $\bar{v}_{n-1}$ . Since  $p_n(\Gamma)$  is symmetric about the x-axis, we only need to consider  $0 \leq t \leq \pi$ , or with  $x = \cos(t)$ ,  $x \in [-1, 1]$ . Then

$$p_n(\exp(it)) = u_n(t) + i \sin(t) \bar{v}_{n-1}(t) \tag{5.351}$$

$$= \sum_{\nu=0}^n c_\nu T_\nu(x) + i \sqrt{1-x^2} \sum_{\nu=0}^{n-1} c_{\nu+1} U_\nu(x) \tag{5.352}$$

$$\equiv q_n(x) + i \sqrt{1-x^2} q_{n-1}(x) \tag{5.353}$$

where  $T_\nu$  and  $U_\nu$  are Chebyshev polynomials of the first and second kind, i.e.

$$T_\nu(x) = \cos(\nu \cos^{-1} x) \tag{5.354}$$

$$U_\nu(x) = \frac{\sin([\nu+1] \cos^{-1} x)}{\sqrt{1-x^2}} \tag{5.355}$$

Hence zeros of  $p_n$  on  $\Gamma/\{\pm 1\}$  coincide with the common zeros of  $q_n$  and  $q_{n-1}$  in  $[-1, 1]$ .

We may apply the Euclidean algorithm starting with  $q_n$  and  $q_{n-1}$  to get the  $\text{gcd}(q_n, q_{n-1}) = q_s$  whose zeros are the common zeros of  $q_n$  and  $q_{n-1}$ . Then we get the authors’

Theorem 2.1: (1)  $p_n$  has zeros of modulus 1 iff either

- a)  $p_n(1) = 0$  or  $p_n(-1) = 0$  or
- b)  $q_n, q_{n-1}$  have a gcd  $q_s$  with some zeros in  $[-1, 1]$

(or both a and b).

(2) If  $q_s \neq 0$  in  $[-1,1]$  then  $q_n, q_{n-1}$  generate the Sturm sequence

$\{q_n, q_{n-1}, \dots, q_s\}$  on  $[-1,1]$ , and the number  $N_\Gamma(p)$  of zeros of  $p_n$  in the interior of  $\Gamma$

$$= SC(q_n(-1), \dots, q_s(-1)) - SC(q_n(1), \dots, q_s(1)) \tag{5.356}$$

where  $SC(a, b, \dots, z)$  means "number of sign changes in the sequence  $a, b, \dots, z$ ". For proof of (2) above see Locher (1993).

We may detect and divide out (by Horner's rule-equation 1.2) zeros of  $p_n$  at  $\pm 1$ , so in the following we assume  $p_n(\pm 1) \neq 0 \neq q_s(\pm 1)$ . If  $q_s$  has zeros  $x_j$  in  $[-1,1]$  then  $\exp(\pm it_j)$  with  $t_j = \cos(x_j)$  are the zeros of  $p_n$  on  $\Gamma$ . Dividing each quadratic factor out by 1.14 we obtain a new polynomial  $p_k$  with no zeros on  $\Gamma$  and whose zeros coincide with the remaining zeros of  $p_n$  outside  $\Gamma$ . By theorem 2.1 part 2 we get the number of zeros of  $p_k$  (and hence  $p_n$ ) in the interior of  $\Gamma$ .

We can get the zeros on and in the interior of  $\Gamma$  using the Chebyshev representation of  $q_n, q_{n-1}$  etc rather than expressing them in powers of  $x$ . We use the relations

$$T_0(x) = U_0(x) \tag{5.357}$$

$$T_1(x) = xU_0(x) \tag{5.358}$$

$$T_\nu(x) = xU_{\nu-1}(x) - U_{\nu-2}(x) \ (\nu = 2, 3, \dots) \tag{5.359}$$

and

$$2T_m U_n = \begin{cases} U_{n+m} + U_{n-m} & (n \geq m) \\ U_{2m-1} & (n = m - 1) \\ U_{m+n} - U_{m-n-2} & (n \leq m - 2) \end{cases} \tag{5.360}$$

Setting  $q_n^{[0]} = q_n, q_{n-1}^{[1]} = q_{n-1}$  we get in the first step of the Euclidean algorithm:

$$\begin{aligned} q_n^{[0]}(x) &= c_0 T_0(x) + c_1 T_1(x) + \sum_{\nu=2}^n c_\nu T_\nu(x) \\ &= c_0 U_0(x) + c_1 x U_0(x) + \sum_{\nu=2}^n c_\nu (x U_{\nu-1}(x) - U_{\nu-2}(x)) \\ &= T_1(x) \sum_{\nu=0}^{n-1} c_{\nu+1} U_\nu(x) - [(c_2 - c_0) U_0(x) + \sum_{\nu=1}^{n-2} c_{\nu+2} U_\nu(x)] \\ &\equiv h_1^{[1]}(x) q_{n-1}^{[1]}(x) - q_{n-2}^{[2]}(x) \end{aligned} \tag{5.361}$$

where in general

$$q_\mu^{[j]} = \sum_{\nu=0}^{\mu} c_\nu^{[j]} U_\nu \quad (j \geq 0), \quad h_k^{[j]} = \sum_{\nu=0}^k b_\nu^{[j]} T_\nu \quad (j \geq 1) \quad (5.362)$$

In particular

$$h_1^{[1]} = b_0^{[1]} + b_1^{[1]} T_1, \quad \text{i.e. } b_0^{[1]} = 0, \quad b_1^{[1]} = 1 \quad (5.363)$$

$$c_\nu^{[0]} = c_\nu \quad (\nu = 0, \dots, n), \quad c_\nu^{[1]} = c_{\nu+1} \quad (\nu = 0, \dots, n-1) \quad (5.364)$$

$$c_0^{[2]} = c_2 - c_0, \quad c_\nu^{[2]} = c_{\nu+2} \quad (\nu = 1, \dots, n-2) \quad (5.365)$$

In later steps of the Euclidean algorithm we have to compute  $h_k^{[j]}$  and  $q_{m-k-1}^{[j+1]}$  from

$$q_m^{[j-1]} = h_k^{[j]} q_{m-k}^{[j]} - q_{m-k-1}^{[j+1]} \quad (2 \leq j \leq s-1) \quad (5.366)$$

Here the subscript refers to the degree of the polynomial and the superscript to the order of creation. Usually  $h_k^{[j]}$  is linear, i.e.  $k = 1$ , so  $m = n - j + 1$ . In that case, comparing coefficients of  $U_\nu$  and using the first of relations 5.360 with  $m = 1$  and  $n = \nu$  i.e.

$$2T_1 U_\nu = U_{\nu+1} + U_{\nu-1} \quad (5.367)$$

we get

$$b_1^{[j]} = 2 \frac{c_m^{[j-1]}}{c_{m-1}^{[j]}}, \quad b_0^{[j]} = \frac{c_{m-1}^{[j-1]} - \frac{1}{2} b_1^{[j]} c_{m-2}^{[j]}}{c_{m-1}^{[j]}} \quad (5.368)$$

$$c_0^{[j+1]} = \frac{1}{2} b_1^{[j]} c_1^{[j]} + b_0^{[j]} c_0^{[j]} - c_0^{[j-1]} \quad (5.369)$$

$$c_\nu^{[j+1]} = \frac{1}{2} b_1^{[j]} c_{\nu+1}^{[j]} + b_0^{[j]} c_\nu^{[j]} + \frac{1}{2} b_1^{[j]} c_{\nu-1}^{[j]} - c_\nu^{[j-1]} \quad (\nu = 1, 2, \dots, m-2) \quad (5.370)$$

For the more general case where  $k > 1$  at some step, see the cited paper.

To get the number of zeros in the interior of  $\Gamma$  we may proceed as follows: if we have found all, say  $\omega$ , zeros of  $p_n(\exp(it))$  (that is of  $q_s(x)$ ) for  $t \in (0, \pi)$  (and we see in the next few paragraphs how to do that), we get the zeros  $z_\nu = \exp(it_\nu)$ ,  $\nu = 0, \dots, \omega - 1 \leq s - 1$  of  $p_n$  on  $\Gamma$ . Now  $\bar{z}_\nu$  is also a zero of  $p_n$ , so we can divide  $p_n$  by the polynomial

$$\Pi_{\nu=0}^{\omega-1} (z - z_\nu)(z - \bar{z}_\nu) \quad (5.371)$$

using equation 14 of Chapter 1  $\omega$  times. This gives a polynomial with no zeros on  $\Gamma$  and applying a Sturm sequence using the above algorithm and Theorem 2.1 (2)

we may find the number of zeros of  $p_n$  in the interior of  $\Gamma$ .

To check whether  $q_s$  (and hence  $q_n, q_{n-1}$ ) has zeros in  $[-1,1]$  or not we generate a Sturm sequence starting with  $q_s$  and  $q'_s$ .

Using

$$U_\nu = \left\{ \begin{array}{l} 2(T_\nu + T_{\nu-2} + \dots + \frac{1}{2}T_0) \quad , \nu \text{ even} \\ 2(T_\nu + T_{\nu-2} + \dots + T_1) \quad , \nu \text{ odd} \end{array} \right\} \quad (5.372)$$

and  $T'_\nu = \nu U_{\nu-1}$

we see that with

$$q_s = \sum_{\nu=0}^s \alpha_\nu U_\nu, \quad (5.373)$$

$$q'_s = \sum_{\nu=0}^{s-1} \alpha'_\nu U_\nu \quad (5.374)$$

we have

$$\alpha'_{s-2\nu} = 2(s - 2\nu + 1)[\alpha_{s-1} + \alpha_{s-3} + \dots + \alpha_{s-2\nu+1}] \quad (\nu = 1, 2, \dots, [\frac{s}{2}]) \quad (5.375)$$

$$\alpha'_{s-2\nu-1} = 2(s - 2\nu)[\alpha_s + \alpha_{s-2} + \dots + \alpha_{s-2\nu}] \quad (\nu = 0, 1, \dots, [\frac{s}{2}]) \quad (5.376)$$

We can use here essentially the same equations as 5.368- 5.370.

Now suppose  $q_{s-k} = \gcd(q_s, q_{s-1}) \neq 0$  in  $[-1,1]$ , then  $\{q_s, q_{s-1}, \dots, q_{s-k}\}$  is a Sturm sequence. Since  $\pm 1$  are not zeros of  $q_s$  we have

$$N_{(-1,1)}(q_s) = SC(q_s(-1), \dots, q_{s-k}(-1)) - SC(q_s(1), \dots, q_{s-k}(1)) \quad (5.377)$$

Thus for  $q_{s-k} \neq \text{const}$  we must decide whether or not any zeros of  $q_{s-k}$  lie in  $[-1,1]$ . We may calculate  $h_k = q_s/q_{s-k}$  and  $h'_k$ , and generate the Sturm sequence starting with  $h_k$  and  $h'_k$ . This gives the number  $\omega$  of zeros of  $h_k$  in  $[-1,1]$ , and hence the number of **distinct** zeros of  $q_s$ , and hence of  $p_n$  on  $|z| = 1, 0 < \arg z < \pi$ . The authors show how to express  $h_k$  and  $h'_k$  as series in  $U_\nu$ , and then we may apply equations 5.368- 5.370 again. Now it has been proved by Locher and Skrzipek (private communication) that the roots of  $q_s$  and hence of  $h_k$  are always real and in  $[-1,1]$ . So, if we start Newton's iterations at 1, it will always converge to a root  $x_0$  in  $[-1,1]$ . We may deflate, by

$$h_{k-1}(x) = \frac{h_k(x)}{x - x_0} = \frac{h_k(x)}{\frac{1}{2}U_1(x) - x_0U_0(x)} \quad (5.378)$$

and similarly find the next lower root  $x_1$  of  $h_{k-1}(x)$  and so on until all  $\omega$  roots are found. If  $s = \omega$ , all zeros of  $q_s$  (and hence of  $p_n$  on  $\Gamma$ ) have been calculated;

otherwise  $q_s$  has some zeros of multiplicity greater than one in  $[-1,1]$ . To calculate the multiplicity  $\mu_j$  of  $x_j$  we divide  $q_s$  by  $(x - x_j)$  until we get a non-zero remainder term; then the multiplicity = number of times we can divide with a zero remainder. Thus

$$z_j = \exp(i \cos^{-1} x_j) \quad (j = 0, \dots, \omega - 1) \tag{5.379}$$

are the zeros of  $p_n$  on  $\Gamma$  (some may be multiple). If we now replace  $p_n(z)$  by

$$p_n^*(z) = \frac{p_n(z)}{\prod_{j=0}^{\omega-1} [(z - z_j)(z - \bar{z}_j)]^{\mu_j}} \tag{5.380}$$

and repeat the algorithm we will get a new  $q_s$  which has no zeros in  $[-1,1]$ . Then Theorem 2.1 (2) will give  $N_\Gamma(p_n)$  = number in interior of  $\Gamma$ ; while number outside =  $n - N_\Gamma(p_n) - \omega$ .

So far we have only shown how to find zeros on or in the unit circle. But if we transform  $p_n(x)$  to  $p_n^*(x) = p_n(rx)$  with coefficients  $c_\nu r^\nu$  we can apply Theorem 2.1 to get either the zeros on  $K_r(0)$  or the number of zeros in the interior of this circle (indeed by applying 5.380 we may get both). We will use a bisection strategy to get the moduli of all zeros  $r_i$  which have distinct moduli. Once this is known we may fix the arguments by finding the real zeros of the gcd of  $q_{n,r_i} = q_n(r_i x)$  and  $q_{n-1,r_i} = q_{n-1}(r_i x)$ . To avoid too much rounding we use the scaled polynomials to get rough values for the zeros, then refine them using the original polynomial and Newton or Bairstow's method.

To start the bisection process we use Gershgorin's theorem to get upper and lower bounds  $r_u$  and  $r_l \neq 0$  for the absolute values of the zeros. Then we use  $r_u$  as initial value to calculate the positive zero  $\tilde{r}_u$  of

$$|c_n|z^n - \sum_{\nu=0}^{n-1} |c_\nu|z^\nu \tag{5.381}$$

and similarly use  $r_l$  to get the positive zero  $\tilde{r}_l$  of

$$\sum_{\nu=1}^n |c_\nu|z^\nu - |c_0| \tag{5.382}$$

Take

$$r^{(0)} = \min(r_u, \tilde{r}_u), \quad r^{(1)} = \max(r_l, \tilde{r}_l) \tag{5.383}$$

Next we test whether any zeros of  $p_n$  lie in  $K_{r^{(0)}}(0)$  or  $K_{r^{(1)}}(0)$ . If so we find their arguments and multiplicities as before, as well as the number inside and outside each of them. If all zeros lie on these circles we are finished. Otherwise we can use algorithm bisect(2) where bisect is defined recursively as follows:

PROCEDURE bisect( $\nu$ )  
 If not all zeros found do:  
 BEGIN

$$r^{(\nu)} = \frac{r^{(\nu-2)} + r^{(\nu-1)}}{2} \tag{5.384}$$

Find the zeros on, inside and outside  $K_{r^{(\nu)}}$ . If there are  $j_1 > 0$  zeros in the annulus between  $K_{r^{(\nu-1)}}(0)$  and  $K_{r^{(\nu)}}(0)$  then

BEGIN if  $|r^{(\nu-1)} - r^{(\nu)}| > \epsilon$  (moderately small) set  $r^{(\nu-2)} = r^{(\nu)}$

ELSE determine the radii of the circles where the remaining zeros lie by Newton or Bairstow (see later).

BISECT( $\nu + 1$ );

END

if there are  $j_2 > 0$  zeros between  $K_{r^{(\nu)}}(0)$  and  $K_{r^{(\nu-2)}}(0)$  then

BEGIN if  $|r^{(\nu-2)} - r^{(\nu)}| > \epsilon$  then

$$r^{(\nu-1)} = r^{(\nu)}$$

ELSE determine moduli of remaining zeros by Newton or Bairstow's method

BISECT( $\nu + 1$ );

END

END

The transformation  $p_n^*(x) = p_n(rx)$  may cause problems if  $n$  is large and  $r$  is very small or very large, for then we may get under- or overflow. It seems that we avoid most of these problems if we use in such cases

$$r^{-\frac{n}{2}} p_n(rx) = \sum_{\nu=0}^n c_\nu r^{\nu-\frac{n}{2}} x^\nu \tag{5.385}$$

(This author suspects that the above device only works for moderate sized  $n$ ).

The above bisection method may be used until we have  $|r^{(\nu-1)} - r^{(\nu)}| < \epsilon$  where  $\epsilon$  is moderately small. Then it is more efficient to use Newton's or Bairstow's method to improve the accuracy as desired. Assume that there is at least one zero of  $p_n$  with modulus  $r \in (r^{(\nu-1)}, r^{(\nu)})$ . The number in this annulus is known since in the above bisection process we have found the numbers  $n_{i_\nu}, n_{i_{\nu-1}}$  in the interior and  $n_{b_\nu}, n_{b_{\nu-1}}$  on the boundary of  $K_{r^{(\nu)}}(0)$  and  $K_{r^{(\nu-1)}}(0)$ . Thus between these circles we have  $n_{i_\nu} - (n_{i_{\nu-1}} + n_{b_{\nu-1}})$  zeros. If there are zeros on the circles we remove them by the process leading to equation 5.380. So we can assume that there are no roots on those circles. The Euclidean algorithm yields the sequence

$$S_{r^{(\nu-1)}} = \{q_{n,r^{(\nu-1)}}^{[0]}, q_{n-1,r^{(\nu-1)}}^{[1]}, \dots\} \tag{5.386}$$

and if we knew  $r$  we could get

$$S_r = \{q_{n,r}^{[0]}, q_{n-1,r}^{[1]}, \dots\} \tag{5.387}$$

For smallish  $\epsilon$  we can approximate  $S_r$  by  $S_{r(\nu-1)}$ , and thus we can approximate the gcd  $q_{n-k,r}^{[j]}$  by  $q_{n-k,r(\nu-1)}^{[j]}$  where the remaining elements of  $S_{r(\nu-1)}$  are close to the zero polynomial. We may find the number of zeros of  $q$  in  $[-1,1]$  by our usual method and can test whether it has zeros at  $\pm 1$ . If there are zeros in  $[-1,1]$  we compute  $x_s$  and  $x_l$  as the smallest and largest in this range by Newton's method, evaluating  $q$  and  $q'$  by Clenshaw's algorithm. By  $x \rightarrow \cos^{-1}x$  we get approximations to the arguments of the zeros of  $p_n$  on  $K_r(0)$  which have the smallest imaginary part. Then we can improve the accuracy of one of these by Newton's or Bairstow's method to give a zero  $\text{rexp}(i\theta)$ . We may then find all the zeros on  $K_r(0)$  by our usual method; next we test whether there are zeros between  $K_{r(\nu-1)}(0)$  and  $K_r(0)$  or between  $K_r(0)$  and  $K_{r(\nu)}(0)$ . If there are we repeat the process until all zeros between  $K_{r(\nu-1)}(0)$  and  $K_{r(\nu)}(0)$  are found. To avoid rounding error effects when applying Sturm's theorem near a zero of  $p_n$  we should use multiple precision.

Tests on a large number of polynomials up to degree about 15 were very successful.

Lang and Frenzel (1994) describe a program which uses Muller's method (see Chapter 7) to compute an estimate of a root of the deflated polynomial that contains all the roots of  $p(x)$  except those already found. Only a few iterations of Muller's method are performed. Note that Muller's method can find complex roots even when initialized with real values. This is in contrast with (for example) Newton's method. In a second step the estimate from Muller's method is used as the initial value for Newton's method, working with the **original** (not deflated) polynomial. This avoids rounding errors introduced by deflation. Their method was compared with the Jenkins-Traub method and with the eigenvalue method used in MATLAB. It gave smaller errors than either, and indeed for degrees above 40 Jenkins-Traub did not work at all. The method described here was faster than the eigenvalue method for all degrees (at degrees above 40 comparison with Jenkins-Traub is meaningless). The Lang-Frenzel program gave near computer accuracy up to degree 10,000. However, this author would point out that some eigenvalue methods developed in the 21st century may be considerably faster than the MATLAB version referred to in Lang and Frenzel's paper.

Tsai and Farouki (2001) describe a collection of C++ functions for operations on polynomials expressed in Bernstein form. This form is used extensively in computer-aided geometric design, as it is much less sensitive to coefficient perturbations than other bases such as the power form. The Bernstein form on  $[-1,1]$  consists in

$$p(x) = \sum_{k=0}^n c_k^n b_k^n(x), \quad b_k^n(x) = \binom{n}{k} (1-x)^{n-k} x^k \quad (5.388)$$

where  $c_k^n$  is a coefficient of the polynomial, while  $\binom{n}{k}$  is the binomial coefficient

$$\frac{n!}{k!(n-k)!} \quad (5.389)$$

Explicit conversions between different bases are ill-conditioned for high degree  $n$ , so it is essential to remain with the Bernstein form from start to end of a calculation. This is not more difficult than the usual power-form. A Bernstein-form polynomial may be evaluated by a sequence of linear interpolations among the coefficients, giving a triangular array  $P_k^{(r)}$  ( $r = 0, 1, \dots, n; k = r, \dots, n$ ). Initially

$$P_k^{(0)} = c_k^n \quad (k = 0, \dots, n) \quad (5.390)$$

and then if the polynomial is to be evaluated at  $x_s$ , we apply, for  $r = 1, \dots, n$ :

$$P_k^{(r)} = (1 - x_s)P_{k-1}^{(r-1)} + x_s P_k^{(r-1)} \quad (k = r, \dots, n) \quad (5.391)$$

Finally

$$p(x_s) = P_n^{(n)} \quad (5.392)$$

In addition the values

$$P_0^{(0)}, P_1^{(1)}, \dots, P_n^{(n)} \text{ and } P_n^{(n)}, P_n^{(n-1)}, \dots, P_n^{(0)} \quad (5.393)$$

are the coefficients on  $[0, x_s]$  and  $[x_s, 1]$  respectively. The authors, in their theorem 2, give conditions for  $p(x)$  to have a unique root in  $[a, b]$ , and for the Newton iteration to converge to it from any point in that interval. Their root-finder employs recursive binary subdivision of  $[0, 1]$  to identify intervals in which the conditions of their theorem 2 are satisfied. If  $p(x)$  has only simple roots on  $[0, 1]$ , this process terminates with up to  $n$  intervals on which Theorem 2 holds and Newton may be applied to give guaranteed quadratic convergence. However at a high level of subdivision, because of rounding errors, we may eventually get erroneous coefficient signs and the method breaks down. It appears that for Chebyshev polynomials this breakdown occurs at degree 50. If the conditions of Theorem 2 are not met on a subinterval smaller than a certain tolerance the process is terminated with an error message. Multiple roots may be determined by the method described in section 4, Chapter 2 of this work; thus we obtain polynomials  $Q_k(x)$  (having only simple roots) to which the above bisection-Newton method may be applied. It is reported that on a set of Chebyshev polynomials the software performs "remarkably" well up to degree 20 (much better than using the power form).

Ellenberger (1960) describes a method in which Bairstow's method is first tried and if that does not work Newton is tried. An Algol program is given in Ellenberger (1961)—see also Alexander (1961) and Cohen (1962).



Hopgood and McKee (1974) describe an i-point iteration method based on Hermite interpolation with a function and derivative evaluation at each point, namely:

$$x_{i+1} = \sum_{j=0}^i h_j(0)x_j + \sum_{j=0}^i \bar{h}_j(0) \left[ \frac{1}{f'(x_j)} \right] \quad (5.394)$$

where

$$h_j(y) = [1 - 2(y - y_j)l'_j(y)]l_j^2(y) \quad (5.395)$$

and

$$\bar{h}_j(y) = (y - y_j)l_j^2(y) \quad (5.396)$$

$$l_j(y) = \frac{\prod_{k=0, \neq j}^i (y - y_k)}{\prod_{k=0, \neq j}^i (y_j - y_k)} \quad (5.397)$$

$$y = f(x), y_k = f(x_k) \quad (k = 0, 1, \dots, i) \quad (5.398)$$

The authors give an algorithm, which they call the m-cycle improved Newton method (*INM*)<sup>m</sup> as follows:

- (1)  $i=0$ , given an initial guess  $x_0$
- (2) Evaluate  $f(x_0)$  and  $f'(x_0)$
- (3) Compute  $x_{i+1} = \sum_{j=0}^i h_j(0)x_j + \sum_{j=0}^i \bar{h}_j(0) \left[ \frac{1}{f'(x_j)} \right]$
- (4) If  $|x_{i+1} - x_i| < \epsilon$  then STOP else GO TO 5
- (5) Evaluate  $f(x_{i+1})$  and  $f'(x_{i+1})$
- (6) If  $i = m-1$  then  $i = 0$ ,  $x_0 = x_m$ ,  $f(x_0) = f(x_m)$ ,  $f'(x_0) = f'(x_m)$  else  $i = i+1$
- (7) GO TO 3

Another variation starts with 2 initial guesses; i.e. we alter step (1) to:

- (1)  $i = 1$  given initial guesses  $x_0, x_1$

This will be called “(*INM*)<sup>m-1</sup> with 2 starting values”. It is shown that the convergence order of (*INM*)<sup>m</sup> and “(*INM*)<sup>m-1</sup> with 2 starting values” are respectively

$$2 \times 3^{m-1} \text{ and } 3^{m-2} + \sqrt{32^{m-4} + 2 \times 3^{m-2}} \quad (5.399)$$

compared with 2<sup>m</sup> for m applications of Newton's method (*NM*)<sup>m</sup>. It follows that for  $m > 1$  the convergence order of (*INM*)<sup>m</sup> is greater than that of (*NM*)<sup>m</sup>. Since the number of function evaluations is the same for both methods, (*INM*)<sup>m</sup> is more efficient.

Cox (1970) describes a bracketing method for real roots based on a generalization of Newton's method with bisection in cases where their Newton-like method fails. Let us start with  $a < \zeta < b$  where  $\zeta$  is a root and  $p(a)p(b) < 0$ . We

find an interpolating function  $y(x) = \frac{(x-c)}{(d_0+d_1x+d_2x^2)}$  such that it and its derivative agree with  $p(x)$  and  $p'(x)$  at  $a$  and  $b$ . Solving for the 4 parameters  $c, d_i$  gives a new approximation to  $\zeta$  namely:

$$c = a + \frac{(b-a)p_ap_b(p_b-p_a) - (b-a)^2p'_ap'_b}{2p_ap_b(p_b-p_a) - (b-a)(p_b^2p'_a + p_a^2p'_b)} \quad (5.400)$$

as  $a \rightarrow \zeta$ , or a similar equation with  $a$  and  $b$  interchanged if  $b \rightarrow \zeta$ , or a formula (equivalent to the above) symmetric in  $a$  and  $b$  for the first few iterations. If  $c$  falls outside  $[a,b]$  we use instead bisection i.e.:

$$c = \frac{1}{2}(a+b) \quad (5.401)$$

If now  $p(c)$  has the same sign as  $p(a)$ , we replace  $a$  by  $c$  and iterate again; otherwise we replace  $b$  by  $c$ . As  $a \rightarrow \zeta$   $c \rightarrow a - \frac{p_a}{p'_a}$ , i.e. the method reduces to Newton's, and so converges quadratically. But when  $a$  is far from  $\zeta$ , and  $p'_a \approx 0$ , Newton fails whereas the new method works by means of 5.401. In some numerical tests of polynomials up to degree 30 the average number of evaluations per root was about 7 (3.5 iterations), while 5.401 was used in about 6% of cases (5.400 in the rest).

## 5.10 Programs

Chapter 6 of Hammer et al (1995) gives an algorithm and C++ program based on the methods of Moore and Hansen described in Section 7 of this Chapter.

The program TOMS/681 implements Newton's method for systems (with bisection). It can probably be applied to a single polynomial.

The program TOMS/812: BPOLY by Tsai and Farouki (2001) contains a library of programs for polynomial operations including root-finding (with the polynomials given in Bernstein form). See Section 9 of this Chapter.

To download either of the above programs, send a message to `netlib@ornl.gov` saying e.g. send Alg681 from TOMS

Ellenberger (1961) gives an Algol program using Bairstow's and Newton's methods. See also Cohen (1962).

Lang and Frenzel (1994) refer to a C program based on their method described in Section 9 of this Chapter. See <http://www.dsp.rice.edu> click on "Software"; "polynomial root finders"; and "The algorithm of Lang and Frenzel"

### 5.11 Miscellaneous Methods Related to Newton's

Luther (1964) describes a method of factorizing  $p(z)$  into

$$s(z)t(z) \tag{5.402}$$

where

$$s(z) = \sum_{l=0}^m \rho_l z^l, \quad t(z) = \sum_{l=0}^{n-m} \beta_l z^l \tag{5.403}$$

We make an initial guess  $p_l^{(0)}$  ( $l = 0, \dots, m$ ) as approximations for  $\rho_l$ . Then  $b_j^{(i)}$  (approximations for  $\beta_j$ ) and  $p_l^{(i)}$  are related by

$$\sum_{l=0}^m p_l^{(i)} b_{j-l}^{(i)} = c_j \quad (j = 0, \dots, n) \tag{5.404}$$

$(b_s^{(k)} = 0 \text{ if } s < 0)$

We obtain a hopefully improved approximation

$$p_j^{(i+1)} = p_j^{(i)} + \gamma^{(i)} \sum_{l=0}^{j-1} p_l^{(i)} b_{j-l}^{(i)} \tag{5.405}$$

where  $\gamma^{(i)}$  is somewhat arbitrary, but depends on the  $p_l$ . The author shows that if the  $p_l^{(0)}$  are close enough to the  $\rho_l$  the method converges.

The case  $m=1$ , with

$$\gamma(p_1) = \frac{1}{f'(z_1)} \tag{5.406}$$

gives Newton's method. For general  $m$  Luther suggests  $\gamma(p_l) = \pm \frac{1}{b_0}$ .

Joseph et al (1989/90) introduce random variables into Newton's method. Let  $W$  be a subset of the real line,  $f$  a real function,  $\{X_k\}$  a sequence of random variables and  $x_k$  the realization of  $X_k$ . let

$$X_{k+1} = x_k - Y_k \text{ when } x_k \in W \tag{5.407}$$

$$Y_k = \frac{f(x_k) + Z_{1k}}{f'(x_k) + Z_{2k}} \tag{5.408}$$

$X_{k+1}$  is uniform over  $W$  if  $x_k \notin W$

where  $Z_{1k}, Z_{2k}$  are independent random variables. In some (but not all) test cases the randomized method converged where the normal Newton's method did not.

Bodmer (1962) gives a method for complex zeros using polar coordinates. Let  $z = re^{i\theta}$ . We seek the roots of

$$p(z) = \sum_{m=0}^n c_m z^m = 0 = C + iS \quad (5.409)$$

where by De Moivre's theorem

$$C = \sum_{m=0}^n c_m r^m \cos(m\theta) \quad (5.410)$$

$$S = \sum_{m=0}^n c_m r^m \sin(m\theta) \quad (5.411)$$

Suppose  $(r_0, \theta_0)$  is an initial guess and  $(r_0 + \Delta r, \theta_0 + \Delta\theta)$  is the true solution. Then expanding C and S by Taylor's series as far as the linear terms we have:

$$0 = C(r_0, \theta_0) + \Delta r C_r(r_0, \theta_0) + \Delta\theta C_\theta(r_0, \theta_0) \quad (5.412)$$

$$0 = S(r_0, \theta_0) + \Delta r S_r(r_0, \theta_0) + \Delta\theta S_\theta(r_0, \theta_0) \quad (5.413)$$

where

$$C_r = \frac{\partial C}{\partial r} = \sum_{m=1}^n m c_m r^{m-1} \cos(m\theta) \quad (5.414)$$

$$C_\theta = \frac{\partial C}{\partial \theta} = - \sum_{m=1}^n m c_m r^m \sin(m\theta) \quad (5.415)$$

$$S_r = \frac{\partial S}{\partial r} = \sum_{m=1}^n m c_m r^{m-1} \sin(m\theta) \quad (5.416)$$

$$S_\theta = \frac{\partial S}{\partial \theta} = \sum_{m=1}^n m c_m r^m \cos(m\theta) \quad (5.417)$$

and

$$C_\theta = -r S_r, \quad S_\theta = r C_r \quad (5.418)$$

Hence by Cramer's rule (and using 5.418)

$$\Delta r = \frac{r(SC_\theta - CS_\theta)}{C_\theta^2 + S_\theta^2}; \quad \Delta\theta = -\frac{(CC_\theta + SS_\theta)}{C_\theta^2 + S_\theta^2} \quad (5.419)$$

where  $C$ ,  $S$ ,  $C_\theta$ ,  $S_\theta$  are evaluated at  $(r_0, \theta_0)$ . They suggest using Graeffe's method initially to obtain an approximate value of  $r$ . Then the corresponding  $\theta_0$  can be found as a zero of

$$|p(z)|^2 = C^2 + S^2 \quad (5.420)$$

At a zero,  $C^2 + S^2$  is a minimum, hence

$$CC_\theta + SS_\theta = 0 \quad (5.421)$$

which can be solved by interpolating on  $\theta$ . For equations with complex coefficients they suggest considering

$$p(z)\bar{p}(z) = 0 \quad (5.422)$$

which has real coefficients so that Graeffe's method can still be applied to find  $r$ .

Beyer (1964) describes a homotopy-version of Newton's method: assume that our equation takes the form

$$f(x, \alpha) = 0 \quad (5.423)$$

where  $\alpha$  is a real parameter. Assume a root of 5.423 is known for  $\alpha = \alpha_0$ , say  $x(\alpha_0)$ . We desire a root for  $\alpha = \alpha_N \neq \alpha_0$ . Suppose that for each  $\alpha$  in the range  $\alpha_0 < \alpha \leq \alpha_N$ ,  $f(x, \alpha)$  has a zero  $x(\alpha)$  and that

$$\frac{\partial f}{\partial x}(x(\alpha), \alpha) \neq 0 \quad (5.424)$$

with a further condition (see the cited reference). The author selects  $\alpha_i$  ( $i = 1, \dots, n$ ) so that

$$\alpha_0 < \alpha_1 < \dots < \alpha_N \quad (5.425)$$

The equations

$$f(x, \alpha_i) = 0 \quad (i = 1, \dots, N) \quad (5.426)$$

are solved successively by Newton's method with

$$x_1(\alpha_{i+1}) = x(\alpha_i) \quad (5.427)$$

To ensure convergence we must take

$$\Delta\alpha_i = \alpha_{i+1} - \alpha_i < \left| \frac{f_x^2}{2f_\alpha f_{xx}} \right| \quad (5.428)$$

where  $f_x = \frac{\partial f}{\partial x}$  etc.

Wu (2005) defines a slightly different homotopy method thus: let

$$H(x, t) \equiv tf(x) + (1-t)g(x) = 0 \quad (5.429)$$

where  $t \in [0, 1]$  and  $f(x)$  is the function whose roots are being sought. So

$$H(x, 0) = g(x), H(x, 1) = f(x) \quad (5.430)$$

and we apply Newton's iterations to  $H(x, t)$ , varying  $t$  from 0 to 1 as in Beyer's method. He suggests using  $g(x) = Cx + K$  or  $Ce^x + K$  where  $C, K$  are non-zero constants. In some numerical tests on a cubic polynomial the homotopy method succeeded for all 4 starting points tried, whereas plain Newton diverged in 3 of them.

Pomientale (1974) describes yet another homotopy method based on Newton, which is modified to avoid points where  $f'(x) = 0$ . It is apparently proved to be globally convergent.

Sharma (2005) describes a hybrid Newton-Steffensen method:

$$x_{i+1} = x_i - \frac{f^2(x_i)}{f'(x_i)(f(x_i) - f(x_i - \frac{f(x_i)}{f'(x_i)}))} \quad (5.431)$$

This is of third order and uses 3 evaluations per step. In tests on 5 functions the hybrid method converged in all cases, whereas in 4 of the cases either Newton or Steffensen (or both) failed by themselves.

Brent (1976) considers root-finding in (variable) multi-precision ("mp") arithmetic. Suppose that the evaluation with error  $\approx O(2^{-n})$  ("to precision  $n$ ") of the function or its derivative(s) takes time

$$w(cn) \approx c^\alpha w(n) \quad (5.432)$$

Here  $\alpha$  varies with the software and/or hardware. The author defines the discrete Newton mp method ( $N_1$ ) as

$$x_{i+1} = x_i - \frac{h_i f(x_i)}{f(x_i + h_i) - f(x_i)} \quad (5.433)$$

He states that to obtain the root to precision  $n$  requires 2 evaluations with precision  $n$ , preceded by two with precision  $\frac{n}{2}$ , etc. Hence the time

$$t(n) \approx 2w(n) + 2w(\frac{n}{2}) + .. \quad (5.434)$$

so the asymptotic constant  $C_{N_1}(\alpha)$  is

$$\frac{2}{1 - 2^{-\alpha}} \quad (5.435)$$

Here  $C(\alpha)$  is defined by

$$t(n) \approx C(\alpha)w(n) \quad (5.436)$$

He also defines a Secant-like method  $S_k$  by

$$x_{i+1} = x_i - f(x_i) \left[ \frac{x_i - x_{i-k}}{f(x_i) - f(x_{i-k})} \right] \tag{5.437}$$

We choose  $k$  to minimize  $C_{S_k}(\alpha)$ , giving the “optimal secant method”  $S_1$  if  $\alpha < 4.5$ , or  $S_2$  if  $\alpha \geq 4.5$ . Finally he defines inverse quadratic interpolation  $Q$  (which is always more efficient than  $S_1$ , but less than  $S_2$  if  $\alpha > 5.1$ ) and an “optimal inverse interpolating method” (I) (see the cited paper). He ranks these methods as follows: I best if  $1 \leq \alpha \leq 5.1$

$S_2$  best if  $5.1 < \alpha \leq 8.7$

$N_1$  best if  $\alpha > 8.7$

Chen (1990), (1992), (1993) describes a variation on Newton's method (called a “cluster-adapted” formula) which converged, usually very fast, for a number of hard examples involving clusters of multiple roots (where Newton failed). It is:

$$z_{i+1} = z_i - n \frac{(Q^{\frac{n}{\mu}} - 1) f(z_i)}{(Q - 1) f'(z_i)} \tag{5.438}$$

where

$$Q = \frac{(\frac{n}{\mu} - 1)}{(\frac{n}{p} - 1)} \tag{5.439}$$

and

$$\mu = \frac{(f'(z_i))^2}{(f'(z_i))^2 - f(z_i)f''(z_i)} \tag{5.440}$$

$p$  is best chosen as follows : Let

$$w = \frac{n}{|\frac{n}{\mu} - 1| + 1} \tag{5.441}$$

$$q = \text{floor}(.5 + w) \tag{5.442}$$

then if  $n > w \geq 1$  set  $p = q$ ,  
otherwise set  $p = 1$ .

As initial guess Chen takes  $A+R$  where the centroid

$$A = \frac{-c_{n-1}}{nc_n}; \tag{5.443}$$

$$R = \left(-\frac{f(A)}{c_n}\right)^{\frac{1}{n}} = \tag{5.444}$$

distance from  $A$  of roots of polynomial  $\phi(z) = c_n[(z - A)^r - R^r]^m$  and  $rm = n$ .

There is a danger of “rebounding” to a far-away point, if an iteration lands near the centroid of a symmetric cluster. This is detected by the following test :

If  $p \geq 2$  and if either

$$(a) \left| \frac{f(z_{i+1})}{f(z_i)} \right| \geq 20, \text{ and/or} \quad (5.445)$$

$$(b) \left| \frac{f(z_{i+2})}{f(z_{i+1})} \right| \geq .75 \quad (5.446)$$

with

$$\frac{|z_{i+2} - z_{i+1}|}{2|z_{i+1} - A|} > \frac{\pi p}{n} \quad (5.447)$$

then  $z_{i+2}$  is judged to have rebounded.. The cure is to replace  $z_{i+2}$  by a new  $z_{i+2}$  given by

$$z_{i+1} + R' e^{i\sigma} \quad (5.448)$$

where

$$R' = (-f(z_{i+1}))^{\frac{1}{p}} \quad (5.449)$$

and  $\sigma$  puts  $z_{i+2}$  on the line between  $z_{i+1}$  and the old  $z_{i+2}$ . In some tests with low-degree polynomials the program based on this method (SCARFS) was 3 times faster than the Jenkins-Traub method and 12 times faster than Muller's.

Traub (1974) gives a sort of combined secant-Newton method as follows:

$$z_0 = x_i; z_1 = z_0 - \frac{f(z_0)}{f'(z_0)} \quad (5.450)$$

$$x_{i+1} = z_1 - \frac{f(z_1)f(z_0)}{[f(z_1) - f(z_0)]^2} \frac{f(z_0)}{f'(z_0)} \quad (5.451)$$

This is of order 4 and requires 3 evaluations per step, so its efficiency is  $\log(\sqrt[3]{4}) = .200$

Chun (2005) describes a family of methods which modify Newton's method to obtain a higher rate of convergence, but using derivatives of a lower order than many other methods of the same convergence rate. This is sometimes an advantage. The family includes

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} - \frac{f(x_{i+1}^*)}{f'(x_i)} \quad (5.452)$$



where

$$x_{i+1}^* = x_i - \frac{f(x_i)}{f'(x_i)} \quad (5.453)$$

This is of convergence rate 3 and efficiency  $\log(\sqrt[3]{3}) = .159$   
Also we have

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} - 2\frac{f(x_{i+1}^*)}{f'(x_i)} + \frac{f(x_{i+1}^*)f'(x_{i+1}^*)}{[f'(x_i)]^2} \quad (5.454)$$

with  $x_{i+1}^*$  as before. This latter method is of order 4 and uses 4 function evaluations, so its efficiency is  $\log(\sqrt[4]{4})$ , the same as Newton's method. Other more complicated methods have even lower efficiency.

Manta et al (2005) give two families of methods similar to Newton's. One is

$$x_{i+1} = x_i - \frac{f(x_i)}{[f'(x_i) \pm pf(x_i)]} \quad (5.455)$$

where the sign is chosen so that  $pf(x_i)$  and  $f'(x_i)$  have the same sign (otherwise  $p$  is arbitrary). This is identical to the method given by Wu (2000) and described in section 4 of this Chapter. Another class of methods is

$$x_{i+1} = x_i - \frac{2f(x_i)}{f'(x_i) \pm \sqrt{f'^2(x_i) + 4p^2f^2(x_i)}} \quad (5.456)$$

where the sign is chosen to make the denominator largest in magnitude. The authors show that convergence is quadratic. Again  $p$  is arbitrary, but in 12 numerical tests of 5.456 with  $p = 1$  this new method converged to the true root in every case, whereas Newton failed in at least 6 of the cases. Moreover, when Newton did converge it often took many more iterations than 5.456.

## 5.12 References for Chapter 5

Adams, D. (1967), A stopping criterion for polynomial root-finding, *Comm. Ass. Comput. Mach.* **10**, 655-658

Aho, A.V., Hopcroft, J.E. and Ullman, J.D. (1976), *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, Mass.

Aitken, A.C. (1926), On Bernoulli's Numerical Solution of Algebraic Equations, *Proc. Roy. Soc. Edinburgh* **46**, 289-305

Akl, S.G. (1989), *The Design and Analysis of Parallel Algorithms*, Prentice-Hall, Englewood Cliffs, N.J., 209-211

- Alefeld, G. (1981), Bounding the Slope of Polynomial Operators and Some Applications, *Computing* **26**, 227-237
- and Potra, F.A. (1988), On Two Higher Order Enclosing Methods of J.W. Schmidt, *Zeit. ang. Math. Mech.* **68**, 331- 337
- and ————— (1995), Algorithm 748: Enclosing Zeros of Continuous Functions, *ACM Trans. Math. Software* **21**, 327-344
- Alexander, W.J. (1961), Certification of Algorithm 30, *Comm. Ass. Comput. Mach.* **4**, 238
- Arthur, D.W. (1972), The Use of Interval Arithmetic to Bound the Zeros of Real Polynomials, *J. Inst. Math. Appl.* **10**, 231-237
- Atkinson, K.E. (1989), *An Introduction to Numerical Analysis (2/E)*, Wiley, New York, p64
- Barna, B. (1951), Über das Newtonische Verfahren zur Annäherung von Wurzeln algebraische Gleichungen, *Publ. Math. Debrecen* **2**, 50-63
- Ben-Israel, A. (1997), Newton's method with modified functions, in "Recent Developments in Optimization Theory and Nonlinear Analysis", eds. Y. Censor and S. Reich, Amer. Math. Soc., Providence, Rhode Island, 39-50
- Beyer, W.A. (1964), A Note on Starting the Newton-Raphson Method, *Comm. Ass. Comput. Mach.* **7**, 442
- Bini, D. and Pan, V.Y. (1998), Computing matrix eigenvalues and polynomial zeros where the output is real, *SIAM J. Computing* **27**, 1099-1115
- Bodmer, W.F. (1962), A method of evaluating the complex zeros of polynomials using polar coordinates, *Proc. Camb. Phil. Soc.* **58**, 52-57
- Brent, R.P. (1976), The complexity of multiple-precision arithmetic, in *The Complexity of Computational Problem-Solving*, eds. R.S. Anderssen and R.P. Brent, University of Queensland Press, 126-165
- Burgstahler, S. (1986), An Algorithm for Solving Polynomial Equations, *Amer. Math. Monthly* **93**, 421-430
- Carniel, R. (1994), A quasi cell-mapping approach to the global analysis of Newton's root-finding algorithm, *Appl. Numer. Math.* **15**, 133-152

Chanabasappa, M.N. (1979), A Note on the Computation of Multiple Zeros of Polynomials by Newton's Method, *BIT* **19**, 134-135

Chen, T.-C. (1990), Iterative zero-finding revisited, in *Computational Techniques and Applications (CTAC-89)*, eds. W.I. Hogarth and B.J. Noye, Hemisphere Publ. Corp., New York, 583-590

——— (1992), Globally Convergent Polynomial Iterative Zero-Finding using APL, *APL Quote Quod* **23** (1), 52-59

——— (1993), SCARFS: An Efficient Polynomial Zero-Finder System, *APL Quote Quod* **24** (1), 47-54

Chun, C. (2005), Iterative Methods Improving Newton's Method by the Decomposition Method, *Comput. Math. Appl.* **50**, 1559-1568

Clegg, D.B. (1981), On Newton's method with a class of rational functions, *J. Comput. Appl. Math.* **7**, 93-100

Cohen, K.J. (1962), Certification of Algorithm 30, *Comm. Ass. Comput. Mach.* **5**, 50

Costabile, F., Gualtieri, M.I., and Luceri, R. (2001), A new iterative method for the computation of the solution of nonlinear equations, *Numer. Algs.* **28**, 87-100

Cosnard, M. and Masse, C. (1983), Convergence presque partout de la méthode de Newton, *C.R. Acad. Sci. Paris* **297**, 549-552

Cox, M.G. (1970), A bracketing technique for computing a zero of a function, *Computer J.* **13**, 101-102

Dargel, R.H., Loscalzo, F.R., and Witt, T.H. (1966), Automatic Error Bounds on Real Zeros of Rational Functions, *Comm. Ass. Comput. Mach.* **9**, 806-809

Dawson, J.E. (1982), A Formula Approximating the Root of a Function, *IMA J. Numer. Anal.* **2**, 371-375

Derr, J.I. (1959), A Unified Process for the Evaluation of the Zeros of Polynomials over the Complex Number Field, *Math. Tables Aids Comput.* **13**, 29-36

Dimitrova, N.S. (1994), On a Parallel Method for Enclosing Real Roots of Nonlinear Equations, in *Advances in Numerical Methods and Applications*, ed. I.T. Dimov et al, World Scientific, Singapore, 78-84

- Dong, C. (1987), A Family of Multipoint Iterative Functions for Finding Multiple Roots of Equations, *Intern. J. Computer Math.* **21**, 363-367
- Donovan, G.C., Miller, A.R., and Moreland, T.J. (1993), Pathological Functions for Newton's Method, *Amer. Math. Monthly* **100**, 53-58
- Ellenberger, K.W. (1960), On Programming the Numerical Solution of Polynomial Equations, *Comm. Ass. Comput. Mach.* **3**, 644-647
- (1961), ALGORITHM 30: Numerical Solution of the Polynomial Equation, *Comm. Ass. Comput. Mach* **4**, 643
- Forsythe, G.E. (1958), Singularity and Near-Singularity in Numerical Analysis, *Amer. Math. Monthly* **65**, 229-240
- Franklin, F. (1881), On Newton's Method of Approximation, *Amer. J. Math.* **4**, 275-276
- Frontini, M. and Sormani, E. (2003), Some variant of Newton's method with third-order onvergence, *Appl. Math. Comput.* **140**, 419-426
- Gargantini, I. (1976), Comparing parallel Newton's method with parallel Laguerre's method, *Comput. Math. Appls.* **2**, 201-206
- Garwick, J.V. (1961), The limit of a converging sequence, *BIT* **1**, 64
- Gilbert, W.J. (1994), Newton's Method for Multiple Roots, *Computers and Graphics* **18**, 227-229
- Grant, J.A. and Hitchins, G.D. (1973), The solution of polynomial equations in interval arithmetic, *Computer J.* **16**, 69-72
- Hammer, R. et al (1995), Chap. 6. Nonlinear equations in one variable, in *C++ Toolbox for Verified Computing I: Basic Numerical Problems*, Springer-Verlag, Berlin, 93-112
- Hansen, E.R. (1968), On Solving Systems of Equations Using Interval Arithmetic, *Math. Comp.* **22**, 374-384
- (1978A), A globally convergent interval method for computing and bounding real roots, *BIT* **18**, 415-424
- (1978B), Interval Forms of Newton's Method, *Computing* **20**, 153-163

————— (1992), *Global Optimization using Interval Analysis*, Marcel Dekker, New York (especially Chap. 7).

————— and Patrick, M. (1976), Estimating the Multiplicity of a Root, *Numer. Math.* **27**, 121-131

Hanson, R.J. (1970), Automatic error bounds for real roots of polynomials having interval coefficients, *Computer J.* **13**, 284-288

He, J.-H. (2004), A Modified Newton-Raphson method, *Comm. Num. Meth. Eng.* **20**, 801-805

Henrici, P. (1971), Circular arithmetic and the determination of polynomial zeros, in *Conference on Applications of Numerical Analysis* (LNM 228), ed. J.L. Morris, Springer-Verlag, Berlin, 86-92

Herzberger, J. (1986), On the R-Order of Some Recurrences with Applications to Inclusion-Methods, *Computing* **36**, 175-180

Hines, J. (1951), On approximating the roots of an equation by iteration, *Math. Mag.* **24**, 123-127

Homeier, H.H.H. (2005), On Newton-type methods with cubic convergence, *J. Comput. Appl. Math.* **176**, 425-432

Hopgood, D.J. and McKee, S. (1974), Improved Root-Finding Methods Derived from Inverse Interpolation, *J. Inst. Maths. Applics.* **14**, 217-228

Householder, A.S. (1970), *The Numerical Treatment of a Single Nonlinear Equation*, McGraw-Hill, New York, Chap. 4, Secs. 0-3

Hubbard, J., Schleicher, D. and Sutherland, S. (2001), How to find all roots of complex polynomials by Newton's method, *Invent. Math.* **146**, 1-33

Igarashi, M. (1982), Zeros of Polynomial and an Estimation of its Accuracy, *J. Inform. Proc.* **5**, 172-175

————— (1985), Practical problems arising for finding roots of nonlinear equations, *Appl. Numer. Math.* **1**, 433-455

Jarratt, P. (1966), Multipoint iterative methods for solving certain equations, *Computer J.* **8**, 393-400

————— (1966B), Some Fourth-Order Multipoint Iterative Methods for Solving

Equations, *Math. Comp.* **20**, 434-437

——— (1970), A review of methods for solving nonlinear algebraic equations in one variable, in *Numerical Methods for Nonlinear Algebraic Equations*, ed P. Rabinowitz, Gordon and Breach, London, 1-26

Joseph, G., Levine, A., and Liukkonen, J. (1989/90), Randomized Newton-Raphson, *Appl. Numer. Math.* **6**, 459-469

King, R.F. (1971), A fifth-order family of modified Newton methods, *BIT* **11**, 404-412

——— (1972), Tangent Methods for Nonlinear Equations, *Numer. Math.* **18**, 298-304

——— (1973), A Family of Fourth-Order Methods for Nonlinear Equations, *SIAM J. Numer. Anal.* **10**, 876-879

——— (1979), An Extrapolation Method of Order Four for Linear Sequences, *SIAM J. Numer. Anal.* **16**, 719-725

——— (1980), An Efficient One-Point Extrapolation Method for Linear Convergence, *Math. Comp.* **35**, 1285-1290

——— (1983A), Improving the Van de Vel Root-Finding Method, *Computing* **30**, 373-378

——— (1983B), Anderson-Bjorck for Linear Sequences, *Math. Comp.* **41**, 591-596

Kirrinnis, P. (1997), Newton Iteration Towards a Cluster of Polynomial Zeros, in *Foundations of Computational Mathematics*, eds. F. Cucker and M. Shub, Springer-Verlag, Berlin, 193-215

Kizner, W. (1964), A Numerical Method for Finding Solutions of Nonlinear Equations, *SIAM J.* **12**, 424-428

Kogan, T.I. (1967), Formulation of higher-order iteration processes, *USSR Comp. Math. Math. Phys.* **7**, 423-424

Kollerstrom, N. (1992), Thomas Simpson and 'Newton's method of approximation': an enduring myth, *British J. Hist. Sci.* **25**, 347-354

Lagouanelle, J.-L. (1966), Sur une méthode de calcul de l'ordre de multiplicité des

zéros d'un polynome, *C.R. Acad. Sci. Paris* **262A**, 626-627

Lagrange, J.L. (1798), *Traité de la Résolution des Équations Numériques*, Paris

Lang, M. and Frenzel, B.-C. (1994), Polynomial Root-Finding, *IEEE Signal Proc. Lett.* **1**, 141-143

Levin, D. (1973), Development of Non-Linear Transformations for Improving Convergence of Sequences, *Intern. J. Comput. Math.* **B3**, 371-388

Lin, Q. and Rokne, J.G. (1995), An interval iteration for multiple roots of transcendental equations, *BIT* **35**, 561-571

Locher, F. (1993), A stability test for real polynomials, *Numer. Math.*, **66**, 33-40

——— and Skrzipek, M.-R. (1995), An Algorithm for Locating All Zeros of a Real Polynomial, *Computing* **54**, 359-375

Luther, M.A. (1964), A Class of Iterative Techniques For the Factorization of Polynomials, *Comm. Ass. Comput. Mach.* **7**, 177-179

Madsen, K. (1973), A Root-Finding Algorithm Based on Newton's Method, *BIT* **13**, 71-75

Mamta et al (2005), On a class of quadratically convergent iteration formulae, *Appl. Math. Comput.* **166**, 633-637

Matthews, J.H. and Fink, K.D. (1999), *Numerical Methods Using Matlab 3/E*, Prentice-Hall

McNamee, J.M. (1988), A Comparison of Methods for Terminating Polynomial Iterations, *J. Comput. Appl. Math.* **21**, 239-244

——— (1998), A Comparison of Methods for Accelerating Convergence of Newton's Method for Multiple Polynomial Roots, *SIGNUM Bull* **33 (2)**, 17-22

Moore, R.E. (1966), *Interval Analysis*, Prentice Hall, Englewood Cliffs, N.J. (especially Chap. 7).

Murakami, T. (1978), Some Fifth Order Multipoint Iterative Formulas for Solving Equations, *J. Inform. Proc.* **1**, 138-139

Nesdore, P.F. (1970), The determination of an algorithm which uses the mixed strategy technique for the solution of single nonlinear equations, in *Numerical Methods*

for *Nonlinear Algebraic Equations*, ed. P. Rabinowitz, Gordon and Breach, 27-46

Neta, B. (1979), A Sixth-Order Family of Methods for Nonlinear Equations, *Intern. J. Computer Math. Sec. B* **7**, 157-161

——— (1981), On a Family of Multipoint Methods for Non-linear Equations, *Intern. J. Computer Math.* **9**, 353-361

Ostrowski, A.M. (1973), *Solution of Equations in Euclidean and Banach Spaces*, Academic Press, New York

Patrick, M.L. (1972), A Highly Parallel Algorithm for Approximating All Zeros of a Polynomial with Only Real Zeros, *Comm. Ass. Comput. Mach.* **11**, 952-955

Petkovic, L.D., Trickovic, S., and Petkovic, M.S. (1997), Slope Methods of Higher Order for the Inclusion of Complex Roots of Polynomials, *Reliable Computing* **3**, 349-362

Petkovic, M.S. (1981), On an Interval Newton Method Derived from Exponential Curve Fitting, *Zeit. ang. Math. Mech.* **61**, 117-119

——— (1987), Some interval iterations for finding a zero of a polynomial with error bounds, *Comput. Math. Appl.* **14**, 479-495

——— and Herzberger, J. (1991), Hybrid inclusion algorithms for polynomial multiple complex zeros in rectangular arithmetic, *Appl. Numer. Math* **7**, 241-262

———, Herceg, D.D., and Ilic, S.M. (1997), *Point estimation Theory and Its Applications*, Institute of Mathematics, Novi Sad, Yugoslavia

Pomentale, T. (1974), Homotopy Iterative Methods for Polynomial Equations, *J. Inst. Maths. Applics.* **13**, 201-213

Presic, M.D. (1978), On Ostrowski's fundamental existence theorem, *Publ. Inst. Math. (Now. Ser.)* **24 (38)**, 125-132

——— (1979), on Ostrowski's fundamental existence theorem in the complex case, *Publ. Inst. Math. Debrecen (Now. Ser.)* **26 (40)**, 229-232

Rall, L.B. (1966), Convergence of the Newton Process to Multiple Solutions, *Numer. Math.* **9**, 23-37

Raphson, J. (1690), *Analysis aequationum universalis ...*, London. Microfilm copy:



University Microfilms, Ann Arbor, MI

Renegar, J. (1987), On the worst-case arithmetic complexity of approximating zeros of polynomials, *J. Complexity* **3**, 90-113

Revol, N. (2003), Interval Newton iteration in multiple precision for the univariate case, *Numer. Algs.* **34**, 417-426

Rokne, J. (1973), Automatic Errorbounds for Simple Zeros of Analytic Functions, *Comm. Ass. Comp. Mach.* **16**, 101-104

——— and Lancaster, P. (1969), Automatic Errorbounds for the Approximate Solution of Equations, *Computing* **4**, 294-303

Schleicher, D. (2002), On the number of iterations of Newton's method for complex polynomials, *Ergod. Th. Dynamic Sys.* **22**, 935-945

Schroder, E. (1870), Ueber unendliche viele Algorithmen zur Auflosung der Gleichungen, *Math. Ann.* **2**, 317-365

Sharma, J.R. (2005), A composite third-order Newton-Steffensen method for solving nonlinear equations, *Appl. Math. Comput.* **169**, 242-246

Shedler, G.S. (1967), Parallel Numerical Methods for the Solution of Equations, *Comm. Ass. Comput. Mach.* **10**, 286-291

Simpson, T. (1740), *Essays... Mathematicks...*, London.

Smale, S. (1986), Newton's method estimates from data at one point, in *The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics*, Ed. R.E. Ewing et al, Springer-Verlag, New York, 185-196

Stoer, J. and Bulirsch, R. (1980), *Introduction to Numerical Analysis*, Springer-Verlag, New York

Tikhonov, O. N. (1976), A generalization of Newton's method of computing the roots of algebraic equations, *Sov. Math. (Iz. Vuz.)* **20 (6)**, 109-111

Traub, J.F. (1964), *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs

——— (1967), On Newton-Raphson Iteration, *Amer. Math. Monthly* **74**, 996-998

- (1974), Theory of Optimal Algorithms, in *Software for Numerical Mathematics*, ed D.J. Evans, Academic Press, London, 1-15
- Tsai, Y.-F. and Farouki, R.T. (2001), Algorithm 812: BPOLY: An Object-Oriented Library of Numerical Algorithms for Polynomials in Bernstein Form, *ACM Trans. Math. Software* **27**, 267-296
- Vander Straeten, M. and Van de Vel, H. (1992), Multiple Root-Finding Methods, *J. Comput. Appl. Math.* **40**, 105-114
- Victory, H.D. and Neta, B. (1983), A Higher Order Method for Multiple Zeros of Nonlinear Functions, *Intern. J. Computer Math.* **12**, 329-335
- Vignes, J. (1978), New methods for evaluating the validity of the results of mathematical computations, *Math. Computers Simul.* **20**, 227-249
- Wallis, J. (1685), *A Treatise of Algebra both Historical and Practical*, London.
- Wang, X. and Han, D. (1989), On dominating sequence method in the point estimate and Smale's theorem, *Sci. Sinica Ser. A*, 905-913
- Weerakom, S. and Fernando, T.G.I. (2000), A Variant of Newton's Method with Accelerated Third-Order Convergence, *Appl. Math. Lett.* **13**, 87-93
- Werner, W. (1982), Some Supplementary Results on the  $1 + \sqrt{2}$  Order Method for the Solution of Nonlinear Equations, *Numer. Math.* **38**, 383-392
- Whiteside, D.T. (ed.) (1967-1976), *The Mathematical Papers of Isaac Newton Vols. I-VII*, Cambridge University Press, Cambridge
- Wiethoff, A. (1996), Enclosing all zeros of a nonlinear function on a parallel computer system, *Zeit. ang. Math. Mech.* **76**, 585-586
- Wu, T.-M. (2005), A study of convergence on the Newton-homotopy continuation method, *Appl. Math. Comp.* **168**, 1169-1174
- Wu, X.-Y. (2000), A new continuation Newton-like method and its deformation, *Appl. Math. Comp.* **112**, 75-78
- Yakoubsohn, J.-C. (2000), Finding a Cluster of Zeros of Univariate Polynomials, *J. Complexity* **16**, 603-638
- Ypma, T.J. (1983), Finding a Multiple Zero by Transformations and Newton-Like Methods, *SIAM Rev.* **25**, 365-378

——— (1995), Historical Development of the Newton- Raphson Method, *SIAM Rev.* **37**, 531-551

# Chapter 6

## Matrix Methods

### 6.1 Methods Based on the Classical Companion Matrix

For many years people solved eigenvalue problems by finding roots of the characteristic polynomial of the matrix in question. In more recent times this process has been reversed: one solves a polynomial by finding a matrix whose characteristic polynomial is identical to the given polynomial, and then finding its eigenvalues.

Brand (1964) defines the classical companion matrix of a monic polynomial  $p(\lambda)$  as follows:

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 0 & 1 \\ -c_0 & -c_1 & \dots & \dots & -c_{n-2} & -c_{n-1} \end{bmatrix} \quad (6.1)$$

it is also given by various authors as

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & \dots & \dots & 0 & -c_0 \\ 1 & 0 & \dots & \dots & 0 & -c_1 \\ 0 & 1 & 0 & \dots & 0 & -c_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 & -c_{n-1} \end{bmatrix} \quad (6.2)$$

and there are other rearrangements of the elements in the literature. It is sometimes referred to as the Frobenius companion matrix, implying that it was discovered by that author, but we have not discovered the original reference. The important property of  $\mathbf{C}$  is that

$$|\mathbf{C} - \lambda\mathbf{I}| = (-1)^n p(\lambda) \quad (6.3)$$

which Brand proves as follows: we write

$$|\mathbf{C} - \lambda\mathbf{I}| = \begin{bmatrix} -\lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & -\lambda & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -\lambda & 1 \\ -c_0 & -c_1 & \dots & \dots & -c_{n-2} & -c_{n-1} - \lambda \end{bmatrix} \quad (6.4)$$

Now multiply columns 2,3,...,n of this determinant by  $\lambda$ ,  $\lambda^2$ , ...,  $\lambda^{n-1}$  and add them to the first column, so that all elements of that column become zero except the last which is now  $-p(\lambda)$ . Since the cofactor of this is  $(-1)^{n-1}$ , we have

$$|\mathbf{C} - \lambda\mathbf{I}| = (-1)^n p(\lambda) \quad (6.5)$$

i.e.  $p(\lambda)$  is the characteristic polynomial of  $\mathbf{C}$ .

$\mathbf{C}$  arises naturally in the theory of differential equations, when one replaces a linear equation

$$f(D)y = 0, \quad (D = \frac{d}{dt}) \quad (6.6)$$

by a system of n first order equations. For example

$$y^{(3)} + ay^{(2)} + by' + cy = 0 \quad (6.7)$$

is replaced by

$$\begin{aligned} y' &= u \\ u' &= v \\ v' &= -cy - bu - av \end{aligned} \quad (6.8)$$

The matrix of coefficients of the right-hand-side is the companion of

$$\lambda^3 + a\lambda^2 + b\lambda + c = 0 \quad (6.9)$$

The equation  $p(\lambda) = 0$  may be written as

$$\mathbf{C}\mathbf{e}(\lambda) = \lambda\mathbf{e}(\lambda) \quad (6.10)$$

where

$$\mathbf{e}^T(\lambda) = (1, \lambda, \lambda^2, \dots, \lambda^{n-1}) \quad (6.11)$$

for the first n-1 equations of 6.10 are of the form  $\lambda^i = \lambda^i$  ( $i = 1, 2, \dots, n-1$ ) (which is always true) and the last is

$$-c_0 - c_1\lambda - \dots - c_{n-1}\lambda^{n-1} = \lambda^n \quad (6.12)$$

Thus, if  $\lambda_i$  is an eigenvalue of  $\mathbf{C}$  (and hence a zero of  $p(\lambda)$ ) 6.12 is satisfied and so 6.10 is also satisfied for  $\lambda = \lambda_i$ . The rank of  $\mathbf{C} - \lambda_i\mathbf{I}$  is always n-1 even when  $\lambda_i$

is a multiple zero; for the minor of element (n,1) has a determinant = 1. So  $\lambda_i$  is associated with only one eigenvector  $\mathbf{e}_i = \mathbf{e}(\lambda_i)$ . When  $\mathbf{C}$  has an eigenvalue  $\lambda_1$  (and hence  $p(\lambda)$  has a root  $\lambda_1$ ) of multiplicity  $k$ , then  $\lambda_1$  satisfies

$$p(\lambda) = p'(\lambda) = \dots = p^{(k-1)}(\lambda) = 0 \tag{6.13}$$

The first of these equations is equivalent to 6.10; the others are equivalent to equations derived from 6.10 by differentiations with respect to  $\lambda$ , such as

$$\mathbf{C}\mathbf{e}^{(j)}(\lambda) = \lambda\mathbf{e}^{(j)}(\lambda) + j\mathbf{e}^{(j-1)}(\lambda) \quad (j = 1, 2, \dots, k - 1) \tag{6.14}$$

So, if  $\lambda_1$  is a  $k$ -fold zero we have

$$\mathbf{C}\mathbf{e}_1 = \lambda_1\mathbf{e}_1 \tag{6.15}$$

and

$$\mathbf{C}\mathbf{e}_j = \lambda_1\mathbf{e}_j + \mathbf{e}_{j-1} \quad (j = 2, \dots, k) \tag{6.16}$$

where  $\mathbf{e}_1$  is the eigenvector and

$$\mathbf{e}_j = \frac{\mathbf{e}^{(j-1)}(\lambda_1)}{(j-1)!} \quad (j = 2, \dots, k) \tag{6.17}$$

are defined as “generalized” eigenvectors. Note that

$$\mathbf{e}_j^T = (0, 0, \dots, 1, \binom{j}{j-1}\lambda_1, \dots, \binom{n-1}{j-1}\lambda_1^{n-j}) \quad (j = 2, \dots, k) \tag{6.18}$$

where the 1 is in position  $j$  (the case  $j = 1$  is given by 6.11 with  $\lambda = \lambda_1$ ). Thus the entire set is linearly independent since the  $k \times n$  matrix formed from their components has rank  $k$ . Brand also shows that the inverse of  $\mathbf{C}$  is

$$\begin{bmatrix} -c_1/c_0 & -c_2/c_0 & \dots & \dots & -c_{n-1}/c_0 & -1/c_0 \\ 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 1 & 0 \end{bmatrix} \tag{6.19}$$

It is the companion of the reciprocal polynomial  $x^n p(\frac{1}{x})$  (rearranged).

Some authors do not assume that  $p(\lambda)$  is monic, in which case the  $c_i$  in 6.1 are each divided by  $c_n$ .

An early application of the companion matrix to actually finding roots is by Mendelsohn (1957). He does this by applying the power method: a vector  $\mathbf{x}^{(0)}$  is repeatedly multiplied by  $\mathbf{C}$  until convergence, i.e.  $\mathbf{C}\mathbf{x}^{(i)} = \lambda_1\mathbf{x}^{(i+1)}$  where  $\mathbf{x}^{(i+1)} \approx \mathbf{x}^{(i)}$  if  $\mathbf{x}^{(i)}$  is normalized so that its last component is 1, and  $\lambda_1$  is the

largest magnitude root.

Krishnamurthy (1960) also applies the power method, but accelerates it by using the Cayley-Hamilton theorem which states that a matrix satisfies its own characteristic polynomial, i.e.

$$-(c_0\mathbf{I} + c_1\mathbf{C} + \dots + c_{n-1}\mathbf{C}^{n-1}) = \mathbf{C}^n \quad (6.20)$$

Thus a power  $\mathbf{C}^m$  ( $m > n$ ) can be expressed as a polynomial in  $\mathbf{C}$  of degree  $n-1$ . We then multiply  $\mathbf{x}^{(0)}$  by  $\mathbf{C}^m$  for large  $m$ . The calculation of  $\mathbf{C}^r$  ( $r = 1, 2, \dots, n-1$ ) is made easy as all the first  $(n-1)$  columns of  $\mathbf{C}^i$  are the same as the second to  $n$ 'th column of  $\mathbf{C}^{i-1}$  ( $i = 2, \dots, n-1$ ). The author shows how to deal with complex or unimodular roots. He mentions that multiple roots are difficult by this method.

Stewart (1970) gives a method which is equivalent to the power method, but more efficient. He starts with an arbitrary polynomial  $h_0$  of degree  $< n$  but  $> 0$ , and applies

$$h_{i+1} = [h_i]^2 \pmod{p} \quad (6.21)$$

He then show that if  $|h_0(r_1)| > |h_0(r_i)|$  ( $r_i \neq r_1$ ) then  $h_k(z)/h_k(0)$  converges to  $\pi_1(z)/\pi_1(0)$ , where

$$\pi_1(z) = \frac{p(z)}{z - r_1} \quad (6.22)$$

and so

$$z - r_1 = \frac{p(z)}{\pi_1(z)} \quad (6.23)$$

Using the Cayley-Hamilton theorem he shows that

$$h_{i+1}(\mathbf{C}) = [h_i(\mathbf{C})]^2 \quad (6.24)$$

or

$$h_i(\mathbf{C}) = [h_0(\mathbf{C})]^{2^i} \quad (6.25)$$

i.e. the vector of coefficients of  $h_i$ , called  $\mathbf{h}_i$

$$= [h_0(\mathbf{C})]^{2^i} \mathbf{e}_1 \quad (6.26)$$

where

$$\mathbf{e}_1^T = (1, 0, \dots, 0) \quad (6.27)$$

Thus 6.21 is a variant of the power method applied to  $h_0(\mathbf{C})$  in which the matrix is squared at each step. For simple roots, the convergence is quadratic; for multiple

it is linear with ratio  $\frac{1}{2}$ .

Hammer et al (1995) give an Algorithm with verified bounds. They use a slightly more general companion matrix than most authors; that is they do NOT assume that the polynomial is monic, so that the  $c_i$  in the companion matrix in the form 6.2 are divided by  $c_n$  for  $i=0, \dots, n-1$ . With  $A$  as the companion matrix thus described they solve the eigenproblem

$$\mathbf{A}\mathbf{q}^* = z^*\mathbf{q}^* \tag{6.28}$$

or

$$(\mathbf{A} - z^*\mathbf{I})\mathbf{q}^* = 0 \tag{6.29}$$

where

$$\mathbf{q}^* = \begin{bmatrix} q_0^* \\ q_1^* \\ \dots \\ q_{n-1}^* \end{bmatrix} \tag{6.30}$$

is the eigenvector corresponding to the eigenvalue  $z^*$ .

Moreover,  $q_0^*, \dots, q_{n-1}^*$  are the coefficients of the deflated polynomial

$$\frac{p(z)}{z - z^*} \tag{6.31}$$

The  $q_i^*$  may be multiplied by an arbitrary factor, so to avoid the divisions in the companion matrix we set  $q_{n-1}^* = c_n$ , and then the others are given by

$$q_{i-1}^* = q_i^* z^* + c_i \quad (i = n - 1, \dots, 1) \tag{6.32}$$

Thus we have a system of nonlinear equations in  $q_0, \dots, q_{n-2}, z$ , and we write

$$\mathbf{x} = \begin{bmatrix} q_0 \\ q_1 \\ \dots \\ q_{n-2} \\ z \end{bmatrix} \equiv \begin{bmatrix} \mathbf{q} \\ z \end{bmatrix} \tag{6.33}$$

(the  $q_i^*, z^*$  are the solutions of 6.32).

The authors solve 6.29

(written as

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \tag{6.34}$$



by applying the simplified Newton's method for a system i.e.

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \mathbf{R}\mathbf{f}(\mathbf{x}^{(i)}) \quad (6.35)$$

where

$$\mathbf{R} = \mathbf{f}'(\mathbf{x}^{(0)})^{-1} \quad (6.36)$$

i.e. the inverse of the Jacobian  $\mathbf{J}$  of  $\mathbf{f}$  at  $\mathbf{x}^{(0)}$ . In our case

$$\mathbf{f}(\mathbf{x}) = f\left(\begin{bmatrix} \mathbf{q} \\ z \end{bmatrix}\right) = (\mathbf{A} - z\mathbf{I}) \begin{bmatrix} \mathbf{q} \\ c_n \end{bmatrix} \quad (6.37)$$

Then

$$\mathbf{J} = \mathbf{f}'(\mathbf{x}) = \left[ (\mathbf{A} - z\mathbf{I}) \begin{bmatrix} \mathbf{q} \\ c_n \end{bmatrix} \right]' \quad (6.38)$$

$$= (\mathbf{A} - z\mathbf{I}) \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \dots & 0 & q_0 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & q_{n-2} \\ 0 & \dots & 0 & c_n \end{bmatrix} \quad (6.39)$$

$$= \begin{bmatrix} -z & 0 & \dots & \dots & 0 & -q_0 \\ 1 & -z & 0 & \dots & 0 & -q_1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 & -z & -q_{n-2} \\ 0 & \dots & \dots & 0 & 1 & -c_n \end{bmatrix} \quad (6.40)$$

The authors show that

$$\mathbf{R} = \mathbf{J}^{(-1)} = \begin{bmatrix} -\frac{w_1}{w_0} & 1 & 0 & \dots & \dots & 0 \\ -\frac{w_2}{w_0} & 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -\frac{w_{n-1}}{w_0} & 0 & \dots & \dots & 0 & 1 \\ -\frac{1}{w_0} & 0 & \dots & \dots & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & z & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix} \dots$$

$$\begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & z \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix} \quad (6.41)$$

where

$$w_{n-1} = c_n; w_i = q_i + zw_{i+1} \quad (i = n-2, \dots, 0) \quad (6.42)$$

If  $\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{\mathbf{q}} \\ \tilde{z} \end{bmatrix}$  is an initial approximation (and the authors do not say how this should be obtained) then we have

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \mathbf{R}f(\mathbf{x}^{(i)}) \quad (6.43)$$

so

$$\mathbf{x}^{(i+1)} - \tilde{\mathbf{x}} = \mathbf{x}^{(i)} - \tilde{\mathbf{x}} - \mathbf{R}f(\tilde{\mathbf{x}} + \Delta_x^{(i)})$$

or

$$\Delta_x^{(i+1)} = \Delta_x^{(i)} - \mathbf{R}f\left(\begin{bmatrix} \tilde{\mathbf{q}} + \Delta_q^{(i)} \\ \tilde{z} + \Delta_z^{(i)} \end{bmatrix}\right) \equiv g\left(\begin{bmatrix} \Delta_q^{(i)} \\ \Delta_z^{(i)} \end{bmatrix}\right) \quad (6.44)$$

or  $g(\Delta_x^{(i)})$  where

$$\Delta_x^{(j)} = \mathbf{x}^{(j)} - \tilde{\mathbf{x}} \quad (j = i, i+1) \quad (6.45)$$

We may write (dropping the superscript<sup>(i)</sup>)

$$\begin{aligned} g(\Delta_x) &= \Delta_x - \mathbf{R}(\mathbf{A} - (\tilde{z} + \Delta_z)\mathbf{I}) \begin{bmatrix} \tilde{\mathbf{q}} + \Delta_q \\ c_n \end{bmatrix} \\ &= \Delta_x - \mathbf{R}[(\mathbf{A} - \tilde{z}\mathbf{I}) \begin{bmatrix} \tilde{\mathbf{q}} \\ c_n \end{bmatrix} - \Delta_z \begin{bmatrix} \tilde{\mathbf{q}} \\ c_n \end{bmatrix} + (\mathbf{A} - \tilde{z}\mathbf{I}) \begin{bmatrix} \Delta_q \\ 0 \end{bmatrix} - \Delta_z \begin{bmatrix} \Delta_q \\ 0 \end{bmatrix}] \quad (6.46) \\ &= \Delta_x - \mathbf{R}(\mathbf{A} - \tilde{z}\mathbf{I}) \begin{bmatrix} \tilde{\mathbf{q}} \\ c_n \end{bmatrix} + \mathbf{R}\Delta_z \begin{bmatrix} \Delta_q \\ 0 \end{bmatrix} - \\ &\quad \mathbf{R} \left[ (\mathbf{A} - \tilde{z}\mathbf{I}) \begin{bmatrix} 1 & 0 & \dots & \dots & \dots \\ 0 & 1 & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix} - \begin{bmatrix} 0 & \dots & 0 & \tilde{q}_0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \tilde{q}_{n-2} \\ 0 & \dots & 0 & c_n \end{bmatrix} \right] \Delta_x \\ &= -\mathbf{R}\mathbf{d} + \mathbf{R}\Delta_z \begin{bmatrix} \Delta_q \\ 0 \end{bmatrix} + (\mathbf{I} - \mathbf{R}\mathbf{J})\Delta_x \quad (6.47) \end{aligned}$$

where

$$\mathbf{d} = (\mathbf{A} - \tilde{z}\mathbf{I}) \begin{bmatrix} \tilde{\mathbf{q}} \\ c_n \end{bmatrix} \quad (6.48)$$

The iterations are repeated until

$$\frac{\|\Delta^{(i+1)}\|_\infty}{\max(\|(q_j^{(i)})_{j=0}^{n-2}\|_\infty, |z^{(i)}|)} \leq \epsilon \quad (e.g. 10^{-7}) \quad (6.49)$$

or the maximum number of iterations is exceeded.

The authors also describe a verification step in which an interval version of Newton's method is used i.e.

$$[\Delta_x]^{(i+1)} = g([\Delta_x]^{(i)}) \quad (6.50)$$

where the  $[\Delta_x]^{(j)}$  are intervals. We obtain interval enclosures for  $[d_i]$ ,  $[w_i]$  by rounding the calculated values up and down as we calculate them. If we replace  $\mathbf{R}$  by an interval matrix  $[\mathbf{R}]$  (by replacing  $w_i$  by  $[w_i]$ ), then  $[\mathbf{R}]$  encloses  $\mathbf{J}^{-1}$  and the term corresponding to  $(\mathbf{I}-\mathbf{R}\mathbf{J})$  in 6.47 drops out, and we obtain

$$g([\Delta_x]) = -[\mathbf{R}][\mathbf{d}] + [\mathbf{R}][\Delta_z] \begin{bmatrix} [\Delta_g] \\ 0 \end{bmatrix} \quad (6.51)$$

Starting from an accurate solution of the non-interval iterations we apply 6.50 with 6.51 until  $[\Delta_x]^{(i+1)}$  is properly included in  $[\Delta_x]^{(i)}$ . Then Schauder's fixed-point theorem tells us that there exists at least one solution of the eigenproblem, i.e.

$$\mathbf{x}^* \in \tilde{\mathbf{x}} + \Delta_x^{(i+1)} \quad (6.52)$$

Other works using the classical companion matrix are referred to in Section 3, under the heading "fast methods of  $O(n^2)$  work".

## 6.2 Other Companion Matrices

have been described by various authors. Some of these can give much more accuracy than the "classical" one described in section 1. For example Schmeisser (1993) modifies Euclid's algorithm to derive a tridiagonal matrix having characteristic polynomial  $p(\lambda)$  (i.e. it is a type of companion matrix). Define  $c(f)$  = leading coefficient of  $f$ , and let

$$f_1(x) = p(x), f_2(x) = \frac{1}{n}p'(x) \quad (6.53)$$

and proceed recursively as follows, for  $i = 1, 2, \dots$ :

If  $f_{i+1}(x) \not\equiv 1$ , then dividing  $f_i$  by  $f_{i+1}$  with remainder  $-r_i$  we have

$$f_i = q_i f_{i+1} - r_i \quad (6.54)$$

and define

$$(i) \text{ if } r_i(x) \not\equiv 0, c_i = c(r_i), f_{i+2}(x) = \frac{r_i(x)}{c_i} \quad (6.55)$$

$$(ii) \text{ if } r_i(x) \equiv 0, c_i = 0, f_{i+2} = \frac{f'_{i+1}}{c(f'_{i+1})} \quad (6.56)$$

If  $f_{i+1}(x) \equiv 1$  terminate the algorithm, defining  $q_i(x) = f_i(x)$ .

The author proves that  $p(x)$  has only real zeros if and only if the modified Euclidean algorithm above yields  $n-1$  nonnegative numbers  $c_1, \dots, c_{n-1}$ , and in this case

$$p(x) = (-1)^n |\mathbf{T} - x\mathbf{I}| \tag{6.57}$$

where

$$\mathbf{T} = \begin{bmatrix} -q_1(0) & \sqrt{c_1} & 0 & \dots & \dots & \dots & 0 \\ \sqrt{c_1} & -q_2(0) & \sqrt{c_2} & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & \sqrt{c_{n-2}} & -q_{n-1}(0) & \sqrt{c_{n-1}} \\ 0 & \dots & \dots & \dots & 0 & \sqrt{c_{n-1}} & -q_n(0) \end{bmatrix} \tag{6.58}$$

Further, the zeros are distinct iff the  $c_i$  are all positive. The eigenvalues of  $\mathbf{T}$  (roots of  $p(x)$ ) may be found by the QR method (see the end of this section).

Schmeisser concentrates on the all-real-root case, but Brugnano and Trigante (1995) apply the Euclidean algorithm to general polynomials. Let  $p_0(x) = p(x)$  and  $p_1(x)$  any other polynomial of degree  $n-1$ . Re-writing 6.54 with slightly different notation we have

$$p_i(x) = q_i(x)p_{i+1}(x) - p_{i+2}(x) \quad (i = 1, 2, \dots) \tag{6.59}$$

terminating when  $i = m - 1$  with  $p_{m+1}(x) = 0$  ( $m$  must be  $\leq n$  since each  $p_{i+1}$  is of lower degree than  $p_i$ ). Then  $p_m(x)$  is the greatest common divisor of  $p_0$  and  $p_1$ , and indeed of  $p_2, \dots, p_{m-1}$  also. Then, if  $p_m(x) \neq \text{const}$ , the functions

$$f_i(x) = \frac{p_i(x)}{p_m(x)} \quad (i = 0, \dots, m) \tag{6.60}$$

are also polynomials. When it happens that

$$\text{deg } p_i(x) = n - i \quad (i = 0, 1, \dots, m = n) \tag{6.61}$$

we say that 6.59 **terminates regularly**. When it does not, i.e. the gcd of  $p_0$  and  $p_1$  is non-trivial, or if  $k \in \{1, \dots, m\}$  exists such that

$$\text{deg } p_i(x) = n - i \quad (i = 0, \dots, k - 1) \tag{6.62}$$

$$\text{deg } p_i(x) < n - i \quad (i \geq k) \tag{6.63}$$

we say that a breakdown occurs at the  $k$ 'th step of 6.59. If 6.61 is satisfied, all the polynomials  $q_i$  are linear, and we may re-write 6.59 as

$$p_i(x) = (x - \alpha_{i+1})p_{i+1}(x) - \beta_{i+1}p_{i+2}(x) \quad (i = 0, \dots, n - 1) \tag{6.64}$$

where  $\beta_i$  is the coefficient of the leading term of  $p_{i+2}(x)$  (so that the latter is now monic). We have also  $p_{n+1}(x) = 0$  and  $p_n(x) \equiv 1$ .

On the other hand, suppose that 6.61 is not satisfied, i.e. 6.59 has a breakdown at the  $k$ 'th step. Thus we obtain only the first  $k$   $p_i$  from 6.64, but we may then switch back to 6.59, at the price of some  $q_i$  being non-linear.

In the regular case one may write 6.64 as

$$x \begin{bmatrix} p_1(x) \\ p_2(x) \\ \dots \\ p_n(x) \end{bmatrix} = \mathbf{T}_n \begin{bmatrix} p_1(x) \\ p_2(x) \\ \dots \\ p_n(x) \end{bmatrix} + \begin{bmatrix} p_0(x) \\ 0 \\ \dots \\ 0 \end{bmatrix} \tag{6.65}$$

where

$$\mathbf{T}_i = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \dots & \dots & 0 \\ 1 & \alpha_2 & \beta_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 & \alpha_i \end{bmatrix} \tag{6.66}$$

from which it follows that the roots of  $p_0(x)$  are the eigenvalues of  $\mathbf{T}_n$  and vice versa. Thus we may transform the polynomial root problem to that of finding eigenvalues of a tridiagonal matrix (in the unlikely event of a breakdown we may repeat the process with a different  $p_1$ ). Again the problem may be solved by the QR method.

However if the roots (or eigenvalues) are multiple, the QR method may be very slow. To avoid this problem we take

$$p_1(x) = \frac{p'_0(x)}{n} \tag{6.67}$$

Then if  $x^*$  is a root of multiplicity  $k$  for  $p_0(x)$ , it is a root of multiplicity  $k-1$  for  $p_m(x)$ , and consequently the roots of

$$f_0(x) = \frac{p_0(x)}{p_m(x)} \tag{6.68}$$

are all simple. If the process 6.59 with 6.67 terminates regularly, it means that  $p_0(x)$  has no multiple roots, since the gcd of  $p_0$  and  $\frac{p'_0}{n}$ , i.e.  $p_n$ , is of degree zero, so  $p_0$  and  $p'_0$  have no common zeros. Hence 6.64 terminates regularly, and  $\mathbf{T}_n$  has simple eigenvalues.

On the other hand, if 6.64 breaks down at the  $r$ 'th step, then there are two cases:

a)  $p_{r+1} \equiv 0$ ; then the zeros of  $p_0(x)$  are given by those of  $f_0(x) = \frac{p_0(x)}{p_r(x)}$  (which will have only simple roots), and also of  $p_r(x)$  (which = the gcd of  $p_0(x), p'_0(x)$ ).  
 b)  $p_{r+1} \not\equiv 0$ ; then we switch to 6.59 and complete the process. Two sub-cases may occur:

(i)  $\text{deg } p_m = 0$ , i.e. the gcd  $p_m$  is a constant, i.e.  $p_0(x)$  and  $p'_0(x)$  have no common factors, i.e. the roots of  $p_0(x)$  are all simple; so we no longer need to use 6.67 and we may choose  $p_1(x)$  randomly and repeat the process 6.64.

(ii)  $\text{deg } p_m > 0$ , then we apply the whole process to  $p_m(x)$  and to  $f_0(x) = \frac{p_0(x)}{p_m(x)}$  (separately).

In the notation of section 3 of Chapter 2,  $p_0$  is represented by  $P_1$ ,  $p_m$  by  $P_2$ , and  $f_0$  by  $Q_1$ . Thus, as stated above,  $f_0$  (or  $Q_1$ ) has only simple roots, and they are the distinct roots of  $p_0$  (or  $P_1$ ). When we repeat the process for  $p_m$  (and later equivalents to  $p_m$ , i.e. the  $P_j$ ) we may compute  $Q_j = \frac{P_j}{P_{j+1}}$  for  $j=1,2,\dots,s$ . Then each  $Q_j$  contains only simple zeros, and they appear in  $p_0$  (or  $P_1$ ) with multiplicity  $\geq j$ . In terms of matrices Brugnano and Trigiante summarize the above as follows: the roots of  $p_0(x)$  are the eigenvalues of a block diagonal matrix

$$\mathbf{T}^* = \begin{bmatrix} \mathbf{T}^{(1)} & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & 0 & \mathbf{T}^{(s)} \end{bmatrix} \tag{6.69}$$

where each block  $\mathbf{T}^{(j)}$  has the form

$$\mathbf{T}^{(j)} = \begin{bmatrix} \alpha_1^{(j)} & \beta_1^{(j)} & 0 & \dots & 0 \\ 1 & \alpha_2^{(j)} & \beta_2^{(j)} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & \alpha_{k_j}^{(j)} \end{bmatrix} \tag{6.70}$$

and has only simple eigenvalues. Here

$$d = k_1 \geq k_2 \geq \dots \geq k_s \geq 1 \tag{6.71}$$

and

$$\sum_{j=1}^s k_j = n \tag{6.72}$$

$s$  = maximum multiplicity of the roots of  $p_n(x)$ , while

$d$  = number of distinct roots. If a root appears in the  $j$ 'th block and not in the  $(j+1)$ 'th, it has exact multiplicity  $j$  and must appear in all the previous blocks.

The authors test their method (incorporated in a Matlab program called **bt-roots**) on several polynomials of moderate degree having roots of multiplicity up to 10. They obtained relative errors at most about  $10^{-14}$  (working in double precision), and perfect multiplicity counts. In contrast the Matlab built-in function

**roots**, which uses the QR method applied to the classical companion matrix, gave errors as high as  $10^{-2}$  in some cases, with no explicit calculation of the multiplicities.

Brugnano (1995) gives a variation on the above in which the Euclidean Algorithm is formulated in terms of vectors and matrices. We take the companion matrix in the form:

$$\mathbf{C} = \begin{bmatrix} -c_{n-1} & -c_{n-2} & \dots & -c_1 & -c_0 \\ 1 & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix} \quad (c_n = 1) \quad (6.73)$$

and let

$$p_1 = \frac{p'(x)}{n} = \sum_{i=1}^n c_i^{(1)} x^{i-1} \quad (6.74)$$

(the normalized derivative of  $p(x)$ ),

$$\mathbf{u}_0 = \mathbf{0}, \mathbf{u}_1 = (1, c_{n-1}^{(1)}, \dots, c_1^{(1)}) \quad (6.75)$$

and compute further  $\mathbf{u}_i$  by

$$\mathbf{C}^T \mathbf{u}_i = \mathbf{u}_{i-1} + \alpha_i \mathbf{u}_i + \beta_i \mathbf{u}_{i+1} \quad (i = 1, \dots, n) \quad (6.76)$$

Here  $\alpha_i$  and  $\beta_i$  have to satisfy

$$\mathbf{e}_i^T \mathbf{u}_{i+1} = 0, \mathbf{e}_{i+1}^T \mathbf{u}_{i+1} = 1 \quad (6.77)$$

where  $\mathbf{e}_i$  is the  $i$ 'th column of  $\mathbf{I}_n$ . These lead to

$$\alpha_i = \mathbf{e}_i^T (\mathbf{C}^T \mathbf{u}_i - \mathbf{u}_{i-1}); \quad (6.78)$$

$$\beta_i = \mathbf{e}_{i+1}^T (\mathbf{C}^T \mathbf{u}_i - \alpha_i \mathbf{u}_i - \mathbf{u}_{i-1}) \quad (6.79)$$

If 6.77 can be satisfied at each step then each  $\mathbf{u}_i$  ( $i=2, \dots, n$ ) has zeros above the  $i$ 'th element, which is 1 (by the second part of 6.77). For example consider the case  $i = 3$ , and suppose the result is true for  $i = 2, 3$  and  $\beta_3 \neq 0$ . Then by re-arranging 6.76 we get

$$\beta_3 \mathbf{u}_4 = \begin{bmatrix} -c_{n-1} & 1 & 0 & \dots & 0 \\ -c_{n-2} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -c_1 & \dots & \dots & 0 & 1 \\ -c_0 & 0 & \dots & \dots & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ \dots \\ \dots \\ \dots \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ \dots \\ \dots \\ \dots \end{bmatrix} - \alpha_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ \dots \\ \dots \end{bmatrix} \quad (6.80)$$

where by 6.78  $\alpha_3$  is the third element of  $\mathbf{C}^T \mathbf{u}_3 - \mathbf{u}_2$ . Hence the third element of  $\mathbf{u}_4$  is zero, and it is obvious that the first and second elements are also 0. Thus the

property is true for  $i = 4$ , and similarly we may prove it in general by induction. It means that the matrix

$$\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n) \quad (6.81)$$

is unit lower triangular.

The first condition in 6.77 can always be satisfied, but the second can only be so if the  $(i+1)$ 'th element of

$$\mathbf{v}_{i+1} = \mathbf{C}^T \mathbf{u}_i - \alpha_i \mathbf{u}_i - \mathbf{u}_{i-1} \quad (6.82)$$

is non-zero (in that case  $\beta_i \neq 0$ ). If this is true as far as  $i = n$  we have

$$\mathbf{v}_{n+1} = \mathbf{C}^T \mathbf{u}_n - \alpha_n \mathbf{u}_n - \mathbf{u}_{n-1} \quad (6.83)$$

where

$$\mathbf{u}_n^T = (0, \dots, 0, 1), \quad \mathbf{u}_{n-1} = (0, \dots, 0, 1, x),$$

$$\alpha_n = \mathbf{e}_n^T \left( \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ x \end{bmatrix} \right) = -x$$

Hence

$$\mathbf{v}_{n+1} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix} + x \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ x \end{bmatrix} = \mathbf{0} \quad (6.84)$$

and so  $\beta_n = 0$ .

Consequently

$$\mathbf{C}^T [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n] = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n] \begin{bmatrix} \alpha_1 & 1 & 0 & \dots & \dots \\ \beta_1 & \alpha_2 & 1 & \dots & \dots \\ 0 & \beta_2 & \alpha_3 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \quad (6.85)$$

or

$$\mathbf{C}^T \mathbf{U} = \mathbf{U} \mathbf{T}_n^T \quad (6.86)$$



where

$$\mathbf{T}_i = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \dots & \dots & 0 \\ 1 & \alpha_2 & \beta_2 & 0 & \dots & 0 \\ 0 & 1 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 & \alpha_i \end{bmatrix} \quad (6.87)$$

From 6.86 we have

$$\mathbf{U}^{-1}\mathbf{C}^T\mathbf{U} = \mathbf{T}_n^T \quad (6.88)$$

i.e.  $\mathbf{T}_n$  is similar to  $\mathbf{C}$ . It can be formed if the second condition in 6.77 is satisfied at each step, in which case we say that the procedure 6.76-6.77 terminates regularly. In that case all the roots of  $p(x)$  are simple, and the QR method applied to  $\mathbf{C}$  or  $\mathbf{T}_n$  gives a good accuracy for all the roots in  $O(n^3)$  flops (later we will see variations which use only  $O(n^2)$  flops). Note that the formation of  $\mathbf{T}_n$  takes only  $O(n^2)$  flops.

If the process 6.76-6.77 does not terminate regularly we say that a **breakdown** has occurred. This may happen in two ways:

(1)  $\mathbf{v}_{i+1} = \mathbf{0}$ ; here we say that the breakdown is **complete**. Then

$$p(x) = d_i(x)p_i(x) \quad (6.89)$$

and  $d_i(x)$  is the characteristic polynomial of  $\mathbf{T}_i$ . Its (all simple) roots are the distinct roots of  $p(x)$ . The same procedure is then applied to  $p_i(x)$ , giving the roots of degree 2 or more, and so on.

(2) Here  $\mathbf{v}_{i+1} \neq \mathbf{0}$  but the  $(i+1)$ -th element does = 0. We call this a partial breakdown. Let  $v_{i+1,i+1+k}$  ( $k \geq 1$ ) be the first non-zero element in  $\mathbf{v}_{i+1}$ , and set  $\beta_i =$  this element. Then define

$$\mathbf{u}_{i+1+k} = \frac{1}{\beta_i}\mathbf{v}_{i+1} \quad (6.90)$$

$$\mathbf{u}_{i+j} = \mathbf{C}^T\mathbf{u}_{i+j+1} \quad (j = k, k-1, \dots, 1) \quad (6.91)$$

and it follows that

$$\mathbf{U}_{i+k+1} = [\mathbf{u}_1, \dots, \mathbf{u}_{i+k+1}] \quad (6.92)$$

is still lower triangular with unit diagonals. Now we can define the following **breakdown step** for 6.76-6.77:

$$\mathbf{C}^T\mathbf{u}_{i+1} = \mathbf{u}_i + \alpha_{i+1}\mathbf{u}_{i+1} + \sum_{j=1}^k \beta_{i+j}\mathbf{u}_{i+j+1} + \beta_{i+k+1}\mathbf{u}_{i+k+2} \quad (6.93)$$

where  $\alpha_{i+1}$  is defined so that

$$\mathbf{e}_{i+1}^T \mathbf{u}_{i+k+2} = 0 \tag{6.94}$$

and the  $\beta_{i+j}$  are defined so that

$$\mathbf{e}_{i+j+1}^T \mathbf{u}_{i+k+2} = 0 \quad (j = 1, \dots, k) \tag{6.95}$$

$$\mathbf{e}_{i+k+2}^T \mathbf{u}_{i+k+2} = 1 \tag{6.96}$$

Then the usual procedure can be restarted with  $\mathbf{u}_{i+k+1}$  and  $\mathbf{u}_{i+k+2}$  in place of  $\mathbf{u}_0$  and  $\mathbf{u}_1$ . If no more breakdowns occur we obtain:

$$\mathbf{C}^T \mathbf{U} = \mathbf{U}(\mathbf{T}'_n)^T \tag{6.97}$$

where  $\mathbf{T}'_n =$

$$\left[ \begin{array}{cccccccccccccccc} \alpha_1 & \beta_1 & 0 & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & 0 \\ 1 & \alpha_2 & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & 0 \\ .. & .. & .. & .. & \beta_{i-1} & .. & .. & .. & .. & .. & .. & .. & .. & 0 \\ .. & .. & .. & 1 & \alpha_i & 0 & .. & .. & 0 & \beta_i & .. & .. & .. & 0 \\ .. & .. & .. & .. & 1 & \alpha_{i+1} & .. & .. & .. & \beta_{i+k} & \beta_{i+k+1} & .. & .. & 0 \\ .. & .. & .. & .. & .. & 1 & 0 & .. & .. & .. & .. & .. & .. & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & 1 & 0 & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & 1 & \alpha_{i+k+2} & \beta_{i+k+2} & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & 1 & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & \beta_{n-1} \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & .. & 1 & \alpha_n & .. \end{array} \right] \tag{6.98}$$

If several partial breakdowns occur, there will be several blocks of the form

$$\left[ \begin{array}{cccc} \alpha_{i+1} & \beta_{i+1} & .. & \beta_{i+k} \\ 1 & 0 & .. & 0 \\ 0 & 1 & .. & 0 \\ .. & .. & .. & .. \\ .. & .. & 0 & 1 \end{array} \right] \tag{6.99}$$

Brugnano next considers the stability of process 6.76 -6.77 and ways of improving it by balancing the matrices  $\mathbf{C}$  and  $\mathbf{T}_n$ . If

$$\mathbf{D} = \left[ \begin{array}{cccc} 1 & 0 & .. & .. \\ 0 & c_{n-1} & 0 & .. \\ .. & .. & .. & .. \\ .. & .. & 0 & c_1 \end{array} \right] \tag{6.100}$$

he shows that

$$\hat{\mathbf{C}} = \mathbf{DCD}^{-1} = \begin{bmatrix} -b_{n-1} & \dots & -b_1 & -b_0 \\ b_{n-1} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \dots & 0 & b_1 & 0 \end{bmatrix} \quad (6.101)$$

(where  $b_i = \frac{c_{i-1}}{c_i}$ ,  $i = 1, \dots, n$ ) has condition number

$$\kappa(\hat{\mathbf{C}}) \leq 4 \frac{\max_i |b_i|}{\min_i |b_i|} \quad (6.102)$$

whereas

$$\kappa(\mathbf{C}) \leq 1 + \max_i |c_i| (1 + |c_0|^{-1} (1 + \max_i |c_i|)) \quad (6.103)$$

If the roots are all near  $\xi$ , where  $\xi \gg 1$ , then

$$\kappa(\mathbf{C}) = 2|\xi|^n, \quad \kappa(\hat{\mathbf{C}}) = 4n^2 \quad (6.104)$$

The second of the above is much smaller than the first e.g. if  $\xi = 10$  and  $n=10$ . He gets similar results for  $|\xi| \ll 1$ , or if some roots are small and others large. In the example  $p(x) = (x+20)^7 + 1$  he finds  $\kappa(\mathbf{C}) \approx 1.7 \times 10^9$ ,  $\kappa(\hat{\mathbf{C}}) \approx 1.4 \times 10^2$ .

Sometimes a complete breakdown may be hard to recognize, because of rounding errors. To ensure proper recognition we may keep track of the

$$\mu_j = \|\mathbf{u}_j\|_\infty \quad (6.105)$$

Then if

$$\mu_1 \dots \mu_i \gg \mu_{i+1} \dots \mu_n \quad (6.106)$$

it means that a complete breakdown has occurred at the  $i$ 'th step.

Several numerical examples gave very accurate roots and multiplicities in most cases, indeed much better accuracy was obtained than by the built-in function **roots** of MAPLE.

Fiedler (1990) has a different approach: he selects distinct numbers  $b_1, \dots, b_n$  Such that  $p(b_i) \neq 0$  for  $i = 1, \dots, n$ . Set

$$v(x) = \prod_{i=1}^n (x - b_i) \quad (6.107)$$

and define the matrix  $\mathbf{A} = [a_{ij}]$  by

$$a_{ij} = -\sigma d_i d_j \text{ if } i \neq j \quad (6.108)$$

$$a_{ii} = b_i - \sigma d_i^2 \text{ if } i = j \tag{6.109}$$

where  $\sigma$  is arbitrary (except non-zero) and

$$\sigma v'(b_i)d_i^2 - p(b_i) = 0 \tag{6.110}$$

Then  $(-1)^n p(x)$  is the characteristic polynomial of the symmetric matrix  $\mathbf{A}$ , and if  $\lambda$  is an eigenvalue of  $\mathbf{A}$ , then

$$\left[ \frac{d_i}{(\lambda - b_i)} \right] \tag{6.111}$$

is the corresponding eigenvector. As a special case, if we have only  $n-1$   $b_i$ ,

$$v(x) = \prod_{i=1}^{n-1} (x - b_i) \tag{6.112}$$

$\mathbf{B}_i = \text{diag}(b_i)$ ,  $\mathbf{g} = [g_i]$  where

$$v'(b_i)g_i^2 + p(b_i) = 0 \text{ (} i = 1, \dots, n - 1 \text{)} \tag{6.113}$$

then

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{g} \\ \mathbf{g}^T & d \end{bmatrix} \text{ (where } d = -c_{n-1} - \sum_{i=1}^{n-1} b_i \text{)} \tag{6.114}$$

has characteristic polynomial  $(-1)^n p(x)$ .

Laszlo (1981) preceded Fiedler with a very similar construction in which  $p(x)$  may be complex and

$$\mathbf{A} = \begin{bmatrix} b_1 & 0 & \dots & 0 & x_1 \\ 0 & b_2 & 0 & \dots & x_2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & b_{n-1} & x_{n-1} \\ y_1 & y_2 & \dots & y_{n-1} & d \end{bmatrix} \tag{6.115}$$

where  $x_i y_i$  takes the place of  $g_i^2$  in 6.113.

Malek and Vaillancourt (1995A) describe an application of Fiedler's method in which the eigenvalues of  $\mathbf{A}$  in 6.114 are estimated by the QR method, with the initial  $b_i$  being eigenvalues of Schmeisser's matrix. A new version of  $\mathbf{A}$  is then constructed using the eigenvalues of the first one for the  $b_i$ , and new eigenvalues calculated. The process is repeated to convergence. As an alternative the initial  $b_i$  may be taken as uniformly distributed on a large circle. Good results were obtained for tests on a number of polynomials of moderate degree, including some of fairly high multiplicity. In some cases extra high precision was needed to give

convergence. In the case of multiple roots, the authors successfully apply the Hull-Mathon procedure described in Chapter 4 section 3.

In a slightly later paper, Malek and Vaillancourt (1995B) point out that the above method requires multiple precision in the case of multiple roots, and they go on to describe a better way of dealing with that case. They find the GCD  $g(x)$  of  $p(x)$  and  $p'(x)$ , by methods described above; then they form

$$q(x) = \frac{p(x)}{g(x)} \quad (6.116)$$

(which has simple roots), and find those roots by iterating Fiedler's method as before. They find the multiplicities by a variation on Schroeder's method, which they ascribe to Lagouanelle. Let  $u(x) = \frac{p(x)}{p'(x)}$ , then the multiplicity is given by

$$m = \frac{1}{u'} \quad (6.117)$$

This gives numerical difficulties when  $x \rightarrow$  a root, as  $p(x)$  and  $p'(x)$  will likely both  $\rightarrow 0$ . So they let

$$v(x) = \frac{p(x)}{g(x)}, w(x) = \frac{p'(x)}{g(x)} \quad (6.118)$$

and prove that

$$m = \text{Lim}_{x \rightarrow \zeta} \frac{w(x)}{v'(x)} \quad (6.119)$$

It is found that the above methods converge most rapidly to large roots, which leads to problems with deflation. So the authors give a method for computing small roots ( $< .01$ ) by finding large roots of the reciprocal equation

$$p^*(x) = x^n p\left(\frac{1}{x}\right) \quad (6.120)$$

In several numerical tests all roots were computed to high accuracy.

In yet a third paper, Malek and Vaillancourt (1995C) compare the use of three companion matrices—the Frobenius, Schmeisser's, and Fiedler's— as starting points for the QR algorithm. For Fiedler's matrix, initial values of  $b_i$  were chosen on a circle of radius 25 centered at the origin, and the number of iterations was set at 5. The reduced polynomial  $q(x)$ , defined above, was used as well as the original  $p(x)$ . For a number of test problems, the Frobenius matrix based on  $p(x)$  gave only a few (or zero) correct digits; when based on  $q(x)$  it was generally good but sometimes very inaccurate. Schmeisser's method was nearly always very accurate whether based on  $p(x)$  or  $q(x)$ ; while Fiedler was moderately good if based on  $p(x)$ , and nearly perfect when based on  $q(x)$ . For condition numbers, Fiedler was easily

the smallest, especially for complex zeros.

Fortune (2002) describes an iterative method which starts by finding the eigenvalues  $s_i$  ( $i = 1, \dots, n$ ) of the Frobenius matrix  $\mathbf{C}$  (by the QR method as usual). These values are used to construct the generalized companion matrix

$$\mathbf{C}(p, \mathbf{s}) = \begin{bmatrix} s_1 & 0 & \dots & \dots & 0 \\ 0 & s_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & s_n \end{bmatrix} - \begin{bmatrix} l_1 & l_2 & \dots & l_n \\ l_1 & l_2 & \dots & l_n \\ \dots & \dots & \dots & \dots \\ l_1 & l_2 & \dots & l_n \end{bmatrix} \quad (6.121)$$

where the Lagrange coefficients  $l_i$  are given by

$$l_i = \left| \frac{p(s_i)}{\prod_{j=1, j \neq i}^n (s_i - s_j)} \right| \quad (6.122)$$

(he proves that 6.121 is in fact a companion matrix). The eigenvalues of  $\mathbf{C}(p, \mathbf{s})$  are computed and used to construct a new  $\mathbf{C}(p, \mathbf{s})$  –the process is repeated till convergence. Fortune describes the implementation (including convergence criteria) in great detail– see the cited paper. In numerical tests, including high degree polynomials, Fortune’s program **eigensolve** was usually much faster than Bini and Fiorentino’s **mpsolve** based on Aberth’s method (see Chapter 4 Section 13). Exceptions include sparse polynomials such as  $x^n - 1$  for large  $n$ .

Many authors who solve the polynomial root problem by finding the eigenvalues of some companion matrix perform the latter task by using the QR method. This was originally invented by Francis (1961), and has been much improved since then. It is described by many authors, such as Ralston and Rabinowitz (1978). We will give a brief description, based on the last-mentioned book.

It is found beneficial to first reduce the matrix by a similarity transform to Hessenberg or tridiagonal form. That is, we find an orthogonal matrix  $\mathbf{P}$  such that

$$\mathbf{P}^* \mathbf{A} \mathbf{P} = \mathbf{H} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & \dots & a_{2n} \\ 0 & a_{32} & a_{33} & \dots & \dots & a_{3n} \\ 0 & 0 & a_{43} & a_{44} & \dots & a_{4n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix} \quad (6.123)$$

$\mathbf{H}$  is known as a Hessenberg matrix. If  $\mathbf{A}$  is symmetric, we may use Givens’ or Householder’s method (see Ralston and Rabinowitz), and  $\mathbf{H}$  is tridiagonal. If  $\mathbf{A}$  is non-symmetric, we may also use those methods or Gaussian elimination. The Frobenius companion matrix is automatically in Hessenberg form, so none of these methods is necessary. Likewise, Schmeisser’s method gives a symmetric tridiagonal matrix, while Brugnano and Trigiante derive a general non-symmetric tridiagonal

matrix. The QR method is used with all of these matrices.

The QR method consists in factorizing

$$\mathbf{A}_i - p_i \mathbf{I} = \mathbf{Q}_i \mathbf{R}_i \quad (6.124)$$

(where  $\mathbf{Q}_i$  is orthogonal and  $\mathbf{R}_i$  is right-triangular) and then reversing the order of the factors to give:

$$\mathbf{A}_{i+1} = \mathbf{R}_i \mathbf{Q}_i + p_i \mathbf{I} \quad (6.125)$$

The shifts  $p_i$  may be chosen in various ways so as to accelerate convergence. It can be proved that nearly always:

$$\mathbf{A}_{i+1} \rightarrow \begin{bmatrix} \lambda_1 & x & x & \dots & \dots & \dots & \dots & x \\ 0 & \lambda_2 & x & \dots & \dots & \dots & \dots & x \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & x \\ 0 & 0 & \dots & 0 & \lambda_m & x & \dots & x \\ 0 & 0 & \dots & \dots & 0 & \mathbf{B}_1 & \dots & x \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & x \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & \mathbf{B}_l \end{bmatrix} \quad (6.126)$$

where  $\lambda_j$  are the real eigenvalues and the  $\mathbf{B}_j$  are  $2 \times 2$  real submatrices whose complex conjugate eigenvalues are eigenvalues of  $\mathbf{A}_1 = \mathbf{A}$ . The work involved in a **single** transformation of type 6.124-6.125 is  $O(n)$  for tridiagonal matrices and  $O(n^2)$  for Hessenberg ones. As several iterations are required per eigenvalue, the total work is  $O(n^2)$  or  $O(n^3)$  respectively. Unfortunately, as we shall see shortly, the QR iterations for a **non**-symmetric tridiagonal matrix eventually cause the previously zero elements in the upper right triangle to become non-zero, leading to a  $O(n^3)$  algorithm. On the other hand a **symmetric** tridiagonal matrix remains in that form under QR, so in that case the algorithm is  $O(n^2)$ .

The factorization in 6.124 is accomplished by premultiplying  $\mathbf{A}_i - p_i \mathbf{I}$  by a series of rotation matrices (also called Givens matrices)

$$S_{j,j+1}^T = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & c_j & -s_j & 0 & \dots & 0 \\ 0 & \dots & 0 & s_j & c_j & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{bmatrix} \quad (6.127)$$

where  $c_j = \cos(\theta_j)$ ,  $s_j = \sin(\theta_j)$ , and  $\theta_j$  is chosen so that after the multiplication the new  $(j+1, j)$  element is 0. This requires

$$s_j a_{jj} + c_j a_{j+1,j} = 0 \quad (6.128)$$

$$\text{i.e. } s_j = -a_{j+1,j}\alpha_j, c_j = a_{jj}\alpha_j \quad (6.129)$$

where

$$\alpha_j = \frac{1}{\sqrt{a_{j+1,j}^2 + a_{jj}^2}} \quad (6.130)$$

The  $a_{kj}$  above have been modified by subtraction of  $p_i$  from the diagonal terms, and by previous multiplications by  $S_{12}^T, S_{23}^T$  etc. The above is repeated for  $j=1,2,\dots,n-1$ , finally giving the desired right-triangular matrix  $\mathbf{R}_i$ . Thus if we set

$$\mathbf{Q}_i^T = S_{n-1,n}^T \dots S_{23}^T S_{12}^T \quad (6.131)$$

we have

$$\mathbf{Q}_i^T (\mathbf{A}_i - p_i \mathbf{I}) = \mathbf{R}_i \quad (6.132)$$

$$\text{or } \mathbf{A}_i - p_i \mathbf{I} = \mathbf{Q}_i \mathbf{R}_i \quad (6.133)$$

Next the reverse product  $\mathbf{R}_i \mathbf{Q}_i$  is formed by postmultiplying  $\mathbf{R}_i$  by  $S_{12}, S_{23}, \dots, S_{n-1,n}$  in turn (and  $p_i \mathbf{I}$  added) to give  $\mathbf{A}_{i+1}$ .

Let us consider for example the case of a  $3 \times 3$  tridiagonal matrix

$$\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \quad (6.134)$$

It may or may not be symmetric. We will assume that the shift has already been applied to the diagonal elements. We will use:

$$S_{12} = \begin{bmatrix} c_1 & s_1 & 0 \\ -s_1 & c_1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, S_{12}^T = \begin{bmatrix} c_1 & -s_1 & 0 \\ s_1 & c_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (6.135)$$

so that

$$S_{12}^T \mathbf{A} = \begin{bmatrix} c_1 & -s_1 & 0 \\ s_1 & c_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \quad (6.136)$$

$$= \begin{bmatrix} c_1 a_{11} - s_1 a_{21} & c_1 a_{12} - s_1 a_{22} & -s_1 a_{23} \\ s_1 a_{11} + c_1 a_{21} & s_1 a_{12} + c_1 a_{22} & c_1 a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \quad (6.137)$$

where

$$s_1 a_{11} + c_1 a_{21} = 0, \text{ i.e. } s_1 = -a_{21}\alpha_1, c_1 = a_{11}\alpha_1 \quad (6.138)$$



$$\alpha_1 = \frac{1}{\sqrt{a_{21}^2 + a_{11}^2}} \quad (6.139)$$

and element in (1,1) position =

$$sq_1 = \sqrt{a_{21}^2 + a_{11}^2} \quad (6.140)$$

Premultiplying the above by  $S_{23}^T$  gives

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & c_2 & -s_2 \\ 0 & s_2 & c_2 \end{bmatrix} \begin{bmatrix} sq_1 & c_1 a_{12} - s_1 a_{22} & -s_1 a_{23} \\ 0 & B & c_1 a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \quad (6.141)$$

where

$$B = (-a_{21}a_{12} + a_{11}a_{22})\alpha_1 \quad (6.142)$$

The above matrix product =

$$\begin{bmatrix} sq_1 & c_1 a_{12} - s_1 a_{22} & -s_1 a_{23} \\ 0 & c_2 B - s_2 a_{32} & c_2 c_1 a_{23} - s_2 a_{33} \\ 0 & s_2 B + c_2 a_{32} & s_2 c_1 a_{23} + c_2 a_{33} \end{bmatrix} \quad (6.143)$$

where

$$s_2 B + c_2 a_{32} = 0 \quad (6.144)$$

i.e.

$$s_2 = -a_{32}\alpha_2, \quad c_2 = B\alpha_2 \quad (6.145)$$

where

$$\alpha_2 = \frac{1}{\sqrt{B^2 + a_{32}^2}} \quad (6.146)$$

and the (2,2) element =

$$sq_2 = \sqrt{B^2 + a_{32}^2} \quad (6.147)$$

Thus the above matrix, which is  $\mathbf{R}_1$ , may be written

$$\begin{bmatrix} sq_1 & c_1 a_{12} - s_1 a_{22} & -s_1 a_{23} \\ 0 & sq_2 & c_2 c_1 a_{23} - s_2 a_{33} \\ 0 & 0 & s_2 c_1 a_{23} + c_2 a_{33} \end{bmatrix} \quad (6.148)$$

Postmultiplying the above by  $S_{12} = \begin{bmatrix} c_1 & s_1 & 0 \\ -s_1 & c_1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$  gives

$$\begin{bmatrix} c_1 sq_1 - s_1(c_1 a_{12} - s_1 a_{22}) & s_1 sq_1 + c_1(c_1 a_{12} - s_1 a_{22}) & -s_1 a_{23} \\ -s_1 sq_2 & c_1 sq_2 & c_2 c_1 a_{23} - s_2 a_{33} \\ 0 & 0 & s_2 c_1 a_{23} + c_2 a_{33} \end{bmatrix} \quad (6.149)$$

Postmultiplying by  $S_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_2 & s_2 \\ 0 & -s_2 & c_2 \end{bmatrix}$  gives  $\mathbf{R}_1 S_{12} S_{23} = \mathbf{R}_1 \mathbf{Q}_1 =$

$$\begin{bmatrix} c_1 s q_1 - s_1(c_1 a_{12} - s_1 a_{22}) & c_2[s_1 s q_1 + c_1(c_1 a_{12} - s_1 a_{22})] + s_1 s_2 a_{23} & E \\ x & x & x \\ 0 & x & x \end{bmatrix} \quad (6.150)$$

where x represents an unspecified non-zero element whose exact value is not important in the present context (which is to determine under what conditions E is 0). In fact

$$\begin{aligned} E &= s_2[s_1 s q_1 + c_1^2 a_{12} - c_1 s_1 a_{22}] - c_2 s_1 a_{23} \quad (6.151) \\ &= (-a_{32} \alpha_2)[-a_{21} \alpha_1 \sqrt{a_{21}^2 + a_{11}^2} + a_{11}^2 a_{12} \alpha_1^2 + \\ & a_{11} a_{21} a_{22} \alpha_1^2] - B \alpha_2 (-a_{21} \alpha_1) a_{23} = \\ & -a_{32} \alpha_2 [-a_{21} + \frac{a_{11}^2 a_{12}}{a_{21}^2 + a_{11}^2} + \frac{a_{11} a_{21} a_{22}}{a_{21}^2 + a_{11}^2}] + \\ & \alpha_1 (a_{11} a_{22} - a_{21} a_{12}) a_{21} a_{23} \alpha_2 \alpha_1 \\ & = -\frac{\alpha_2}{a_{21}^2 + a_{11}^2} [a_{32} \{a_{11}^2 (a_{12} - a_{21}) + a_{21} (a_{11} a_{22} - a_{21}^2)\} \\ & - a_{23} a_{21} (a_{11} a_{22} - a_{21} a_{12})] \quad (6.152) \end{aligned}$$

= 0 if  $a_{12} = a_{21}$  and  $a_{23} = a_{32}$  (i.e. if  $\mathbf{A}$  is symmetric). In general, if  $\mathbf{A}$  is non-symmetric, E will be non-zero, i.e. a new non-zero has been introduced in the upper-right part of  $\mathbf{A}_2$ . Next we will consider a  $4 \times 4$  matrix, with a view to showing how the right upper triangle fills in with non-zeros, even for a matrix which is tridiagonal initially. We will indicate by a "x" those elements which are non-zero after each multiplication by a rotation matrix, and by (0) those elements which are set to 0 by the most recent multiplication. We have:

$$S_{12}^T \mathbf{A} = \begin{bmatrix} c_1 & -s_1 & 0 & 0 \\ s_1 & c_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x & x & 0 & 0 \\ x & x & x & 0 \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} = \begin{bmatrix} x & x & x & 0 \\ (0) & x & x & 0 \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} \quad (6.153)$$

Now premultiplying by  $S_{23}^T$  gives:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_2 & -s_2 & 0 \\ 0 & s_2 & c_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x & x & x & 0 \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} = \begin{bmatrix} x & x & x & 0 \\ 0 & x & x & x \\ 0 & (0) & x & x \\ 0 & 0 & x & x \end{bmatrix} \quad (6.154)$$

Similarly premultiplying by  $S_{34}^T$  zeros out the (4,3) element, and then postmultiplying by  $S_{12}$  gives:

$$\begin{bmatrix} x & x & x & 0 \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & (0) & x \end{bmatrix} \begin{bmatrix} c_1 & s_1 & 0 & 0 \\ -s_1 & c_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} x & x & x & 0 \\ x & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \quad (6.155)$$

Next postmultiplying by  $S_{23}$  gives:

$$\begin{bmatrix} x & x & x & 0 \\ x & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_2 & s_2 & 0 \\ 0 & -s_2 & c_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} x & x & x & 0 \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \quad (6.156)$$

and then postmultiplying by  $S_{34}$  we have:

$$\begin{bmatrix} x & x & x & 0 \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_3 & s_3 \\ 0 & 0 & -s_3 & c_3 \end{bmatrix} = \begin{bmatrix} x & x & x & 0 \\ x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \end{bmatrix} \quad (6.157)$$

The above matrix is what we have called  $\mathbf{A}_2$ . The reader may like to verify that another QR iteration to give  $\mathbf{A}_3$  will fill in the (1,4) element, and for the general case of order  $n$ ,  $n-1$  iterations will fill in the whole of the upper right triangle, so that later iterations will take  $O(n^2)$  flops, and the whole process will thus take  $O(n^3)$  flops. On the other hand if  $\mathbf{A}_1$  is symmetric, then  $\mathbf{A}_2 = \mathbf{R}_1\mathbf{Q}_1 = \mathbf{Q}_1^T\mathbf{A}_1\mathbf{Q}_1$ , which is also symmetric, and so on for all  $\mathbf{A}_i$ . But since the lower triangle below the first sub-diagonal does not fill in, neither does the upper triangle.

Goedecker (1994) has compared a companion matrix eigenvalue subroutine (using the QR method) with two well-known “conventional” methods (i.e. not using matrices). They are the IMSL rootfinder ZPORC (based on the Jenkins-Traub method), and the NAG rootfinder C02AGF (based on Laguerre’s method). He ran tests involving several types of polynomials on two serial machines and a vector machine. He reached the following conclusions regarding several important aspects of root-finding:

- 1) **Reliability.** For high degree polynomials overflow occurs frequently in the conventional rootfinders, but never in the matrix-based routine.
- 2) **Accuracy.** For low order, all three routines gave high accuracy. For higher-order polynomials with simple roots, the conventional methods rapidly lost accuracy, whereas QR still gave good accuracy. For multiple roots the QR method gave the best performance.
- 3) **Speed.** For low degree, the QR method (although of order  $n^3$  versus order  $n^2$  for the others) was fastest. Even for high degree, the QR method is not much slower than the others, and in fact is faster on the vector machine for “reasonable” values of  $n$ .

### 6.3 Methods with $O(N^2)$ Operations

Several authors in the late 20th or early 21st century have described fast methods (i.e. faster than those of sections 1-2) using  $O(n^2)$  operations to find all the roots. We will start with Uhlig (1999). He applies the Euclidean algorithm, in a similar manner to Bruignano and Trigiante, to find a (generally unsymmetric) tridiagonal matrix whose characteristic polynomial is  $p(x)$ . He finds that the QR method behaves badly for unsymmetric tridiagonal matrices, so he performs a similarity transformation which converts the tridiagonal generalized companion matrix to a complex symmetric one, as follows: Let

$$\tilde{\mathbf{T}} = \mathbf{D}\mathbf{T}\mathbf{D}^{-1} \quad (6.158)$$

where

$$\mathbf{T} = \begin{bmatrix} a_1 & b_1 & 0 & \dots & 0 \\ c_1 & a_2 & b_2 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & c_{n-2} & a_{n-1} & b_{n-1} \\ 0 & \dots & 0 & c_{n-1} & a_n \end{bmatrix} \quad (6.159)$$

and

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & d_n \end{bmatrix} \quad (6.160)$$

Then

$$\tilde{\mathbf{T}} = \begin{bmatrix} d_1 a_1 d_1^{-1} & d_1 b_1 d_2^{-1} & 0 & \dots \\ d_2 c_1 d_1^{-1} & d_2 a_2 d_2^{-1} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & d_n a_n d_n^{-1} \end{bmatrix} \quad (6.161)$$

We require this to be symmetric, i.e.

$$\frac{d_2 c_1}{d_1} = \frac{d_1 b_1}{d_2}, \text{ i.e. } \frac{d_2^2}{d_1^2} = \frac{b_1}{c_1} \quad (6.162)$$

So  $d_1$  is arbitrary and

$$d_2 = d_1 \sqrt{\frac{b_1}{c_1}} \quad (6.163)$$

and in general

$$d_{i+1} = d_i \sqrt{\frac{b_i}{c_i}} \quad (i = 1, \dots, n-1) \quad (6.164)$$

We see that if  $b_i c_i < 0$  then  $d_i$  will be complex. 6.161 gives

$$\tilde{\mathbf{T}} = \begin{bmatrix} a_1 & \sqrt{b_1 c_1} & 0 & \dots & \dots \\ \sqrt{b_1 c_1} & a_2 & \sqrt{b_2 c_2} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & a_{n-1} & \sqrt{b_{n-1} c_{n-1}} \\ 0 & \dots & 0 & \sqrt{b_{n-1} c_{n-1}} & a_n \end{bmatrix} = \begin{bmatrix} a_1 & \beta_1 & 0 & \dots & \dots \\ \beta_1 & a_1 & \beta_2 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \beta_{n-2} & a_{n-1} & \beta_{n-1} \\ 0 & \dots & 0 & \beta_{n-1} & a_n \end{bmatrix} \tag{6.165}$$

with  $\beta_i = \sqrt{b_i c_i}$ . As pointed out in section 2, eigenvalues of  $\tilde{\mathbf{T}}$  can be obtained in  $O(n^2)$  operations.

The tridiagonalization process yields a block diagonal matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & 0 & \dots & \dots \\ 0 & \mathbf{T}_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \mathbf{T}_k \end{bmatrix} \tag{6.166}$$

If  $k > 1$  we can compute several approximations for any multiple root from several blocks, and averaging them gives better accuracy due to smoothing of random errors. Uhlig explains that his version of Euclid’s algorithm for tridiagonal matrices takes  $O(n^2)$  operations.

The polynomial  $p(x)$  is scaled to give

$$\tilde{p}(x) = x^n + \frac{c_{n-1}}{c} x^{n-1} + \frac{c_{n-2}}{c^2} x^{n-2} + \dots + \frac{c_0}{c^n} \tag{6.167}$$

which has roots  $\frac{\zeta_i}{c}$ . Uhlig computes the optimal power of 2 for  $c$  that nearly equalizes the coefficients of  $\tilde{p}(x)$ , and re-scales the computed roots at the end. He finds better precision and multiplicity estimates with this technique.

We will use standard Givens rotations (possibly complex), but observe that

$$\cos(\theta_k) = \frac{a_k}{\sqrt{a_k^2 + \beta_k^2}} \tag{6.168}$$

and

$$\sin(\theta_k) = \frac{\beta_k}{\sqrt{a_k^2 + \beta_k^2}} \tag{6.169}$$

can be very large if

$$a_k \approx i\beta_k \tag{6.170}$$

This effect can be monitored and bounded, as will be explained later. Uhlig treats the QR iteration a little differently from our treatment in section 2; he writes:

$$\hat{\mathbf{T}} = (\mathbf{G}_{m-1} \dots \mathbf{G}_1) \tilde{\mathbf{T}} (\mathbf{G}_1^{-1} \dots \mathbf{G}_{m-1}^{-1}) \tag{6.171}$$

where the  $\mathbf{G}_j$  are ‘‘Givens rotations’’, the same as  $S_{j,j+1}$  of 6.127 but with a different notation (in line with the different sources) and  $\tilde{\mathbf{T}}$  is given by 6.165. This may also be written

$$\mathbf{G}_{m-1} (\dots (\mathbf{G}_2 (\mathbf{G}_1 \tilde{\mathbf{T}} \mathbf{G}_1^{-1}) \mathbf{G}_2^{-1}) \dots) \mathbf{G}_{m-1}^{-1} \tag{6.172}$$

Thus he forms in turn  $(\mathbf{G}_1 \tilde{\mathbf{T}} \mathbf{G}_1^{-1})$ ,  $\mathbf{G}_2 (\mathbf{G}_1 \tilde{\mathbf{T}} \mathbf{G}_1^{-1}) \mathbf{G}_2^{-1}$ , etc. So we have at the first step (with

$$\mathbf{T}_0 = \tilde{\mathbf{T}}, \quad -s_1 a_1 + c_1 \beta_1 = 0, \quad s_1^2 + c_1^2 = 1) \tag{6.173}$$

$$\mathbf{G}_1 \mathbf{T}_0 \mathbf{G}_1^{-1} = \begin{bmatrix} c_1 & s_1 & 0 & \dots & \dots & \dots \\ -s_1 & c_1 & 0 & \dots & \dots & \dots \\ \dots & 0 & 1 & 0 & \dots & \dots \\ \dots & \dots & 0 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 & \beta_1 & 0 & \dots & \dots \\ \beta_1 & a_2 & \beta_2 & 0 & \dots \\ 0 & \beta_2 & a_3 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \tag{6.174}$$

$$\begin{bmatrix} c_1 & -s_1 & 0 & \dots & \dots & \dots \\ s_1 & c_1 & 0 & \dots & \dots & \dots \\ \dots & 0 & 1 & 0 & \dots & \dots \\ \dots & \dots & 0 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 & 1 \end{bmatrix} \tag{6.174}$$

$$= \begin{bmatrix} \hat{a}_1 & -s_1^2 \beta_1 + s_1 c_1 a_2 & s_1 \beta_2 & 0 & \dots \\ -s_1^2 \beta_1 + s_1 c_1 a_2 & -c_1 s_1 \beta_1 + c_1^2 a_2 & c_1 \beta_2 & \dots & \dots \\ s_1 \beta_2 & c_1 \beta_2 & a_3 & \dots & \dots \\ 0 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \tag{6.175}$$

$$= \begin{bmatrix} \hat{a}_1 & s_1 P_2 & s_1 \beta_2 & 0 & \dots \\ s_1 P_2 & c_1 P_2 (= \gamma_2) & c_1 \beta_2 & 0 & \dots \\ s_1 \beta_2 & c_1 \beta_2 & a_3 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} = \mathbf{T}_1 \text{ (say)} \tag{6.176}$$

where

$$P_2 = -s_1 \beta_1 + c_1 a_2, \quad \gamma_2 = c_1 P_2, \quad \hat{a}_1 = a_1 + a_2 - \gamma_2 \tag{6.177}$$

(the last due to preservation of trace under similarity). Now at the next step when we premultiply by  $\mathbf{G}_2$  we need to eliminate the (3,2) element of  $\mathbf{T}_1$ , so we must have

$$-s_2(c_1P_2) + c_2(c_1\beta_2) = 0 \quad (6.178)$$

i.e.  $\frac{s_2}{c_2} = \frac{c_1\beta_2}{c_1P_2}$  and hence

$$-s_2(s_1P_2) + c_2(s_1\beta_2) = 0 \quad (6.179)$$

i.e. the (3,1) element is also zeroed out. Postmultiplying by  $\mathbf{G}_2^{-1}$  makes the (3,2) element non-zero again, but the (3,1) element (and by symmetry the (1,3) element) remains 0. In general we apply the transformation

$$\mathbf{T}_i = \mathbf{G}_i \mathbf{T}_{i-1} \mathbf{G}_i^{-1} \quad (6.180)$$

where

$$\mathbf{T}_{i-1} = \begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \hat{a}_{i-1} & s_{i-1}P_i & s_{i-1}\beta_i & \dots & \dots \\ \dots & s_{i-1}P_i & \gamma_i (= c_{i-1}P_i) & c_{i-1}\beta_i & \dots & \dots \\ \dots & s_{i-1}\beta_i & c_{i-1}\beta_i & a_{i+1} & \beta_{i+1} & \dots \\ \dots & 0 & \dots & \beta_{i+1} & a_{i+2} & \dots \end{bmatrix} \quad (6.181)$$

so as to zero out the (i+1,i) element in  $\mathbf{G}_i \mathbf{T}_{i-1}$ , i.e.

$$-s_i(c_{i-1}P_i) + c_i(c_{i-1}\beta_i) = 0, \text{ i.e. } \frac{s_i}{c_i} = \frac{\beta_i}{P_i} \quad (6.182)$$

$$s_i = \frac{\beta_i}{\sqrt{P_i^2 + \beta_i^2}}, \quad c_i = \frac{P_i}{\sqrt{P_i^2 + \beta_i^2}} \quad (6.183)$$

Thus we obtain

$$\mathbf{T}_i = \begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \hat{a}_{i-1} & \hat{\beta}_{i-1} & 0 & \dots & \dots \\ \dots & \hat{\beta}_{i-1} & \hat{a}_i & s_iP_{i+1} & s_i\beta_{i+1} & \dots \\ \dots & 0 & s_iP_{i+1} & c_iP_{i+1} (= \gamma_{i+1}) & c_i\beta_{i+1} & \dots \\ \dots & 0 & s_i\beta_{i+1} & c_i\beta_{i+1} & a_{i+2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \quad (6.184)$$

with

$$P_{i+1} = -s_i c_{i-1} \beta_i + c_i a_{i+1}, \quad \hat{\beta}_{i-1} = s_{i-1} \sqrt{P_i^2 + \beta_i^2} \quad (6.185)$$

$$\gamma_{i+1} = c_i P_{i+1} \text{ and } \hat{a}_i = \gamma_i + a_{i+1} - \gamma_{i+1} \quad (6.186)$$

Now we are really only interested in changes affecting the diagonal entries  $\hat{a}_i$  and the products  $B_i = \hat{\beta}_i^2$  of off-diagonal elements, so we may avoid the time-consuming operations of taking square-roots by setting

$$R = P_i^2 + \beta_i^2, \quad c_i^2 = \frac{P_i^2}{R}, \quad s_i^2 = \frac{\beta_i^2}{R} \quad (6.187)$$

$$B_{i-1}(= \hat{\beta}_{i-1}^2) = s_{i-1}^2 R, \quad \gamma_{i+1} = c_i^2 a_{i+1} - s_i^2 \gamma_i \quad (6.188)$$

$$P_{i+1}^2 = \frac{\gamma_{i+1}^2}{c_i^2} \text{ (if } c_i \neq 0), = c_{i-1}^2 \beta_i^2 \text{ (if } c_i = 0) \quad (6.189)$$

$$\hat{a}_i = \gamma_i + a_{i+1} - \gamma_{i+1} \quad (6.190)$$

6.188 can be derived as follows:  $\gamma_{i+1} = c_i P_{i+1} = c_i(-s_i c_{i-1} \beta_i + c_i a_{i+1})$  (by 6.185)

$$= c_i^2 a_{i+1} - s_i c_i c_{i-1} \beta_i = c_i^2 a_{i+1} - s_i c_{i-1} (s_i P_i) \text{ (by 6.182)}$$

$$= c_i^2 a_{i+1} - s_i^2 \gamma_i \text{ (by definition of } \gamma_i).$$

The equations 6.187-6.190 may be applied recursively for  $i=2, \dots, m-1$  to give a complete (modified) QR transformation in about  $12(m-1)$  adds, multiplies and divides (and no square roots) for an  $m \times m$  matrix. This is about half the count for a standard QR application using complex Givens transformations. Since  $\approx 3m$  QR transformations are required to get all the eigenvalues, we have  $O(m^2)$  operations for that task.

We have previously referred to the possibility of large complex Givens matrix coefficients if 6.170 applies. Uhlig (1997) in his section 7 shows that if  $|\cos(\theta_k)|$  and  $|\sin(\theta_k)|$  are bounded by 100, the coefficients will return to “moderate” levels at the next iteration, and eigenvalues are preserved to many correct digits. If  $|\cos(\theta_k)|$  and  $|\sin(\theta_k)|$  become  $> 100$ , Uhlig applies “exceptional shifts” which avoid the problem. He does not explain in detail how this is done.

In numerical tests the program **pzero** constructed by Uhlig usually performed well, with comparable or better accuracy than other programs such as **btr** of Brugnano’s. More importantly, it ran about  $\frac{2}{2}$  times faster than the QR method applied to the companion matrix for  $p(x)$ , thus verifying that it is of order  $O(n^2)$ .

Bini, Daddi, and Gemignani (2004) apply the QR method to a modified companion matrix in  $O(n^2)$  operations. They define the companion matrix in the form:

$$\mathbf{F} = \begin{bmatrix} 0 & \dots & \dots & 0 & f_1 \\ 1 & 0 & \dots & 0 & f_2 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & 0 & f_{n-1} \\ 0 & \dots & 0 & 1 & f_n \end{bmatrix} \quad (6.191)$$



where  $f_i = -c_{i-1}$ .

They apply the shifted QR iteration

$$\mathbf{A}_k - \alpha_k \mathbf{I} = \mathbf{Q}_k \mathbf{R}_k \quad (6.192)$$

$$\mathbf{A}_{k+1} = \mathbf{R}_k \mathbf{Q}_k + \alpha_k \mathbf{I} = \mathbf{Q}_k^H \mathbf{A}_k \mathbf{Q}_k \quad (6.193)$$

(with  $\mathbf{A}_0 = \mathbf{F}$  and  $\mathbf{M}^H =$  conjugate transpose to matrix  $\mathbf{M}$ ) which is equivalent to

$$\mathbf{A}_k = \mathbf{P}_k^H \mathbf{F} \mathbf{P}_k \quad (6.194)$$

(with  $\mathbf{P}_k = \mathbf{Q}_0 \dots \mathbf{Q}_{k-1}$ ) and all the  $\mathbf{A}$ , like  $\mathbf{F}$ , are in upper Hessenberg form. We can show that

$$\mathbf{F}^{-1} = \begin{bmatrix} -\frac{f_2}{f_1} & 1 & 0 & \dots & 0 \\ -\frac{f_3}{f_1} & 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ -\frac{f_n}{f_1} & 0 & \dots & 0 & 1 \\ \frac{1}{f_1} & 0 & \dots & 0 & 0 \end{bmatrix} \quad (6.195)$$

(for multiplying by  $\mathbf{F}$  gives  $\mathbf{I}$ ) and we may also verify that

$$\mathbf{F} = \mathbf{F}^{-H} + \mathbf{U} \mathbf{V}^H \quad (6.196)$$

where

$$\mathbf{U} = \begin{bmatrix} 1 & f_1 \\ 0 & f_2 \\ \dots & \dots \\ 0 & f_n \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \frac{f_2}{f_1} & 0 \\ \frac{f_3}{f_1} & 0 \\ \dots & \dots \\ \frac{f_n}{f_1} & 0 \\ \frac{1}{f_1} & 1 \end{bmatrix} \quad (6.197)$$

Combining 6.196 and 6.194 gives

$$\mathbf{A}_k = \mathbf{A}_k^{-H} + \mathbf{U}_k \mathbf{V}_k^H \quad (k = 0, 1, \dots) \quad (6.198)$$

where

$$\mathbf{U}_k = \mathbf{Q}_{k-1}^H \mathbf{U}_{k-1}, \quad \mathbf{V}_k = \mathbf{Q}_{k-1}^H \mathbf{V}_{k-1} \quad (6.199)$$

(Bini et al have  $\mathbf{Q}_k^H$  in the above, but we think that is a misprint).

The authors show that the entries in the upper right part of  $\mathbf{A}_k^{-H}$  are given by

$$(\mathbf{A}_k^{-H})_{i,j} = \frac{1}{y_n^{(k)}} x_i^{(k)} y_j^{(k)} \quad (i \leq j) \quad (6.200)$$

where

$$\mathbf{x}^{(k)} = (x_i^{(k)}) = \mathbf{A}_k^{-H} \mathbf{e}_n \quad (6.201)$$

is the last column of  $\mathbf{A}_k^{-H}$  and

$$\mathbf{y}^{(k)T} = (y_i^{(k)}) = \mathbf{e}_1^T \mathbf{A}_k^{-H} \tag{6.202}$$

is the first row of  $\mathbf{A}_k^{-H}$ . It follows that  $\mathbf{A}_k = (a_{ij}^{(k)})$  is given by

$$a_{ij}^{(k)} = \begin{cases} 0 & \text{for } i > j + 1 \\ \frac{1}{y_n^{(k)}} x_i^{(k)} y_j^{(k)} + u_{i1}^{(k)} \bar{v}_{j1}^{(k)} + u_{i2}^{(k)} \bar{v}_{j2}^{(k)} & \text{for } i \leq j \end{cases} \tag{6.203}$$

In fact the  $\mathbf{A}_k$  can be determined by  $7n-1$  parameters, namely the subdiagonals  $(b_1^{(k)}, \dots, b_{n-1}^{(k)})$ ,  $\mathbf{x}^{(k)}$ ,  $\mathbf{y}^{(k)}$ , columns  $\mathbf{u}_1^{(k)}$ ,  $\mathbf{u}_2^{(k)}$  of  $\mathbf{U}$ , and columns  $\mathbf{v}_1^{(k)}$  and  $\mathbf{v}_2^{(k)}$  of  $\mathbf{V}$ . Let us denote in particular  $\mathbf{a}^{(k)}$  = vector of diagonals of  $\mathbf{A}_k$  and  $\mathbf{d}^{(k)}$ ,  $\mathbf{g}^{(k)}$  the diagonal and superdiagonal of  $\mathbf{R}_k$ . The reduction to upper triangular form of  $\mathbf{A}_k - \alpha_k \mathbf{I}$  can be obtained by a sequence of Givens rotations

$$\mathbf{G}_i^{(k)} = \begin{bmatrix} \mathbf{I}_{i-1} & 0 & 0 & 0 \\ 0 & c_i^{(k)} & s_i^{(k)} & 0 \\ 0 & -s_i^{(k)} & c_i^{(k)} & 0 \\ 0 & 0 & 0 & \mathbf{I}_{n-i-1} \end{bmatrix} \tag{6.204}$$

where  $c_i^{(k)}$  is real,  $|c_i^{(k)}|^2 + |s_i^{(k)}|^2 = 1$  and  $\mathbf{I}_i$  is the  $i \times i$  identity matrix.  $\mathbf{G}_1^{(k)}$  is chosen so that the (2,1) element of  $\mathbf{G}_1^{(k)}(\mathbf{A}_k - \alpha_k \mathbf{I})$  is 0. Thus only the elements in the first two rows of  $\mathbf{G}_1^{(k)}(\mathbf{A}_k - \alpha_k \mathbf{I})$  are changed from  $\mathbf{A}_k - \alpha_k \mathbf{I}$ . And, for  $j > 2$  the new elements are given by

$$\frac{1}{y_n^{(k)}} \hat{x}_i^{(k)} y_j^{(k)} + \hat{u}_{i1}^{(k)} \bar{v}_{j1}^{(k)} + \hat{u}_{i2}^{(k)} \bar{v}_{j2}^{(k)} \quad (i = 1, 2; j > i) \tag{6.205}$$

where

$$\begin{bmatrix} \hat{x}_1^{(k)} \\ \hat{x}_2^{(k)} \end{bmatrix} = \mathbf{G}_1^{(k)} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix}$$

with similar relations for  $\hat{u}_{il}^{(k)}$  ( $i, l = 1, 2$ ). Moreover the  $2 \times 2$  leading submatrix of  $\mathbf{G}_1^{(k)}(\mathbf{A}_k - \alpha_k \mathbf{I})$  is given by

$$\begin{bmatrix} d_1^{(k)} & g_1^{(k)} \\ 0 & d_2^{(k)} \end{bmatrix} = \mathbf{G}_1^{(k)} \begin{bmatrix} a_1^{(k)} - \alpha_k & \frac{1}{y_n^{(k)}} \hat{x}_1^{(k)} y_2^{(k)} + \hat{u}_{11}^{(k)} \bar{v}_{21}^{(k)} + \hat{u}_{12}^{(k)} \bar{v}_{22}^{(k)} \\ b_1^{(k)} & a_2^{(k)} - \alpha_k \end{bmatrix} \tag{6.206}$$

The values of  $d_1^{(k)}$ ,  $\hat{x}_1^{(k)}$ ,  $\hat{u}_{11}^{(k)}$ , and  $\hat{u}_{12}^{(k)}$  are not modified by later Givens rotations, while  $\hat{x}_2^{(k)}$ ,  $\hat{u}_{21}^{(k)}$ ,  $\hat{u}_{22}^{(k)}$  and  $d_2^{(k)}$  are modified only at the second step. At the  $i$ 'th step,  $G_i^{(k)}$  is chosen so that the  $(i+1,i)$  element of  $G_{i-1}^{(k)} \dots G_1^{(k)} (\mathbf{A}_k - \alpha_k)$  becomes zero. At the end, when the  $(n,n-1)$  element has been zeroed, we have  $\mathbf{R}_k$  as follows:

$$r_{ij}^{(k)} = \left\{ \begin{array}{ll} d_i^{(k)} & \text{for } i = j \\ g_i^{(k)} & \text{for } i = j - 1 \\ \frac{1}{y_n^{(k)}} \hat{x}_i^{(k)} y_j^{(k)} + \hat{u}_{i1}^{(k)} \bar{v}_{j1}^{(k)} + \hat{u}_{i2}^{(k)} \bar{v}_{j2}^{(k)} & \text{for } i < j - 1 \\ 0 & \text{for } i > j \end{array} \right\} \quad (6.207)$$

Thus  $\mathbf{R}_k$  can be stored with  $8n-1$  parameters. The authors combine the above into their

“ALGORITHM 1”: **Input**  $\mathbf{U}_k$ ,  $\mathbf{V}_k$ ,  $\mathbf{x}^{(k)}$ ,  $\mathbf{y}^{(k)}$ ,  $\mathbf{b}^{(k)}$  and  $\alpha_k$

**Output**  $s_i^{(k)}$ ,  $c_i^{(k)}$  ( $i = 1, \dots, n-1$ ),  $\mathbf{d}^{(k)}$ ,  $\hat{\mathbf{x}}^{(k)}$ ,  $\hat{\mathbf{U}}_k$

**Computation 1.** Let

$$\mathbf{a}^{(k)} = \left( \frac{1}{y_n^{(k)}} x_i^{(k)} y_i^{(k)} + u_{i1}^{(k)} \bar{v}_{i1}^{(k)} + u_{i2}^{(k)} \bar{v}_{i2}^{(k)} \right)_{i=1, \dots, n} \quad (6.208)$$

$$\hat{\mathbf{b}}^{(k)} = \mathbf{b}^{(k)}; \hat{\mathbf{U}}_k = \mathbf{U}_k \quad (6.209)$$

$$\hat{\mathbf{x}}^{(k)} = \mathbf{x}^{(k)}$$

2. Set  $\mathbf{a}^{(k)} = \mathbf{a}^{(k)} - \alpha_k(1, 1, \dots, 1)$

3. For  $i = 1, \dots, n-1$  do:

(a) (Compute  $\mathbf{G}_i$ )

$$(i) \gamma_i = \frac{1}{\sqrt{|a_i^{(k)}|^2 + |b_i^{(k)}|^2}}, \nu_i = \frac{a_i^{(k)}}{|a_i^{(k)}|}, (\nu_i = 1 \text{ if } a_i^{(k)} = 0)$$

$$\theta_i = \frac{\gamma_i}{\nu_i}$$

$$(ii) c_i^{(k)} = a_i^{(k)} \theta_i, s_i^{(k)} = b_i^{(k)} \theta_i$$

(b) (Update  $\mathbf{R}_k$ )

$$(i) d_i^{(k)} = \frac{\gamma_i}{y_n^{(k)}} \hat{x}_i^{(k)} y_{i+1}^{(k)} + \hat{u}_{i1}^{(k)} \bar{v}_{i+1,1}^{(k)} + \hat{u}_{i2}^{(k)} \bar{v}_{i+1,2}^{(k)}$$

$$(ii) g_i^{(k)} = c_i^{(k)} t + s_i^{(k)} a_{i+1}^{(k)}$$

$$(iii) d_{i+1}^{(k)} = -s_i^{(k)} t + c_i^{(k)} a_{i+1}^{(k)}$$

(c) (Update  $\hat{\mathbf{U}}_k$ )

$$(i) t = c_i^{(k)} \hat{u}_{i1}^{(k)} + s_i^{(k)} \hat{u}_{i+1,1}^{(k)}, \hat{u}_{i+1,1}^{(k)} = -\bar{s}_i^{(k)} \hat{u}_{i1}^{(k)} + c_i^{(k)} \hat{u}_{i+1,1}^{(k)}, \hat{u}_{i1}^{(k)} = t,$$

$$(ii) t = c_i^{(k)} \hat{u}_{i2}^{(k)} + s_i^{(k)} \hat{u}_{i+1,2}^{(k)}, \hat{u}_{i+1,2}^{(k)} = -\bar{s}_i^{(k)} \hat{u}_{i2}^{(k)} + c_i^{(k)} \hat{u}_{i+1,2}^{(k)}, \hat{u}_{i2}^{(k)} = t.$$

(d) (Update  $\hat{\mathbf{x}}^{(k)}$ )

$$(i) t = c_i^{(k)} \hat{x}_i^{(k)} + s_i^{(k)} \hat{x}_{i+1}^{(k)},$$

$$(ii) \hat{x}_{i+1}^{(k)} = -s_i^{(k)} \hat{x}_i^{(k)} + c_i^{(k)} \hat{x}_{i+1}^{(k)},$$

$$(iii) \hat{x}_i^{(k)} = t.$$

4. End do

Now since

$$G_{n-1}^{(k)} \dots G_2^{(k)} G_1^{(k)} (\mathbf{A}_k - \alpha_k \mathbf{I}) = \mathbf{R}_k \quad (6.210)$$

we have that the  $\mathbf{Q}_k$  in  $\mathbf{A}_k - \alpha_k \mathbf{I} = \mathbf{Q}_k \mathbf{R}_k$  is given by

$$\mathbf{Q}_k = (G_{n-1}^{(k)} \dots G_1^{(k)})^H = G_1^{(k)H} \dots G_{n-1}^{(k)H} \quad (6.211)$$

Also, it is proved (in effect) by Gill et al (1974) that if  $c_i^{(k)}$  (real) and  $s_i^{(k)}$  (perhaps complex) are the Givens rotation parameters, and

$$\mathbf{D}^{(k)} = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & -s_1^{(k)} & 0 & \dots & 0 \\ 0 & 0 & s_1^{(k)} s_2^{(k)} & 0 & \dots \\ 0 & \dots & \dots & 0 & (-1)^{n-1} s_1^{(k)} s_2^{(k)} \dots s_{n-1}^{(k)} \end{bmatrix} \quad (6.212)$$

$$\mathbf{p}^{(k)} = \mathbf{D}^{(k)-1} [1, c_1^{(k)}, c_2^{(k)}, \dots, c_{n-1}^{(k)}]^T \quad (6.213)$$

$$= \left[ 1, -\frac{c_1^{(k)}}{s_1^{(k)}}, \frac{c_2^{(k)}}{s_1^{(k)} s_2^{(k)}}, \dots, (-1)^{n-1} \frac{c_{n-1}^{(k)}}{s_1^{(k)} s_2^{(k)} \dots s_{n-1}^{(k)}} \right] \quad (6.214)$$

$$\mathbf{q}^{(k)} = \mathbf{D}^{(k)} [c_1^{(k)}, c_2^{(k)}, \dots, c_{n-1}^{(k)}, 1] \quad (6.215)$$

$$= [c_1^{(k)}, -s_1^{(k)} c_2^{(k)}, s_1^{(k)} s_2^{(k)} c_3^{(k)}, \dots, (-1)^{n-1} s_1^{(k)} \dots s_{n-1}^{(k)}] \quad (6.216)$$

then

$$\mathbf{Q}_k = \begin{bmatrix} q_1^{(k)} p_1^{(k)} & q_2^{(k)} p_1^{(k)} & \dots & q_n^{(k)} p_1^{(k)} \\ \bar{s}_1^{(k)} & q_2^{(k)} p_2^{(k)} & \dots & q_n^{(k)} p_2^{(k)} \\ 0 & \bar{s}_2^{(k)} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & q_{n-1}^{(k)} p_{n-1}^{(k)} & q_n^{(k)} p_{n-1}^{(k)} \\ \dots & \dots & \bar{s}_{n-1}^{(k)} & q_n^{(k)} p_n^{(k)} \end{bmatrix} \quad (6.217)$$

i.e.

$$q_{ij}^{(k)} = \begin{cases} 0 & \text{if } i > j + 1 \\ \bar{s}_i^{(k)} & \text{if } i = j + 1 \\ (-1)^{i+j} c_{i-1}^{(k)} c_j^{(k)} \prod_{l=i}^{j-1} s_l^{(k)} & \text{if } i \leq j \end{cases} \quad (6.218)$$

with  $c_0^{(k)} = c_n^{(k)} = 1$

With the above algorithm and the reverse RQ step (see later) we can design an algorithm for performing a single QR-step in  $O(n)$  operations. The factorization

into  $\mathbf{Q}_k \mathbf{R}_k$  can be done in about  $35n$  operations. The reverse  $\mathbf{R}_k \mathbf{Q}_k + \alpha_k \mathbf{I}$  step proceeds as follows:

**Given**  $\mathbf{s}^{(k)}$ ,  $\mathbf{c}^{(k)}$  defining the Givens rotations and hence

$\mathbf{Q}_k$ ;  $\mathbf{x}^{(k)}$  and  $\mathbf{y}^{(k)}$ ;  $\mathbf{U}_k$ ,  $\mathbf{V}_k$ ;  $\mathbf{d}^{(k)}$ ,  $\mathbf{g}^{(k)}$ ,  $\hat{\mathbf{x}}^{(k)}$  and  $\hat{\mathbf{U}}_k$  defining (together with  $\mathbf{y}^{(k)}$  and  $\mathbf{V}_k$ )  $\mathbf{R}_k$  via 6.207;  $\alpha_k$ .

**Compute** (i)  $\mathbf{a}^{(k+1)}$ , the diagonal elements of  $\mathbf{A}_{k+1}$ .

We have  $\mathbf{A}_{k+1} = \mathbf{R}_k \mathbf{Q}_k + \alpha_k \mathbf{I}$  where  $\mathbf{R}_k$  is right triangular and  $\mathbf{Q}_k$  Hessenberg so that

$$\begin{aligned} a_{ii}^{(k+1)} &= r_{ii}^{(k)} q_{ii}^{(k)} + g_i^{(k)} q_{i+1,i}^{(k)} + \alpha_k = \\ r_{ii}^{(k)} c_{i-1,i}^{(k)} + g_i^{(k)} \bar{s}_i^{(k)} + \alpha_k \quad (i = 1, \dots, n) \end{aligned} \quad (6.219)$$

(N.B.  $r_{ii} = d_i$ ).

We also compute  $\mathbf{b}^{(k+1)}$  where

$$b_i^{(k+1)} = d_{i+1}^{(k)} \bar{s}_i^{(k)} \quad (i = 1, \dots, n) \quad (6.220)$$

(ii) Compute  $\mathbf{U}_{k+1}$  and  $\mathbf{V}_{k+1}$  (such that 6.198 holds for  $\mathbf{A}_{k+1}$ ) by applying  $n-1$  Givens rotations to the two columns of  $\mathbf{U}_k$  and  $\mathbf{V}_k$  using  $24(n-1)$  operations.

(iii)  $\mathbf{x}^{(k+1)}$  and  $\mathbf{y}^{(k+1)}$ : the authors suggest 5 different algorithms for computing these quantities. We will describe their third and fifth methods. Method 3 uses

$$\mathbf{A}_{k+1} = \mathbf{R}_k \mathbf{A}_k \mathbf{R}_k^{-1} \quad (6.221)$$

to deduce that

$$\mathbf{x}^{(k+1)} = \mathbf{A}_{k+1}^{-H} \mathbf{e}_n = \mathbf{R}_k^{-H} \mathbf{A}_k^{-H} \mathbf{R}_k^H \mathbf{e}_n = \bar{r}_{nn}^{(k)} \mathbf{R}_k^{-H} \mathbf{x}^{(k)} \quad (6.222)$$

and

$$\mathbf{y}^{(k+1)} = \mathbf{e}_1^T \mathbf{A}_{k+1}^{-H} = \mathbf{e}_1^T \mathbf{R}_k^{-H} \mathbf{A}_k^{-H} \mathbf{R}_k^H = \frac{1}{\hat{r}_{11}^{(k)}} \mathbf{y}^{(k)T} \mathbf{R}_k^H \quad (6.223)$$

6.222 is equivalent to the triangular system of special form

$$\mathbf{R}_k^H \mathbf{x}^{(k+1)} = \hat{r}_{nn} \mathbf{x}^{(k)} \quad (6.224)$$

which can be solved in  $O(n)$  operations by a method similar to Algorithm 2 below. The vector-matrix multiplication in 6.223 can also be performed in  $O(n)$  operations, according to Bini et al although they do not give details of how to do this. Their “method 5” applies if  $\alpha_k = 0$ , in which case  $\mathbf{A}_{k+1} = \mathbf{R}_k \mathbf{Q}_k$  so that  $\mathbf{A}_{k+1}^{-1} = \mathbf{Q}_k^{-1} \mathbf{R}_k^{-1}$  and  $\mathbf{A}_{k+1}^{-H} = \mathbf{R}_k^{-H} \mathbf{Q}_k^{-H}$ . Then

$$\mathbf{x}^{(k+1)} = \mathbf{R}_k^{-H} \mathbf{Q}_k \mathbf{e}_n \quad (6.225)$$

$$\mathbf{y}^{(k+1)} = \frac{1}{\bar{r}_{11}^{(k)}} \mathbf{Q}_k^H \mathbf{e}_1 \quad (6.226)$$

To avoid stability problems in the calculation of  $\mathbf{x}^{(k+1)}$  and  $\mathbf{y}^{(k+1)}$  Bini et al recommend writing them as

$$\mathbf{x}^{(k+1)} = (-1)^{n-1} (s_1^{(k)} \dots s_{n-1}^{(k)}) \mathbf{D}^{(k)-1} \mathbf{z}^{(k+1)} \quad (6.227)$$

$$\mathbf{y}^{(k+1)} = \overline{\mathbf{D}}^{(k)} \mathbf{w}^{(k+1)} \quad (6.228)$$

where

$$\mathbf{w}^{(k+1)} = \frac{1}{\overline{r}_{11}^{(k)}} (c_1^{(k)}, \dots, c_{n-1}^{(k)}, 1)^T \quad (6.229)$$

$$\mathbf{z}^{(k+1)} = \mathbf{D}^{(k)} \mathbf{R}_k^{-H} \mathbf{D}^{(k)-1} (1, c_1^{(k)}, \dots, c_{n-1}^{(k)})^T \quad (6.230)$$

These can be derived as follows: by 6.226  $\mathbf{y}^{(k+1)} = \frac{1}{\overline{r}_{11}^{(k)}} \mathbf{Q}_k^H \mathbf{e}_1$

= (by 6.217)  $\frac{1}{\overline{r}_{11}^{(k)}} \overline{\mathbf{Q}}^{(k)} p_1^{(k)} =$  (by 6.214 and 6.215)

$$\frac{1}{\overline{r}_{11}^{(k)}} \overline{\mathbf{D}}^{(k)} (c_1^{(k)}, \dots, c_{n-1}^{(k)}, 1)^T = \overline{\mathbf{D}} \mathbf{w}^{(k+1)}.$$

The derivation of 6.227 is similar.

6.229 can be implemented easily in  $n$  divisions; for 6.230 we will use the method of Algorithm 2 below.

In the QR factorization step of Algorithm 1 we will replace  $\mathbf{x}^{(k)}$ ,  $\hat{\mathbf{x}}^{(k)}$ , and  $\mathbf{y}^{(k)}$  by  $\mathbf{z}^{(k)}$ ,  $\hat{\mathbf{z}}^{(k)}$  and  $\mathbf{w}^{(k)}$  respectively. The relations between these are given by 6.227 and 6.228 with  $k+1$  replaced by  $k$  (and  $\mathbf{x}$  by  $\hat{\mathbf{x}}$  etc.). This modifies Algorithm 1 in stage 1, 3(b) and 3(d) which become:

$$1. \mathbf{a}^{(k)} = \left( \frac{1}{w_n^{(k)}} z_i^{(k)} w_i^{(k)} + u_{i1}^{(k)} \overline{v}_{i1}^{(k)} + u_{i2}^{(k)} \overline{v}_{i2}^{(k)} \right)_{i=1..n}, \hat{\mathbf{b}}^{(k)} = \mathbf{b}^{(k)}, \hat{\mathbf{U}}_k = \mathbf{U}_k, \hat{\mathbf{z}}^{(k)} = \mathbf{z}^{(k)}$$

3(b) (Update  $\mathbf{R}_k$ )

$$(i) d_i^{(k)} = \gamma_i^{(k)}, t = \frac{1}{w_n^{(k)}} \hat{z}_i^{(k)} w_{i+1}^{(k)} s_i^{(k-1)} + \hat{u}_{i1}^{(k)} \overline{v}_{i+1,1}^{(k)} + \hat{u}_{i2}^{(k)} \overline{v}_{i+1,2}^{(k)}$$

$$(ii) g_i^{(k)} = c_i^{(k)} t + s_i^{(k)} a_{i+1}^{(k)},$$

$$(iii) d_{i+1}^{(k)} = -s_i^{(k)} t + c_i^{(k)} a_{i+1}^{(k)}.$$

(d) (Update  $\hat{\mathbf{z}}^{(k)}$ )

$$(i) t = c_i^{(k)} \hat{z}_i^{(k)} - \frac{s_i^{(k)}}{s_{i-1}^{(k)}} \hat{z}_{i+1}^{(k)},$$

$$(ii) \hat{z}_{i+1}^{(k)} = s_i^{(k)} s_{i-1}^{(k-1)} \hat{z}_i^{(k)} + c_i^{(k)} \hat{z}_{i+1}^{(k)},$$

$$(iii) \hat{z}_i^{(k)} = t.$$

The cost of this is still  $O(n)$ . Possible overflow in  $\frac{s_i^{(k)}}{s_{i-1}^{(k)}}$  does not occur, since if  $s_{i-1}^{(k)} \rightarrow 0$  then convergence has taken place and we deflate and continue with a smaller matrix.

For the RQ step, we have to compute  $\mathbf{z}^{(k+1)}$  from 6.230 by solving

$$\mathbf{D}^{(k)} \mathbf{R}_k^H \mathbf{D}^{(k)-1} \mathbf{z}^{(k+1)} = \mathbf{b} = (1, c_1^{(k)}, \dots, c_{n-1}^{(k)}) \tag{6.231}$$

The coefficient matrix above, obtained from 6.207, is given by

$$\left\{ \begin{array}{ll} d_i^{(k)} & \text{if } i = j \\ -s_i^{(k)} \overline{g}_i^{(k)} & \text{if } i = j + 1 \\ (s_j^{(k)} \dots s_{i-1}^{(k)}) (\frac{1}{\overline{w}_n^{(k)}} \overline{w}_i^{(k)} \overline{z}_j^{(k)} (\overline{s}_j^{(k)-1} \dots \overline{s}_{i-1}^{(k)-1}) + & \\ (-1)^{i+j} (v_{i1}^{(k)} \overline{u}_{j1}^{(k)} + v_{i2}^{(k)} \overline{u}_{j2}^{(k)}) & \text{if } i > j + 1 \\ 0 & \text{if } i < j \end{array} \right\} \tag{6.232}$$

To solve 6.231 we use Algorithm 2 as follows:

**Input:**  $\mathbf{d}^{(k)}$ ,  $\mathbf{g}^{(k)}$ ,  $\hat{\mathbf{z}}^{(k)}$ ,  $\mathbf{w}^{(k)}$ ,  $\hat{\mathbf{U}} = (\hat{u}_{ij}^{(k)})$ ,  $\mathbf{V} = (v_{ij}^{(k)})$ , the Givens parameters  $s_i^{(k-1)}$  and  $c_i^{(k-1)}$ ;  $\mathbf{b}$  (r.h.s. of 6.231) and  $s_i^{(k)}$  ( $i = 1, \dots, n - 1$ ).

**Output:** Solution  $\mathbf{z}^{(k+1)}$  of 6.231

**Computation**

Set  $z_1^{(k+1)} = \frac{b_1}{d_1^{(k)}}$ ,  $z_2^{(k+1)} = \frac{(b_2 + \overline{g}_1^{(k)} s_1^{(k)}) z_1^{(k+1)}}{d_2^{(k)}}$ ,  $\gamma_{2,1} = \gamma_{2,2} = \phi_2 = 0$

For  $i=3, \dots, n$  do:

1.  $\gamma_{ij} = s_{i-1}^{(k)} (-\gamma_{i-1,j} + \overline{u}_{i-2,j}^{(k)} s_{i-2}^{(k)} z_{i-2}^{(k+1)})$ ,  $j = 1, 2$
2.  $\phi_i = s_{i-1}^{(k)} \overline{s}_{i-1}^{(k-1)} (\phi_{i-1} - \overline{z}_{i-2}^{(k)} \overline{s}_{i-2}^{(k-1)} s_{i-2}^{(k)} z_{i-2}^{(k+1)})$ ,
3.  $z_i^{(k+1)} = \frac{(b_i + \overline{g}_{i-1}^{(k)} s_{i-1}^{(k)} z_{i-1}^{(k+1)} - v_{i1}^{(k)} \gamma_{i1} - v_{i2}^{(k)} \gamma_{i2} + \hat{u}_i^{(k)} \frac{\phi_i}{\overline{w}_n^{(k)}})}{\overline{d}_i^{(k)}}$

End do.

Numerical experiments were performed with polynomials such as

(1) Wilkinson’s i.e.  $p(x) = \prod_{i=1}^n (x - i)$  for  $n = 10, 20$

(2)  $p(x) = (x^{n-m} + 1) \prod_{i=2}^{m+1} (x - \frac{1}{i})$  for  $m=20$ ,  $n \leq 10^6$ .

For (1) with  $n=20$ , the algorithm failed when  $\alpha_0 = 0$ , but it gave results correct to at least 2 decimal places when  $\alpha_0 = 22$ . For (2) the cost grows linearly with  $n$ , while the error is almost independent of  $n$ . In some cases breakdowns due to underflow/overflow have been encountered. The authors conclude that the algorithm is not robust and needs more investigation.

Bini, Gemignani and Pan (2004a) describe an inverse power method for a generalized companion matrix. Suppose we have  $n$  distinct values  $s_1, \dots, s_n$ ; then we define a rank-one matrix  $E_d$  with diagonal entries

$$d_i = \frac{p(s_i)}{q_i(s_i)} \tag{6.233}$$

where

$$q_i(x) = \prod_{j \neq i} (x - s_j) \tag{6.234}$$

and an associated companion matrix

$$\mathbf{C} = \mathbf{D}_s - \mathbf{E}_d \quad (6.235)$$

where  $\mathbf{D}_s = \begin{bmatrix} s_1 & 0 & \dots \\ \dots & \dots & \dots \\ \dots & 0 & s_n \end{bmatrix}$  (i.e. it is a diagonal matrix with (i,i) element  $s_i$ ) and  $\mathbf{E}_d$  is somewhat arbitrary, but Elsner (1973) proposes

$$\mathbf{E}_d = \begin{bmatrix} d_1 & d_2 & \dots & d_n \\ d_1 & d_2 & \dots & d_n \\ \dots & \dots & \dots & \dots \\ d_1 & d_2 & \dots & d_n \end{bmatrix} \quad (6.236)$$

The authors quote Carstensen (1991) as proving that

$$\det(x\mathbf{I} - \mathbf{C}) = p(x) \quad (6.237)$$

i.e.  $\mathbf{C}$  is indeed a companion matrix.

Now the inverse power method for a general matrix  $\mathbf{C}$  is defined as follows: let  $z^{(0)}$  be a sufficiently close approximation to an eigenvalue  $z_j$  of  $\mathbf{C}$  and let

$$\mathbf{v} = \sum_{k=1}^n a_k \mathbf{v}_k \quad (6.238)$$

with  $\|\mathbf{v}\|_2 = 1$ , where  $\mathbf{v}_k$  ( $k = 1, \dots, n$ ) are the eigenvectors of  $\mathbf{C}$  and  $a_k \neq 0$ . Let

$$\mathbf{x}^{(0)} = \mathbf{v}, \quad (6.239)$$

$$\mathbf{y}^{(i)} = (\mathbf{C} - z^{(i-1)}\mathbf{I})^{-1}\mathbf{x}^{(i-1)} \quad (6.240)$$

$$\mathbf{x}^{(i)} = \frac{\mathbf{y}^{(i)}}{\|\mathbf{y}^{(i)}\|_2} \quad (6.241)$$

$$z^{(i)} = \mathbf{x}^{(i)T}\mathbf{C}\mathbf{x}^{(i)} \quad (6.242)$$

All the above 3 equations are repeated for  $i = 1, 2, \dots$

Then the pairs  $(\mathbf{y}^{(i)}, z^{(i)})$  rapidly converge to an eigenvector/ eigenvalue pair  $(\mathbf{v}_j, z_j)$  (under certain conditions). The authors in their section 3 describe several methods of choosing the initial  $z$ . They prove that for  $\mathbf{C}$  as in 6.235 and 6.236 the vector products  $\mathbf{C}\mathbf{x}$  and  $(\mathbf{C} - z\mathbf{I})^{-1}\mathbf{x}$  can be performed in  $O(n)$  flops. For the  $i$ 'th component of

$$\mathbf{C}\mathbf{x} = s_i x_i - \left(\sum d_j x_j\right), \quad (6.243)$$

where  $\sum d_j x_j$  only needs to be done once for all  $i$  ( $2n$  flops) and then 6.243 takes 2 flops for each  $i$ . Also

$$(\mathbf{C} - z\mathbf{I})^{-1} = ([\mathbf{D} - z\mathbf{I}] - \mathbf{1d}^T)^{-1} \quad (6.244)$$



where  $\mathbf{1}^T = [1, 1, \dots, 1]$  and  $\mathbf{d}^T = [d_1, d_2, \dots, d_n]$ . They apply the Sherman-Morrison-Woodbury formula (see Golub and Van Loan (1996) p50) i.e.

$$(\mathbf{A} + \mathbf{U}\mathbf{V}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_{k \times k} + \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^T\mathbf{A}^{-1} \quad (6.245)$$

where  $\mathbf{U}$  and  $\mathbf{V}$  are  $n \times k$ . They set  $\mathbf{A} = \mathbf{D} - z\mathbf{I}$ ,  $\mathbf{U} = -\mathbf{1}$ , and  $\mathbf{V} = \mathbf{d}$  so that 6.245 becomes

$$(\mathbf{C} - z\mathbf{I})^{-1} = ([\mathbf{D} - z\mathbf{I}] - \mathbf{1}\mathbf{d}^T)^{-1} = (\mathbf{D} - z\mathbf{I})^{-1} + (\mathbf{D} - z\mathbf{I})^{-1}\mathbf{1}(\mathbf{I}_{1 \times 1} - \mathbf{d}^T(\mathbf{D} - z\mathbf{I})^{-1}\mathbf{1})^{-1}\mathbf{d}^T(\mathbf{D} - z\mathbf{I})^{-1} \quad (6.246)$$

$$= (\mathbf{I}_{n \times n} + \frac{1}{1 - \tau}(\mathbf{D} - z\mathbf{I})^{-1}\mathbf{1}\mathbf{d}^T)(\mathbf{D} - z\mathbf{I})^{-1} \quad (6.247)$$

where

$$\tau = \mathbf{d}^T(\mathbf{D} - z\mathbf{I})^{-1}\mathbf{1} \quad (6.248)$$

Hence

$$(\mathbf{C} - z\mathbf{I})^{-1}\mathbf{v} = (\mathbf{D} - z\mathbf{I})^{-1}\mathbf{v} + \frac{\sigma}{1 - \tau}(\mathbf{D} - z\mathbf{I})^{-1}\mathbf{1} \quad (6.249)$$

where

$$\sigma = \mathbf{d}^T(\mathbf{D} - z\mathbf{I})^{-1}\mathbf{v} \quad (6.250)$$

Hence we can compute  $\mathbf{y} = (\mathbf{C} - z\mathbf{I})^{-1}\mathbf{v}$  by performing  $n$  reciprocations,  $4n$  multiplications, and  $4n$  additions (or subtractions). For complex data this gives  $37n + O(1)$  real flops. In more detail the algorithm proceeds as follows:

1. Compute

$$\mathbf{g} = (\mathbf{D} - z\mathbf{I})^{-1}\mathbf{1} \quad (6.251)$$

2. Compute

$$\mathbf{u} = \mathbf{g} * \mathbf{v} \quad (6.252)$$

where  $*$  denotes the componentwise product.

3. Compute

$$\tau = \sum_{i=1}^n d_i g_i \quad (6.253)$$

4. Compute

$$\sigma = \sum_{i=1}^n d_i u_i \quad (6.254)$$

5. Compute

$$\mathbf{y} = \mathbf{u} + \frac{\sigma}{1 - \tau}\mathbf{g} \quad (6.255)$$

A random initial eigenvector is usually a good choice for  $\mathbf{v}$ . Or, if we have an approximation  $z_j$  to one of the eigenvalues we may use

$$\mathbf{v}_j = [1, z_j, z_j^2, \dots, z_j^{n-1}]^T \tag{6.256}$$

or

$$\mathbf{v}_j = \left[ \frac{1}{s_1 - z_j}, \frac{1}{s_2 - z_j}, \dots, \frac{1}{s_n - z_j} \right]^T \tag{6.257}$$

When a zero  $z_j$  is closely approximated, we may deflate (i.e. compute  $\frac{p(x)}{(x-z_j)}$ ) at a cost of  $2n-2$  multiplications and  $n-1$  subtractions. The authors state that 6.242 may be replaced by a cheaper calculation

$$z^{(i)} = \frac{(\mathbf{C}\mathbf{y}^{(i)})_j}{y_j^{(i)}} \tag{6.258}$$

where  $j$  is such that  $y_j^{(i)} \neq 0$ . Then for  $\mathbf{C}$  given by 6.235 and 6.236 this becomes

$$z^{(i)} = s_j - \frac{\sum_{k=1}^n d_k y_k^{(i)}}{y_j^{(i)}} \tag{6.259}$$

Deflation of a calculated eigenvalue is fairly inexpensive. Let  $z$  be the computed eigenvalue, and let  $s_n$  be the initial approximation closest to  $z$  (we may reorder the  $s_i$  to achieve this). Let

$$\hat{\mathbf{s}} = (\hat{s}_1, \dots, \hat{s}_{n-1}, \hat{s}_n)^T \equiv (s_1, \dots, s_{n-1}, z)^T \tag{6.260}$$

and

$$\hat{d}_i = \frac{p(\hat{s}_i)}{\prod_{j \neq i} (\hat{s}_i - \hat{s}_j)} \tag{6.261}$$

Then as  $p(z) = 0$ , the last column of  $\mathbf{C}_{\hat{\mathbf{s}}, \hat{\mathbf{d}}} = \mathbf{D}_{\hat{\mathbf{s}}} - \mathbf{E}_{\hat{\mathbf{s}}}$  is given by  $(0, 0, \dots, 0, z)$ . Hence

$$|\mathbf{C}_{\hat{\mathbf{s}}, \hat{\mathbf{d}}} - x\mathbf{I}| = (z - x)|\mathbf{G} - x\mathbf{I}| \tag{6.262}$$

where  $\mathbf{G}$  is the leading  $(n - 1) \times (n - 1)$  principal submatrix of  $\mathbf{C}_{\hat{\mathbf{s}}, \hat{\mathbf{d}}}$ ; so the latter coincides with the generalized companion matrix associated with the deflated polynomial  $\frac{p(x)}{(x-z)}$  and the vector  $(s_1, \dots, s_{n-1})^T$ . This matrix is defined by the vectors  $(s_1, \dots, s_{n-1})^T$  and  $(\hat{d}_1, \dots, \hat{d}_{n-1})^T$ . The former is already known, but the latter needs to be calculated, which can be done by

$$\hat{d}_i = d_i \frac{s_i - s_n}{s_i - z} \quad (i = 1, \dots, n - 1) \tag{6.263}$$

Thus we can deflate with  $2(n-1)$  subtractions,  $(n-1)$  divisions, and  $(n-1)$  multiplications. The shifted inverse power and deflation process may be repeated for  $k =$

$n, n-1, \dots, 1$  (for  $k = 1$  the eigenvalue is  $s_1 - d_1$ ).

The above method is particularly efficient if only one or a few eigenvalues of  $\mathbf{C}$  (roots of  $p(x)$ ) are required. The authors quote Wilkinson's (1963) criterion for judging whether a given approximation  $\xi$  to a zero of  $p(x)$  is a zero of a slightly perturbed polynomial: let  $fl(p(\xi))$  be the value obtained by computing  $p(\xi)$  by means of Horner's rule

$$u_0 = c_n, u_{i+1} = \xi u_i + c_{n-i-1} \quad (i = 0, \dots, n-1), p(\xi) = u_n \quad (6.264)$$

in floating point arithmetic with machine precision  $\mu$ . If

$$|fl(p(\xi))| \leq \delta \sum_{i=0}^n |c_i| |\xi|^i \quad (6.265)$$

where

$$\delta = (12n + 3)\mu \quad (6.266)$$

then there exists a polynomial

$$\tilde{p}(x) = \sum_{i=0}^n \tilde{c}_i x^i \quad (6.267)$$

such that

$$\tilde{c}_i = c_i(1 + \epsilon_i), |\epsilon_i| \leq \delta \text{ and } \tilde{p}(\xi) = 0 \quad (6.268)$$

If 6.265 is not satisfied, then for any polynomial  $\tilde{p}(x)$  such that  $|\epsilon_i| \leq \frac{\delta}{3}$  we have  $\tilde{p}(\xi) \neq 0$ . If 6.265 is satisfied, we say that  $\xi$  is a  $\delta$ -approximated zero of  $p(x)$ . The authors' "Algorithm 7.2" computes approximations  $\xi_1, \dots, \xi_n$  to the zeros of  $p(x)$  satisfying 6.265 for each  $\xi = \xi_i$  ( $i = 1, \dots, n$ ). The computation is as follows:

1. Compute initial approximations  $s_1, \dots, s_n$  by previously mentioned methods. Set  $m = n$ ,  $\delta = (12n + 3)\mu$ .
  2. While  $m > 0$  do:
    - 2a. Compute  $d_1, \dots, d_n$  by 6.233 and check if  $s_i$  is a  $\delta$ -approximate zero of  $p(x)$ .
    - 2b. Sort the  $s_i$  so that  $\delta$ -approximated components are at the bottom and components not yet  $\delta$ -approximated are ordered with non-increasing modulus.
    - 2c. Let  $m$  = number of components not yet  $\delta$ -approximated.
    - 2d. Apply the shifted inverse power method to the  $m \times m$  generalized companion matrix  $\mathbf{C}_{s,d} = \mathbf{C}_s - \mathbf{E}_d$  defined by  $s_1, \dots, s_m, d_1, \dots, d_m$  and output approximations  $\xi_1, \dots, \xi_m$ . Set  $s_i = \xi_i$  ( $i = 1, \dots, m$ )
- End While.

In numerical experiments the new method was often much faster than the WDK method, for example it found the roots of  $x^{2000} - 1$  ten times faster than the WDK

method did.

Bini, Gemignani, and Pan (2004b) give a method based on evaluating  $p(z)$  at the  $n$ 'th roots of unity, i.e.  $\omega^j$  where

$$\omega = \exp\left(\frac{2\pi\sqrt{-1}}{n}\right) \tag{6.269}$$

It is assumed that  $p(z)$  can be evaluated at any point  $z$  without explicit knowledge of the coefficients. By applying the Lagrange interpolation formula at the nodes  $\omega_j \equiv \omega^j$  ( $j = 0, \dots, n - 1$ ) we get

$$p(z) - c_n z^n = \sum_{i=0}^{n-1} (p(\omega_i) - c_n \omega_i^n) \frac{\prod_{j \neq i} (z - \omega_j)}{\prod_{j \neq i} (\omega_i - \omega_j)} \tag{6.270}$$

$$= (z^n - 1) \sum_{i=0}^{n-1} \frac{(p(\omega_i) - c_n)}{(z - \omega_i) \prod_{j \neq i} (\omega_i - \omega_j)} \tag{6.271}$$

(as  $\omega_i^n = \exp\left(\frac{2\pi i \sqrt{-1}}{n} n\right) = 1$ )

The product above =

$$\prod_{j \neq i} \omega_i^{n-1} (1 - \omega_{j-i}) \tag{6.272}$$

$$= \frac{\omega_i^n}{\omega_i} \prod_{j \neq i} (1 - \omega_{j-i}) = \frac{n}{\omega_i} \tag{6.273}$$

(For  $\prod_{j \neq i} (z - \omega_{j-i}) = \prod_{l=1}^{n-1} (z - \omega_l) = \frac{\prod_{l=0}^{n-1} (z - \omega_l)}{(z - 1)} = \frac{z^n - 1}{z - 1} = z^{n-1} + z^{n-2} + \dots + z + 1 \rightarrow n$  as  $z \rightarrow 1$ ).

Hence

$$p(z) - c_n z^n = (z^n - 1) \sum_{i=0}^{n-1} \frac{\omega_i (p(\omega_i) - c_n)}{n(z - \omega_i)} \tag{6.274}$$

Hence

$$p(z) = c_n z^n + (z^n - 1) \sum_{i=0}^{n-1} \frac{\omega_i p(\omega_i)}{n(z - \omega_i)} - c_n (z^n - 1) \sum_{i=0}^{n-1} \frac{\omega_i}{n(z - \omega_i)} \tag{6.275}$$

But  $\frac{1}{z^n - 1} = \sum_{i=0}^{n-1} \frac{\prod_{j \neq i} \frac{1}{(\omega_i - \omega_j)}}{z - \omega_i}$

$$= \sum_{i=0}^{n-1} \frac{1}{z - \omega_i} \prod_{j=0, j \neq i}^{n-1} \frac{\omega_i}{\omega_i^n (1 - \omega_{j-i})} = \sum_{i=0}^{n-1} \frac{\omega_i}{n(z - \omega_i)} \tag{6.276}$$

Hence last term in 6.275 =  $-c_n$ .

Hence

$$p(z) = (z^n - 1)c_n + \sum_{i=0}^{n-1} \frac{\omega_i p(\omega_i)}{n(z - \omega_i)} \quad (6.277)$$

Putting  $z = 0$  gives

$$p(0) = -c_n - \sum_{i=1}^{n-1} \frac{\omega_i p(\omega_i)}{n(-\omega_i)} \quad (6.278)$$

i.e.

$$c_n = \sum_{i=0}^{n-1} \frac{p(\omega_i)}{n} - p(0) \quad (6.279)$$

The root-finding problem for  $p(z)$  in the form 6.277 can be expressed as the computation of the eigenvalues of a generalized companion matrix

$$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & \omega & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & \omega^{n-1} \end{bmatrix} - \frac{1}{nc_n} \begin{bmatrix} p(1) \\ p(\omega) \\ \dots \\ p(\omega^{n-1}) \end{bmatrix} [ 1 \quad \omega \quad \dots \quad \omega^{n-1} ] \quad (6.280)$$

For we will show that  $\hat{\mathbf{A}}$  and  $\mathbf{F}$  have the same eigenvalues, while it is known that the eigenvalues of  $\mathbf{F}$  are zeros of  $p(z)$ . To show this, recall that

$$\mathbf{F} = \begin{bmatrix} 0 & \dots & \dots & 0 & -\frac{c_0}{c_n} \\ 1 & 0 & \dots & 0 & -\frac{c_1}{c_n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & -\frac{c_{n-1}}{c_n} \end{bmatrix} \quad (6.281)$$

We may re-write this as

$$\begin{bmatrix} 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 \end{bmatrix} + \begin{bmatrix} -1 - \frac{c_0}{c_n} \\ -\frac{c_1}{c_n} \\ \dots \\ -\frac{c_{n-1}}{c_n} \end{bmatrix} [ 0 \quad \dots \quad 0 \quad 1 ] \quad (6.282)$$

$$\equiv \mathbf{Z} + \mathbf{p}\mathbf{e}_n^T \quad (6.283)$$

where

$$\mathbf{p} = \begin{bmatrix} -1 - \frac{c_0}{c_n} \\ -\frac{c_1}{c_n} \\ \dots \\ -\frac{c_{n-1}}{c_n} \end{bmatrix}, \mathbf{e}_n^T = [ 0 \quad \dots \quad 0 \quad 1 ]$$

Now let

$$\Omega = (\omega^{(i-1)(j-1)}) = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & \omega & \omega^2 & \dots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \dots & \omega^{2n-2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \omega^{n-1} & \omega^{2n-2} & \dots & \omega^{(n-1)(n-1)} \end{bmatrix} \tag{6.284}$$

$$\mathbf{V} = \frac{1}{\sqrt{n}}\Omega \tag{6.285}$$

and

$$\hat{\mathbf{D}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \omega & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \omega^{n-1} \end{bmatrix} \tag{6.286}$$

Then

$$\mathbf{V}^H \hat{\mathbf{D}} \mathbf{V} = \frac{1}{n} \begin{bmatrix} 1 & 1 & \dots & \dots & 1 \\ 1 & \bar{\omega} & \bar{\omega}^2 & \dots & \bar{\omega}^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \bar{\omega}^{n-1} & \dots & \dots & \bar{\omega}^{(n-1)(n-1)} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ \omega & \omega^2 & \dots & \omega^n \\ \omega^2 & \omega^4 & \dots & \omega^{2n} \\ \dots & \dots & \dots & \dots \\ \omega^{n-1} & \omega^{2n-2} & \dots & \omega^{(n-1)n} \end{bmatrix} \tag{6.287}$$

= (since  $\bar{\omega} = \omega^{-1}$ )

$$\frac{1}{n} \begin{bmatrix} 1 & 1 & \dots & \dots & 1 \\ 1 & \omega^{-1} & \omega^{-2} & \dots & \omega^{-(n-1)} \\ 1 & \omega^{-2} & \omega^{-4} & \dots & \omega^{-(2n-2)} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \omega^{-(n-1)} & \dots & \dots & \omega^{-(n-1)(n-1)} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & \dots & \dots & 1 \\ \omega & \omega^2 & \dots & \omega^{n-1} & 1 \\ \omega^2 & \omega^4 & \dots & \omega^{2n-2} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ \omega^{n-1} & \omega^{n-2} & \dots & \dots & 1 \end{bmatrix} = \quad (6.288)$$

$$\frac{1}{n} \begin{bmatrix} 0 & 0 & \dots & \dots & 0 & n \\ n & 0 & \dots & \dots & 0 & 0 \\ 0 & n & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & n & 0 \end{bmatrix} = \mathbf{Z} \quad (6.289)$$

(for example the (1,1) element in the product in 6.288 =  $1 + \omega + \omega^2 + \dots + \omega^{n-1} = \frac{\omega^n - 1}{\omega - 1} = \frac{1-1}{\omega-1} = 0$ ; while the (1,n) element =  $1 + 1 + \dots + 1 = n$ ). Hence

$$\mathbf{F} = \mathbf{V}^H \hat{\mathbf{D}} \mathbf{V} + \mathbf{p} \mathbf{e}_n^T = \mathbf{V}^H (\hat{\mathbf{D}} + \mathbf{V} \mathbf{p} \mathbf{e}_n^T \mathbf{V}^H) \mathbf{V} \quad (6.290)$$

Also (again since  $\bar{\omega} = \omega^{-1}$ ), we have

$$\mathbf{e}_n^T \Omega^H = [1, \omega, \omega^2, \dots, \omega^{n-1}] \equiv \hat{\mathbf{v}}^T \quad (6.291)$$

(that is, we will call the above  $\hat{\mathbf{v}}^T$ ) and

$$\Omega \mathbf{p} = -\frac{1}{c_n} [p(1), p(\omega), \dots, p(\omega^{n-1})]^T \quad (6.292)$$

(we will call the vector in 6.292  $\hat{\mathbf{u}}^T$ )

Then 6.280 may be re-written as

$$\hat{\mathbf{A}} = \hat{\mathbf{D}} - \frac{1}{nc_n} \hat{\mathbf{u}} \hat{\mathbf{v}}^T \quad (6.293)$$

It follows that

$$\mathbf{F} = \mathbf{V}^H \hat{\mathbf{A}} \mathbf{V} \quad (6.294)$$

and so  $\mathbf{F}$  and  $\hat{\mathbf{A}}$  have the same eigenvalues, as claimed. Note that  $\hat{\mathbf{D}}$  has complex entries, but it is desirable, in order to achieve a fast solution, to have a matrix in the form of a **real** diagonal plus rank-one matrix (the latter not necessarily real). In fact  $\hat{\mathbf{D}}$  has entries on the unit circle, so we will use the Mobius transformation

$$M(z) = \frac{\delta z - \beta}{-\gamma z + \alpha}, \quad \alpha\delta - \beta\gamma \neq 0 \quad (6.295)$$

which for appropriate choices of the parameters maps the unit circle into the real axis. If  $\alpha \mathbf{I} - \gamma \hat{\mathbf{A}}$  is non-singular then  $\mathbf{A} = M(\hat{\mathbf{A}})$  has the required form i.e. real diagonal plus rank-one. Now the inverse of a Mobius transformation is also a Mobius transformation given by

$$M^{-1}(z) = \frac{\alpha z + \beta}{\gamma z + \delta} \quad (6.296)$$

It is shown by Van Barel et al (2005) in their Theorem 4.3 that if

$$\gamma = |\gamma|e^{i\theta_\gamma}, \delta = |\delta|e^{i\theta_\delta}, \alpha = |\alpha|e^{i\hat{\theta}} \text{ and } \beta = |\beta|e^{i\hat{\theta}} \quad (6.297)$$

where

$$\hat{\theta} = \tilde{\theta} + \theta_\gamma - \theta_\delta \quad (6.298)$$

then

$$M(z) = \frac{\delta z - \beta}{-\gamma z + \alpha} \quad (6.299)$$

maps the unit circle (except the point  $z = \frac{\alpha}{\gamma}$ ) onto the real axis. Assume that  $\alpha\mathbf{I} - \gamma\hat{\mathbf{A}}$  is non-singular and  $\omega^j \neq \frac{\alpha}{\gamma}$  ( $j = 0, \dots, n-1$ ) Then

$$M(\hat{\mathbf{A}}) = (\delta\hat{\mathbf{A}} - \beta\mathbf{I})(\alpha\mathbf{I} - \gamma\hat{\mathbf{A}})^{-1} \quad (6.300)$$

$$= [(\delta\hat{\mathbf{D}} - \beta\mathbf{I}) - \frac{\delta}{nc_n}\hat{\mathbf{u}}\hat{\mathbf{v}}^T][(\alpha\mathbf{I} - \gamma\hat{\mathbf{D}}) + \frac{\gamma}{nc_n}\hat{\mathbf{u}}\hat{\mathbf{v}}^T]^{-1} \quad (6.301)$$

By 6.245

$$[(\alpha\mathbf{I} - \gamma\hat{\mathbf{D}}) + \frac{\gamma}{nc_n}\hat{\mathbf{u}}\hat{\mathbf{v}}^T]^{-1} = (\alpha\mathbf{I} - \gamma\hat{\mathbf{D}})^{-1}(\mathbf{I} - \theta\hat{\mathbf{u}}\hat{\mathbf{v}}^T) \quad (6.302)$$

where

$$\theta = \frac{\gamma}{nc_n + \gamma\mathbf{v}^T\hat{\mathbf{u}}} \quad (6.303)$$

and

$$\mathbf{v} = (\alpha\mathbf{I} - \gamma\hat{\mathbf{D}})^{-1}\hat{\mathbf{v}} \quad (6.304)$$

Replacing the LHS of 6.302 by the RHS in 6.301 gives

$$M(\hat{\mathbf{A}}) = M(\hat{\mathbf{D}}) - \theta M(\hat{\mathbf{D}})\hat{\mathbf{u}}\hat{\mathbf{v}}^T - \frac{\delta}{nc_n}\hat{\mathbf{u}}\hat{\mathbf{v}}^T + \frac{\delta\theta}{nc_n}\hat{\mathbf{u}}(\mathbf{v}^T\hat{\mathbf{u}})\mathbf{v}^T \quad (6.305)$$

$$= M(\hat{\mathbf{D}}) - (\theta M(\hat{\mathbf{D}})\hat{\mathbf{u}} + \frac{\delta}{nc_n}\hat{\mathbf{u}} - \frac{\delta\theta}{nc_n}\hat{\mathbf{u}}(\mathbf{v}^T\hat{\mathbf{u}}))\mathbf{v}^T \quad (6.306)$$

Letting

$$\mathbf{u} = \theta M(\hat{\mathbf{D}})\hat{\mathbf{u}} + \frac{\delta}{nc_n}\hat{\mathbf{u}} - \frac{\delta\theta}{nc_n}\hat{\mathbf{u}}(\hat{\mathbf{v}}^T\hat{\mathbf{u}}) \quad (6.307)$$

Finally

$$\mathbf{A} = M(\hat{\mathbf{A}}) = \mathbf{D} - \mathbf{u}\mathbf{v}^T \quad (6.308)$$



is in the required form of a **real** diagonal plus rank-one matrix, where

$$\mathbf{D} = \text{diag}[M(1), M(\omega), \dots, M(\omega^{n-1})] \in R^{n \times n} \tag{6.309}$$

Each eigenvalue  $\eta_j$  of  $\mathbf{A}$  is related to the corresponding eigenvalue  $\lambda_j$  of  $\hat{\mathbf{A}}$  by

$$\eta_j = M(\lambda_j) \neq -\frac{\delta}{\gamma} \tag{6.310}$$

Once we have computed the  $\eta_j$  we may find the  $\lambda_j$  (roots of  $p(z)$ ) by

$$\lambda_j = \frac{\alpha\eta_j + \beta}{\gamma\eta_j + \delta}, \quad (j = 1, \dots, n) \tag{6.311}$$

We may summarize the above in the following algorithm called **FastRoots(p)** which outputs a vector  $\lambda$  of approximations to the roots of  $p(z)$ :

1. Evaluate  $p(z)$  at  $1, \omega, \dots, \omega^{n-1}$  where

$$\omega = \cos\frac{2\pi}{n} + \sqrt{-1}\sin\frac{2\pi}{n} \tag{6.312}$$

2. Compute the leading coefficient  $c_n$  of  $p(z)$  by 6.279
3. Form the vectors  $\hat{\mathbf{u}}$  and  $\hat{\mathbf{v}}^T$  by 6.291 and 6.292.
4. Choose random complex numbers  $\gamma$  and  $\delta$ .
5. Choose a random real number  $\tilde{\theta} \in [0, 1]$ .
6. Define  $\alpha$  and  $\beta$  by 6.297 and 6.298.
7. Compute  $(\mathbf{D})_{ii} = M(\omega^{i-1})$  for  $i=1, \dots, n$ .
8. Compute  $\mathbf{u}$  and  $\mathbf{v}$  by 6.307 and 6.304.
9. Compute approximations  $\eta_j$  of the eigenvalues of

$$\mathbf{D} - \mathbf{u}\mathbf{v}^T \tag{6.313}$$

10. Approximate the  $\lambda_j$  by 6.311.

The most time-consuming part of the above is step 9. Bini, Gemignani and Pan (2005) describe a method for finding eigenvalues of the type of matrix considered here (as well as others), which takes  $O(n^2)$  time. They summarize this method in their (2004b) paper, and we will reproduce their summary here. For more details see later (and their 2005 paper). They define a class of “generalized semi-separable matrices”  $C_n$  by stating that  $\mathbf{A} = (a_{ij})$  belongs to  $C_n$  if there exist real numbers  $d_1, \dots, d_n$ , complex numbers  $t_2, \dots, t_{n-1}$  and possibly complex vectors  $\mathbf{u} = [u_1, \dots, u_n]^T$ ,  $\mathbf{v} = [v_1, \dots, v_n]^T$ ,  $\mathbf{z} = [z_1, \dots, z_n]^T$  and  $\mathbf{w} = [w_1, \dots, w_n]^T$  such that

$$a_{ii} = d_i + z_i\bar{w}_i \quad (i = 1, \dots, n) \tag{6.314}$$

$$a_{ij} = u_i t_{ij}^\times \bar{v}_j \quad (i = 2, \dots, n; j = 1, \dots, i - 1) \tag{6.315}$$

$$a_{ij} = \bar{u}_j \bar{t}_{ji}^\times v_i + z_i \bar{w}_j - \bar{z}_j w_i \quad (j = 2, \dots, n; i = 1, \dots, j - 1) \tag{6.316}$$

where

$$t_{i,j}^\times = t_{i-1} \dots t_{j+1} \text{ for } i - 1 \geq j + 1 \tag{6.317}$$

and otherwise

$$t_{i,i-1}^\times = 1 \tag{6.318}$$

If we set  $\mathbf{z} = \mathbf{u}$ ,  $\mathbf{w} = \mathbf{v}$  and  $t_i = 1$  (all  $i$ ) we obtain the form  $\mathbf{D} + \mathbf{u}\mathbf{v}^H$  as required in the present work. The authors prove that the shifted QR algorithm preserves the structure of semi-separable matrices given by 6.314-6.316, and that an iteration can be performed in  $O(n)$  flops, so that the complete eigenvalue problem takes only  $O(n^2)$ .

Numerical tests were performed with the above algorithm on several difficult polynomials. Most results were accurate, except for the Wilkinson polynomial  $(z - 1)(z - 2) \dots (z - 19)(z - 20)$ . The case

$$p(z) = (z^n - 1) \left( c_n + \sum_{i=0}^{n-1} \frac{\omega_i p(\omega_i)}{n(z - \omega_i)} \right) \tag{6.319}$$

with  $p(\omega_i)$  and  $p(0)$  random complex numbers was particularly interesting: for  $n = 2^{2+m}$  with  $m = 1, \dots, 7$  the tests confirm that the time is indeed quadratic in  $n$ .

Returning to the 2005 paper, the authors define  $\text{triu}(\mathbf{B}, p) =$  the upper triangular part of  $\mathbf{B}$  formed by the elements on and above the  $p$ 'th diagonal of  $\mathbf{B}$  (i.e. the diagonal which is  $p$  positions above and to the right of the main diagonal). Similarly  $\text{tril}(\mathbf{B}, p)$  is formed by the elements on and **below** the  $p$ 'th diagonal. If  $\mathbf{A}$  is a matrix in the form of 6.314- 6.316, then

$$\text{tril}(\mathbf{A}, -1) = \begin{bmatrix} 0 & 0 & \dots & \dots & 0 \\ u_2 \bar{v}_1 & 0 & \dots & \dots & 0 \\ u_3 t_2 \bar{v}_1 & u_3 \bar{v}_2 & 0 & \dots & 0 \\ u_4 t_3 t_2 \bar{v}_1 & u_4 t_3 \bar{v}_2 & u_4 \bar{v}_3 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ u_n t_{n-1} \dots t_2 \bar{v}_1 & u_n t_{n-1} \dots t_3 \bar{v}_2 & \dots & u_n \bar{v}_{n-1} & 0 \end{bmatrix} \tag{6.320}$$

We also denote the above by

$$L(\{u_i\}_{i=2}^n, \{\bar{v}_i\}_{i=1}^{n-1}, \{t_i\}_{i=2}^{n-1}) \tag{6.321}$$

and also

$$R(\{\bar{u}_i\}_{i=2}^n, \{v_i\}_{i=1}^{n-1}, \{\bar{t}_i\}_{i=2}^{n-1}) \equiv (L(\{u_i\} \text{etc}))^H \tag{6.322}$$

Note that  $L(\{u_i\}_{i=2}^n, \{\bar{v}_i\}_{i=1}^{n-1}, 0)$  is a lower bidiagonal matrix with main diagonal 0 and sub-diagonal  $\eta_i = u_i \bar{v}_{i-1}$  ( $i = 2, \dots, n$ ). This is called *Subdiag*( $\{\eta_i\}_{i=2}^n$ ).

Let  $\mathbf{A}$  be of type 6.314-6.316 and denote  $\mathbf{x}_i = [z_i, w_i]$  and  $\mathbf{y}_i = [w_i, -z_i]$  ( $i = 1, \dots, n$ ). Then

$$\begin{aligned} \text{triu}(\mathbf{A}, 1) - R(\{\bar{u}_i\}_{i=2}^n, \{v_i\}_{i=1}^{n-1}, \{\bar{t}_i\}_{i=2}^{n-1}) = \\ \begin{bmatrix} 0 & \mathbf{x}_1 \mathbf{y}_2^H & \dots & \mathbf{x}_1 \mathbf{y}_n^H \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \mathbf{x}_{n-1} \mathbf{y}_n^H \\ 0 & \dots & \dots & 0 \end{bmatrix} \end{aligned} \tag{6.323}$$

We will be using Givens rotations such as

$$G(\gamma) = \frac{1}{\sqrt{1+|\gamma|^2}} \begin{bmatrix} 1 & \gamma \\ -\bar{\gamma} & 1 \end{bmatrix} \tag{6.324}$$

$$= \begin{bmatrix} \phi & \psi \\ -\bar{\psi} & \phi \end{bmatrix} \tag{6.325}$$

where  $\gamma$  and  $\psi$  are in general complex and  $\phi$  is real, while  $|\psi|^2 + |\phi|^2 = 1$ . We also have

$$G(\infty) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{6.326}$$

We can find  $\gamma$  so that  $G(\gamma)$  transforms a vector  $\begin{bmatrix} a \\ b \end{bmatrix}$  into  $\begin{bmatrix} \rho \\ 0 \end{bmatrix}$  where  $|\rho| = \sqrt{a^2 + b^2}$ . If  $a \neq 0$ , set  $\bar{\gamma} = \frac{b}{a}$ , else  $\gamma = \infty$ . Then we define the  $n \times n$  Givens rotation  $G_{k,k+1}(\gamma)$  in coordinates  $k, k+1$  by

$$\begin{bmatrix} \mathbf{I}_{k-1} & 0 & 0 \\ 0 & G(\gamma) & 0 \\ 0 & 0 & \mathbf{I}_{n-k-1} \end{bmatrix} \tag{6.327}$$

Recall that the QR iteration, which can be written

$$\mathbf{A}_{s+1} = \mathbf{R}_s \mathbf{A}_s \mathbf{R}_s^{-1} \tag{6.328}$$

where  $\mathbf{R}_s$  is right-triangular, yields  $\mathbf{A}_{s+1}$  tending to upper triangular or block upper-triangular form, so that the eigenvalues of  $\mathbf{A} = \mathbf{A}_0$  can be deduced. It can be proved that the structure of  $\mathbf{A}_0$  is preserved by the QR iterations: the authors first prove, in their theorem 3.1, that the structure of the lower triangular part is preserved. That is, each matrix satisfies

$$\text{tril}(\mathbf{A}_s, -1) = L(\{u_i^{(s)}\}_{i=2}^n, \{\bar{v}_i^{(s)}\}_{i=1}^{n-1}, \{t_i^{(s)}\}_{i=2}^{n-1}) \tag{6.329}$$

for suitable  $u_i^{(s)}, v_i^{(s)}, t_i^{(s)}$ . The proof is by induction on  $s$ . The case  $s = 0$  follows from the definition of  $\mathbf{A}_0$ . Assume that 6.329 holds for some  $u_i^{(s)}$  etc, and then prove the theorem for  $s+1$ . Let  $\mathbf{R}_s = (r_{ij}^{(s)})$ ,  $\mathbf{R}_s^{-1} = \mathbf{W}_s = (w_{ij}^{(s)})$  and  $\mathbf{A}_{s+\frac{1}{2}} = \mathbf{A}_s \mathbf{W}_s$ . The last matrix is obtained by linearly combining the columns of  $\mathbf{A}_s$ , i.e. it =

$$\begin{bmatrix} 0 & 0 & \dots & 0 \\ u_2^{(s)} \overline{v}_1^{(s)} & 0 & \dots & 0 \\ u_3^{(s)} t_2^{(s)} \overline{v}_1^{(s)} & u_3^{(s)} \overline{v}_2^{(s)} & 0 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \begin{bmatrix} w_{11}^{(s)} & w_{12}^{(s)} & \dots \\ 0 & w_{22}^{(s)} & \dots \\ 0 & 0 & \dots \end{bmatrix} \tag{6.330}$$

$$= \begin{bmatrix} 0 & 0 & \dots \\ u_2^{(s)} (w_{11}^{(s)} \overline{v}_1^{(s)}) & \dots & \dots \\ u_3^{(s)} t_2^{(s)} (w_{11}^{(s)} \overline{v}_1^{(s)}) & u_3^{(s)} [w_{12}^{(s)} t_2^{(s)} \overline{v}_1^{(s)} + w_{22}^{(s)} \overline{v}_2^{(s)}] & \dots \\ \dots & \dots & \dots \end{bmatrix} \tag{6.331}$$

Thus we may write

$$\text{tril}(\mathbf{A}_{s+\frac{1}{2}}, -1) = L(\{u_i^{(s+\frac{1}{2})}\}_{i=2}^n, \{\overline{v}_i^{(s+\frac{1}{2})}\}_{i=1}^{n-1}, \{t_i^{(s+\frac{1}{2})}\}_{i=2}^{n-1}) \tag{6.332}$$

where  $u_j^{(s+\frac{1}{2})} = u_j^{(s)}$ ,  $t_j^{(s+\frac{1}{2})} = t_j^{(s)}$  and  $\overline{v}_1^{(s+\frac{1}{2})} = w_{11}^{(s)} \overline{v}_1^{(s)}$  while

$$\overline{v}_j^{(s+\frac{1}{2})} = \sum_{k=2}^j w_{k-1,j}^{(s)} t_j^{(s)} \dots t_k^{(s)} \overline{v}_{k-1}^{(s)} + w_{jj}^{(s)} \overline{v}_j^{(s)} \quad (j = 2, \dots, n-1) \tag{6.333}$$

Similarly the rows of  $\mathbf{A}_{s+1} = \mathbf{R}_s \mathbf{A}_{s+\frac{1}{2}}$  are linear combinations of rows of  $\mathbf{A}_{s+\frac{1}{2}}$  and we get that

$$\text{tril}(\mathbf{A}_{s+1}, -1) = L(\{u_i^{(s+1)}\}_{i=2}^n, \{\overline{v}_i^{(s+1)}\}_{i=1}^{n-1}, \{t_i^{(s+1)}\}_{i=2}^{n-1}) \tag{6.334}$$

where  $v_j^{(s+1)} = v_j^{(s+\frac{1}{2})}$  (given by 6.333),  $t_j^{(s+1)} = t_j^{(s)}$ , and  $u_n^{(s+1)} = r_{nn}^{(s)} u_n^{(s)}$ , while

$$u_{n-j}^{(s+1)} = \sum_{k=0}^{j-1} r_{n-j,n-k}^{(s)} t_{n-j}^{(s)} \dots t_{n-k-1}^{(s)} u_{n-k}^{(s)} + r_{n-j,n-j}^{(s)} u_{n-j}^{(s)} \tag{6.335}$$

$(j = 1, \dots, n-2)$

Next, we can prove that each  $\mathbf{A}_s$  is of the form 6.314- 6.316, i.e.

$$a_{ii}^{(s)} = d_i^{(s)} + z_i^{(s)} \overline{w}_i^{(s)} \quad (i = 1, \dots, n) \tag{6.336}$$

$$a_{ij}^{(s)} = u_i^{(s)} t_{ij}^{(s)} \times \overline{v}_j^{(s)} \quad (i = 2, \dots, n; j = 1, \dots, i-1) \tag{6.337}$$

$$a_{ij}^{(s)} = \overline{u}_j^{(s)} \overline{t}_{ji}^{(s)} \times v_i^{(s)} + z_i^{(s)} \overline{w}_j^{(s)} - \overline{z}_j^{(s)} w_i^{(s)} \quad (j = 2, \dots, n; i = 1, \dots, j-1) \tag{6.338}$$

For 6.337 has already been proved (it is the same as 6.334), and we may show that  $\mathbf{B} \equiv \mathbf{A} - \mathbf{z}\mathbf{w}^H$  is Hermitian so that

$$a_{ij}^{(s)} - z_i^{(s)}\bar{w}_j^{(s)} = \bar{a}_{ji}^{(s)} - \bar{z}_j^{(s)}w_i^{(s)} \tag{6.339}$$

For  $i < j$  we have by 6.337 with  $i, j$  reversed that

$$a_{ji}^{(s)} = u_j^{(s)}\bar{t}_{ji}^{(s)}\bar{v}_i^{(s)}; \bar{a}_{ji}^{(s)} = \bar{u}_j^{(s)}\bar{t}_{ji}^{(s)}v_i^{(s)} \tag{6.340}$$

Substituting in 6.339 gives  $a_{ij}^{(s)} = \bar{u}_j^{(s)}\bar{t}_{ji}^{(s)}v_i^{(s)} + z_i^{(s)}\bar{w}_j^{(s)} - \bar{z}_j^{(s)}w_i^{(s)}$ , which is 6.338. For  $i = j$  we can deduce from 6.339 that the imaginary part of  $a_{ii}^{(s)}$  coincides with that of  $z_i^{(s)}\bar{w}_i^{(s)}$  (for if  $a_{ii}^{(s)} = R + iI$  and  $z_i^{(s)}\bar{w}_i^{(s)} = \rho + i\sigma$  we have  $R + iI = R - iI + (\rho + i\sigma) - (\rho - i\sigma)$ ; hence  $2iI = 2i\sigma$ ). So  $a_{ii}^{(s)} = d_i^{(s)} + z_i^{(s)}\bar{w}_i^{(s)}$ , which is 6.336.

We have also

$$\begin{aligned} \mathbf{A}_{s+1} &= \mathbf{P}_s^H \mathbf{A}_0 \mathbf{P}_s = \mathbf{P}_s^H (\mathbf{B}_0 + \mathbf{z}^{(0)}\mathbf{w}^{(0)H}) \mathbf{P}_s = \\ &= \mathbf{P}_s^H \mathbf{B}_0 \mathbf{P}_s + \mathbf{z}^{(s+1)}\mathbf{w}^{(s+1)H} \end{aligned} \tag{6.341}$$

which shows that

$$\mathbf{z}^{(s+1)} = \mathbf{P}_s^H \mathbf{z}^{(0)} = \mathbf{Q}_s^H \mathbf{z}^{(s)} \tag{6.342}$$

(where  $\mathbf{P}_s = \mathbf{Q}_0\mathbf{Q}_1\dots\mathbf{Q}_s$ ) and

$$\mathbf{w}^{(s+1)H} = \mathbf{w}^{(0)H} \mathbf{P}_s = \mathbf{w}^{(s)H} \mathbf{Q}_s \tag{6.343}$$

which give easy rules for updating  $\mathbf{z}^{(s)}$  and  $\mathbf{w}^{(s)}$  at each QR step. If, for a certain index  $\hat{s}$ ,  $\mathbf{R}_{\hat{s}}$  and  $\mathbf{A}_{\hat{s}} - \sigma_{\hat{s}}\mathbf{I}_n$  are singular (or nearly so), then  $\sigma_{\hat{s}}$  is an eigenvalue of  $\mathbf{A}_0$ , and we may deflate.

We need an efficient procedure to derive  $\mathbf{A}_1$  from  $\mathbf{A}_0$  and so on by the QR factorization. Using the structure of  $\text{tril}(\mathbf{A}_0, -1)$  we may express  $\mathbf{Q}_0$  as the product of  $2n-3$  Givens rotations (compared with  $O(n^2)$  for a general matrix). First we reduce  $\mathbf{A}_0$  to upper Hessenberg form

$$\mathbf{H}_0 = G_{2,3}(\gamma_{n-2})\dots G_{n-1,n}(\gamma_1)\mathbf{A}_0 = \hat{\mathbf{Q}}_0\mathbf{A}_0 \tag{6.344}$$

Then we reduce  $\mathbf{H}_0$  to upper triangular form

$$\mathbf{R}_0 = G_{n-1,n}(\gamma_{2n-3})\dots G_{12}(\gamma_{n-1})\mathbf{H}_0 \tag{6.345}$$

Thus

$$\begin{aligned} \mathbf{R}_0 &= G_{n-1,n}(\gamma_{2n-3})\dots G_{12}(\gamma_{n-1})G_{23}(\gamma_{n-2})\dots G_{n-1,n}(\gamma_1)\mathbf{A}_0 = \\ &= \mathbf{Q}_0^H \mathbf{A}_0 \end{aligned} \tag{6.346}$$

where

$$\mathbf{Q}_0^H = G_{n-1,n}(\gamma_{2n-3}) \dots G_{n-1,n}(\gamma_1) \quad (6.347)$$

The Givens matrices are chosen to zero the various entries of  $\text{tril}(\mathbf{A}_0, -2)$ ; e.g.  $\gamma_1$  can be chosen so that

$$G(\gamma_1) \begin{bmatrix} u_{n-1}^{(0)} \\ u_n^{(0)} t_{n-1}^{(0)} \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{u}_{n-1}^{(0)} \\ 0 \end{bmatrix} \quad (6.348)$$

Thus, because of the special structure (see 6.320), the whole of the  $n$ 'th row through the  $(n, n-2)$  element is zeroed. The general case  $G(\gamma_i)$  is similar. Consequently the reduction takes  $O(n)$  flops, in contrast to the  $O(n^2)$  required for a general matrix. The same is true for the reduction to upper triangular form. In fact the authors show that each QR step requires  $120n$  multiplications and  $28n$  storage. For further details see the cited paper.

Numerical tests were performed, firstly on arrowhead matrices of order  $n = 2^s$  for  $s = 3, \dots, 8$ ; then Hermitian diagonal-plus-semiseparable matrices; and finally on the Chebyshev-comrade matrices of order  $n$ . The latter problem is related to the task of finding roots of a polynomial represented as a series of Chebyshev polynomials given by

$$p_0(z) = 1; p_j(z) = w^j + \left(\frac{1}{2w}\right)^j \quad (j = 1, 2, \dots) \quad (6.349)$$

where

$$z = w + \frac{1}{2w} \quad (6.350)$$

In the last-mentioned problem the coefficients were random complex values with real and imaginary parts in the range  $[-1, 1]$ . The tests confirm that the time is  $O(n^2)$ , the error is very small, and about 6 iterations are required per eigenvalue.

## 6.4 Methods Designed for Multiple Roots

Some of the methods described in previous sections of this Chapter work (to a certain extent) for multiple roots, but not as well as they do for simple roots. In contrast the methods to be described in the present section are designed specifically to work accurately in the case of multiple roots. The first is due to Zeng (2003, 2004a, 2004b). We will follow the treatment in (2004b). Zeng presents a combination of two algorithms for computing multiple roots and multiplicity structures (i.e. a list of the number of times each distinct root is repeated). It accurately calculates polynomial roots of high multiplicity without using multiprecision arithmetic (as is usually required), even if the coefficients are inexact. This is the first work to do that, and is a remarkable achievement. Traditionally it has been believed that

there is an “attainable accuracy” in computing multiple roots: i.e. to compute an  $m$ -fold root correct to  $k$  digits requires a precision of  $mk$  digits in the coefficients and machine numbers—hence the need for multiple precision in the common case that  $mk > 16$ . Even worse, when coefficients are truncated (as is usual), multiple roots are turned into clusters, and no amount of multiple precision will turn them back into multiple roots.

However Kahan (1972) proved that if multiplicities are preserved, the roots may be well-behaved, i.e. not nearly as hypersensitive as they would otherwise be. Polynomials with a fixed multiplicity structure are said to form a **pejorative manifold**. For a polynomial on such a manifold multiple roots are insensitive to perturbations which preserve the multiplicity structure, unless it is near a submanifold of higher multiplicities.

In light of the above, Zeng proposes his Algorithm I (see below) that transforms the singular root-finding problem into a regular non-linear least squares problem on a pejorative manifold. To apply this algorithm, we need initial root approximations as well as knowledge of the multiplicity structure. To accomplish this, Zeng proposes a numerical GCD-finder (for the GCD  $u$  of  $p$  and  $p'$ ) which uses a series of Sylvester matrices. It finds the smallest singular value of each of these matrices, and extracts the degree of  $u$  and the coefficients of the GCD decomposition ( $v$  and  $w$ , where  $p = uv$  and  $p' = uw$ ). Finally it applies the Gauss-Newton iteration to refine the approximate GCD. This GCD-finder constitutes the main part of his Algorithm II, which computes the multiplicity structure and initial root approximations.

While most reported numerical experiments do not even reach multiplicity 10, Zeng successfully tested his algorithms on polynomials with root multiplicities up to 400 without using multiprecision arithmetic. To quote him: “We are aware of no other reliable methods that calculate multiple roots accurately by using standard machine precision”. On the question of speed, there exist general-purpose root-finders using  $O(n^2)$  flops, such as those discussed in Section 3 of this Chapter. But the barrier of “attainable accuracy” may prevent these from calculating multiple roots accurately when the coefficients are inexact, even if multiple precision arithmetic is used. Zeng’s methods overcome this barrier at a cost of  $O(n^3)$  flops in standard arithmetic. This may not be too high a price (in fact for moderate  $n$  it may be faster than an  $O(n^2)$  method using multiple precision).

We will now discuss some preliminary material, before describing Algorithm I and II in detail. The numbers of Lemmas etc will be as in Zeng’s paper. If the polynomial

$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_0 \quad (6.351)$$

then the same letter in boldface (e.g.  $\mathbf{p}$ ) denotes the coefficient vector

$$\mathbf{p} = [c_n \ c_{n-1} \ \dots \ c_0]^T \quad (6.352)$$

The degree  $n$  of  $p(x)$  is called  $\text{deg}(p)$ . For a pair of polynomials  $p$  and  $q$ , their greatest common divisor is called  $\text{GCD}(p,q)$ .

**Definition 2.1.** For any integer  $k \geq 0$ , we define the matrix

$$\mathbf{C}_k(p) = \begin{bmatrix} c_n & 0 & \dots & \dots \\ c_{n-1} & c_n & 0 & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & c_n \\ c_0 & \dots & \dots & c_{n-1} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & c_0 \end{bmatrix} \tag{6.353}$$

(it is assumed that the above matrix has  $k+1$  columns and hence  $n+k+1$  rows). It is called the  $k$ 'th order **convolution matrix** associated with  $p(x)$ .

**Lemma 2.2.** Let  $f$  and  $g$  be polynomials of degree  $n$  and  $m$  respectively, with

$$h(x) = f(x)g(x) \tag{6.354}$$

Then  $\mathbf{h}$  is the **convolution** of  $\mathbf{f}$  and  $\mathbf{g}$  defined by

$$\mathbf{h} = \text{conv}(\mathbf{f},\mathbf{g}) = \mathbf{C}_m(\mathbf{f})\mathbf{g} = \mathbf{C}_n(\mathbf{g})\mathbf{f} \tag{6.355}$$

**Proof.** We see that for example from 6.354

$$h_{n+m} = f_n g_m; h_{m+n-1} = f_{n-1} g_m + f_n g_{m-1} \tag{6.356}$$

etc until

$$h_n = f_{n-m} g_m + f_{n-m+1} g_{m-1} + \dots + f_n g_0 \tag{6.357}$$

and so on. This agrees with 6.355.

**Definition 2.3.** Let  $p'$  be the derivative of  $p$ ; then for  $k = 1, 2, \dots, n-1$  the matrix of size  $(n+k) \times (2k+1)$

$$\mathbf{S}_k(p) = [ \mathbf{C}_k(p') \mid \mathbf{C}_{k-1}(p) ] \tag{6.358}$$

is called the  $k$ 'th Sylvester discriminant matrix.

**Lemma 2.4.** With  $p$  and  $p'$  as before, let  $u = \text{GCD}(p, p')$ . For  $j = 1, \dots, n$ , let  $\sigma_j$  be the smallest singular value of  $\mathbf{S}_j(p)$ . Then the following are equivalent:

- (a)  $\text{deg}(u) = k$ ,
- (b)  $p$  has  $m = n-k$  distinct roots,
- (c)  $\sigma_1, \sigma_2, \dots, \sigma_{m-1} > 0, \sigma_m = \sigma_{m+1} = \dots = \sigma_n = 0$ ,

**Proof** that (a) is equivalent to (b): Assume that  $p(x)$  has  $m$  distinct roots of multiplicities  $l_1, \dots, l_m$ ; then we have

$$p(x) = (x - \zeta_1)^{l_1} (x - \zeta_2)^{l_2} \dots (x - \zeta_m)^{l_m} \tag{6.359}$$

where  $l_i \geq 1$  ( $i = 1, \dots, m$ ) and  $\sum_{i=1}^m l_i = n$ . Then

$$\text{GCD}(p, p') \equiv u = (x - \zeta_1)^{l_1-1} \dots (x - \zeta_m)^{l_m-1} \tag{6.360}$$



Hence  $\text{deg}(u) = k = l_1 - 1 + l_2 - 1 + \dots + l_m - 1 = \sum l_i - m = n - m$ , so  $m = n - k$ . For the proof that (a) is equivalent to (c) Zeng refers to Rupprecht (1999), Proposition 3.1.

**Lemma 2.5.** Let  $p, p', u$  and  $k$  be as before. Let  $v$  and  $w$  be polynomials that satisfy

$$u(x)v(x) = p(x); u(x)w(x) = p'(x) \tag{6.361}$$

Then (a)  $v$  and  $w$  are coprime (i.e. they have no common factors);

(b) the column rank of  $\mathbf{S}_m(p)$  is deficient by one;

(c) the normalized vector  $\begin{bmatrix} \mathbf{v} \\ -\mathbf{w} \end{bmatrix}$  is the right singular vector of  $\mathbf{S}_m(p)$  associated with its smallest singular value  $\sigma_m$  (which is zero);

(d) if  $\mathbf{v}$  is known, the coefficient vector  $\mathbf{u}$  of  $u(x) = \text{GCD}(p, p')$  is the solution to the linear system

$$\mathbf{C}_m(v)\mathbf{u} = \mathbf{p} \tag{6.362}$$

**Proof.** (a) follows by definition of the GCD: if  $v$  and  $w$  had a common factor it would be included in  $u$ . Now

$$\mathbf{S}_m(p) \begin{bmatrix} \mathbf{v} \\ -\mathbf{w} \end{bmatrix} = \mathbf{C}_m(p')\mathbf{v} - \mathbf{C}_{m-1}(p)\mathbf{w} = \mathbf{0} \tag{6.363}$$

because it is the coefficient vector of  $p'v - pw \equiv (uw)v - (uv)w \equiv 0$ . Let  $\hat{v}$  of size  $m+1$  and  $\hat{w}$  of size  $m$  be two other coefficient vectors of two other polynomials  $\hat{v}$  and  $\hat{w}$  which also satisfy

$$\mathbf{C}_m(p')\hat{v} - \mathbf{C}_{m-1}(p)\hat{w} = \mathbf{0} \tag{6.364}$$

Then we also have  $(uw)\hat{v} - (uv)\hat{w} = 0$ , so that  $w\hat{v} = v\hat{w}$  (as  $u$  cannot be trivial; at worst it = 1). Hence, since  $v$  and  $w$  are coprime, the factors of  $v$  must also be factors of  $\hat{v}$ , so  $\hat{v} = cv$  for some polynomial  $c$ . But  $v$  and  $\hat{v}$  have the same degree  $m (=n-k)$  so  $c$  must be a constant. Also  $\hat{w} = w(\frac{\hat{v}}{v}) = cw$ . Therefore the single vector  $\begin{bmatrix} \mathbf{v} \\ -\mathbf{w} \end{bmatrix}$  forms the basis for the null-space of  $\mathbf{S}_m(p)$ . Consequently, both parts (b) and (c) follow. (d) follows from Lemma 2.2 with  $h$  replaced by  $p$ ,  $f$  by  $v$ , and  $g$  by  $u$ .

**Lemma 2.6 .** Let  $\mathbf{A}$  be an  $n \times k$  matrix with  $n \geq k$  having two smallest singular values  $\hat{\sigma} > \tilde{\sigma}$ . Let  $\mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{A}$  be the QR decomposition of  $\mathbf{A}$ , where  $\mathbf{Q}$  is  $n \times n$  and unitary, and  $\mathbf{R}$  is  $k \times k$  and upper triangular. From any complex vector  $\mathbf{x}_0$  that is not orthogonal to the right singular subspace associated with  $\tilde{\sigma}$ , we generate the sequences  $\{s_j\}$  and  $\{\mathbf{x}_j\}$  by the inverse iteration: Solve

$$\mathbf{R}^H \mathbf{y}_j = \mathbf{x}_{j-1} \text{ (} \mathbf{y}_j \text{ complex and size } k \text{)} \tag{6.365}$$

Solve

$$\mathbf{R}\mathbf{z}_j = \mathbf{y}_j \quad (\mathbf{z}_j \text{ complex and size } k) \quad (6.366)$$

Calculate

$$\mathbf{x}_j = \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}, \quad s_j = \|\mathbf{R}\mathbf{x}_j\|_2 \quad (6.367)$$

Then

$$\lim_{j \rightarrow \infty} s_j = \lim_{j \rightarrow \infty} \|\mathbf{A}\mathbf{x}_j\|_2 = \tilde{\sigma} \quad (6.368)$$

and

$$s_j = \|\sigma_j\|_2 + O(\tau^2)$$

where

$$\tau = \left( \frac{\tilde{\sigma}}{\hat{\sigma}} \right)^2 \quad (6.369)$$

and if  $\tilde{\sigma}$  is simple,  $\mathbf{x}_j \rightarrow$  the right singular vector of  $\mathbf{A}$  associated with  $\tilde{\sigma}$ . Zeng refers to Van Huffel (1991) for a proof.

He next discusses the Gauss-Newton iteration as a method for solving non-linear least squares problems. Let

$$G(\mathbf{z}) = \mathbf{a} \quad (6.370)$$

where  $\mathbf{a}$ ,  $\mathbf{z}$  are of size  $n$ ,  $m$  ( $n > m$ ). This is an over-determined system so we seek a weighted least squares solution. If

$$\mathbf{W} = \begin{bmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \omega_n \end{bmatrix} \quad (\omega_i > 0) \quad (6.371)$$

and for  $\mathbf{v}$  any vector of size  $n$ ,

$$\|\mathbf{v}\|_W = \|\mathbf{W}\mathbf{v}\|_2 = \sqrt{\sum_{j=1}^n \omega_j^2 v_j^2} \quad (6.372)$$

then our objective is to find

$$\min_{\mathbf{z}} \|G(\mathbf{z}) - \mathbf{a}\|_W^2 \quad (6.373)$$

**Lemma 2.7.** Let  $F$  be an analytic function from  $C^m$  to  $C^n$ , having Jacobian  $\hat{\mathbf{J}}(\mathbf{z})$ . If there is a neighborhood  $\Omega$  of  $\tilde{\mathbf{z}}$  in  $C^m$  such that

$$\|F(\tilde{\mathbf{z}})\|_2 \leq \|F(\mathbf{z})\|_2 \quad (6.374)$$

for all  $\mathbf{z} \in \Omega$  then

$$\hat{\mathbf{J}}^H F(\bar{\mathbf{z}}) = \mathbf{0} \quad (6.375)$$

For proof Zeng quotes Dennis and Schnabel (1983) for the real case, and states that the complex case is identical except for using the Cauchy-Riemann equation. Now let  $\mathbf{J}(\mathbf{z})$  be the Jacobian of  $G(\mathbf{z})$ . To find a local minimum of  $\|F(\mathbf{z})\|_2 \equiv \|\mathbf{W}[G(\mathbf{z}) - \mathbf{a}]\|_2$  with  $\hat{\mathbf{J}}(\mathbf{z}) = \mathbf{W}\mathbf{J}(\mathbf{z})$ , we seek  $\bar{\mathbf{z}}$  of size  $m$  such that

$$\begin{aligned} \hat{\mathbf{J}}(\bar{\mathbf{z}})^H F(\bar{\mathbf{z}}) &= [\mathbf{W}\mathbf{J}(\bar{\mathbf{z}})]^H \mathbf{W}[G(\bar{\mathbf{z}}) - \mathbf{a}] = \\ \mathbf{J}(\bar{\mathbf{z}})^H \mathbf{W}^2[G(\bar{\mathbf{z}}) - \mathbf{a}] &= \mathbf{0} \end{aligned} \quad (6.376)$$

**Lemma 2.8.** Let  $\Omega \in C^m$  be a bounded open convex set and let  $F$  (a function from  $C^m$  to  $C^n$ ) be analytic in an open set  $D \supset \bar{\Omega}$ . Let  $\hat{\mathbf{J}}$  be the Jacobian of  $F(\mathbf{z})$ . Suppose there exist  $\bar{\mathbf{z}} \in \Omega$  such that

$$\hat{\mathbf{J}}(\bar{\mathbf{z}})^H F(\bar{\mathbf{z}}) = \mathbf{0} \quad (6.377)$$

with  $\hat{\mathbf{J}}(\bar{\mathbf{z}})$  of full rank. Let  $\sigma$  be the smallest singular value of  $\hat{\mathbf{J}}(\bar{\mathbf{z}})$ , and  $\delta > 0$  be such that

$$\|[\hat{\mathbf{J}}(\mathbf{z}) - \hat{\mathbf{J}}(\bar{\mathbf{z}})]^H F(\bar{\mathbf{z}})\|_2 \leq \delta \|\mathbf{z} - \bar{\mathbf{z}}\|_2 \quad (6.378)$$

for all  $\mathbf{z} \in \Omega$

If  $\delta < \sigma^2$ , then for any  $c \in [\frac{1}{\sigma}, \frac{\sigma}{\delta}]$  there exists  $\epsilon > 0$  such that for all  $\mathbf{z}_0 \in \Omega$  with  $\|\mathbf{z}_0 - \bar{\mathbf{z}}\|_2 < \epsilon$ , the sequence generated by the Gauss-Newton iteration

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \hat{\mathbf{J}}(\mathbf{z}_k)^+ F(\mathbf{z}_k) \quad (6.379)$$

where

$$\hat{\mathbf{J}}(\mathbf{z}_k)^+ = [\hat{\mathbf{J}}(\mathbf{z}_k)^H \hat{\mathbf{J}}(\mathbf{z}_k)]^{-1} \hat{\mathbf{J}}(\mathbf{z}_k)^H \quad (6.380)$$

for  $k = 0, 1, \dots$  is well-defined inside  $\Omega$ , converges to  $\bar{\mathbf{z}}$ , and satisfies

$$\|\mathbf{z}_{k+1} - \bar{\mathbf{z}}\|_2 \leq \frac{c\delta}{\sigma} \|\mathbf{z}_k - \bar{\mathbf{z}}\|_2 + \frac{c\alpha\gamma}{2\sigma} \|\mathbf{z}_k - \bar{\mathbf{z}}\|_2^2 \quad (6.381)$$

where  $\alpha > 0$  is the upper bound of  $\|\hat{\mathbf{J}}(\mathbf{z})\|_2$  on  $\bar{\Omega}$ , and  $\gamma > 0$  is the Lipschitz constant of  $\hat{\mathbf{J}}(\mathbf{z})$  in  $\Omega$ , i.e.

$$\|\hat{\mathbf{J}}(\mathbf{z} + \mathbf{h}) - \hat{\mathbf{J}}(\mathbf{z})\|_2 \leq \gamma \|\mathbf{h}\|_2 \quad (6.382)$$

for all  $\mathbf{z}, \mathbf{z} + \mathbf{h} \in \Omega$ .

**Proof:** Zeng refers to Dennis and Schnabel (Theorem 10.2.1) for the real case, and states that the complex case is a “straightforward generalization” of the real case.

We now turn to the details of Algorithm I. It assumes that the multiplicity structure is known: we shall deal with the problem of finding this later. A condition number will be introduced to measure the sensitivity of multiple roots; when this is moderate the multiple roots can be calculated accurately. A polynomial of degree  $n$  corresponds to a (complex) vector of size  $n$ :

$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_0 \sim \mathbf{a} = [a_{n-1}, \dots, a_0]^T = \left[ \frac{c_{n-1}}{c_n}, \dots, \frac{c_0}{c_n} \right]^T \tag{6.383}$$

For a partition of  $n$ , i.e. an array of positive integers  $l_1, l_2, \dots, l_m$  with  $l_1 + l_2 + \dots + l_m = n$ , a polynomial  $p$  that has roots  $\zeta_1, \dots, \zeta_m$  with multiplicities  $l_1, \dots, l_m$  can be written as

$$\frac{1}{c_n} p(x) = \prod_{j=1}^m (x - \zeta_j)^{l_j} = x^n + \sum_{j=1}^n g_{n-j}(\zeta_1, \dots, \zeta_m) x^{n-j} \tag{6.384}$$

where each  $g_j$  is a polynomial in  $\zeta_1, \dots, \zeta_m$ . We have the correspondence

$$p \sim G_l(\mathbf{z}) \equiv \begin{bmatrix} g_{n-1}(\zeta_1, \dots, \zeta_m) \\ g_{n-2}(\zeta_1, \dots, \zeta_m) \\ \dots \\ g_0(\zeta_1, \dots, \zeta_m) \end{bmatrix} \text{ where } \mathbf{z} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \dots \\ \zeta_m \end{bmatrix} \tag{6.385}$$

**Definition 3.1.** An ordered array of positive integers  $l = [l_1, \dots, l_m]$  is called a **multiplicity structure** of degree  $n$  if  $l_1 + \dots + l_m = n$ . For given  $l$ , the collection of vectors  $\Pi_l = \{G_l(\mathbf{z}) | \mathbf{z} \text{ is of size } m\}$  is called the **pejorative manifold** of  $l$ , and  $G_l$  is called the **coefficient operator** associated with  $l$ . For example consider polynomials of degree 3. Firstly, for  $l = [1, 2]$  we have  $(x - \zeta_1)(x - \zeta_2)^2 = x^3 + (-\zeta_1 - 2\zeta_2)x^2 + (2\zeta_1\zeta_2 + \zeta_2^2)x + (-\zeta_1\zeta_2^2)$ , i.e. a polynomial with one simple root  $\zeta_1$  and one double root  $\zeta_2$  corresponds to

$$G_{[1,2]}(\mathbf{z}) = \begin{bmatrix} -\zeta_1 - 2\zeta_2^2 \\ 2\zeta_1\zeta_2 + \zeta_2^2 \\ -\zeta_1\zeta_2^2 \end{bmatrix}, \mathbf{z} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

The vectors  $G_{[1,2]}(\mathbf{z})$  for all  $\mathbf{z}$  form the pejorative manifold  $\Pi_{[1,2]}$ . Similarly

$$\Pi_{[3]} = \{(-3\zeta, 3\zeta^2, -\zeta^3) | \zeta \in C\}$$

when  $l = [3]$ .  $\Pi_{[3]}$  is a submanifold of  $\Pi_{[1,2]}$  that contains all polynomials with a single triple root.  $\Pi_{[1,1,\dots,1]} = C^n$  is the vector space of **all** polynomials of degree  $n$ .

We now consider methods of solving the least-squares problem. If  $l = [l_1, \dots, l_m]$  is a multiplicity structure of degree  $n$ , with the corresponding pejorative manifold

$\Pi_l$ , and the polynomial  $p \sim \mathbf{a} \in \Pi_l$ , then there is a vector  $\mathbf{z} \in C^m$  such that  $G_l(\mathbf{z}) = \mathbf{a}$ , i.e.

$$\begin{bmatrix} g_{n-1}(\zeta_1, \dots, \zeta_m) \\ g_{n-2}(\zeta_1, \dots, \zeta_m) \\ \dots \\ g_0(\zeta_1, \dots, \zeta_m) \end{bmatrix} = \begin{bmatrix} a_{n-1} \\ \dots \\ a_0 \end{bmatrix} \text{ or } \mathbf{G}_l(\mathbf{z}) = \mathbf{a} \tag{6.386}$$

In general this system is over-determined except for  $l = [1, 1, \dots, 1]$ . Let  $\mathbf{W} = \text{diag}(\omega_0, \dots, \omega_{n-1})$  as in 6.371 (with a change of numbering of the  $\omega_i$ ) and let  $\|\cdot\|_W$  denote the weighted 2-norm defined in 6.372. We seek a weighted least-squares solution to 6.386 by solving the minimization problem

$$\begin{aligned} \text{Min}_{\mathbf{z} \in C^m} \|\mathbf{G}_l(\mathbf{z}) - \mathbf{a}\|_W^2 &\equiv \text{Min}_{\mathbf{z} \in C^m} \|\mathbf{W}(\mathbf{G}_l(\mathbf{z}) - \mathbf{a})\|_2^2 \equiv \\ \text{Min}_{\mathbf{z} \in C^m} \left\{ \sum_{j=0}^{n-1} \omega_j^2 |G_j(\mathbf{z}) - a_j|^2 \right\} &\tag{6.387} \end{aligned}$$

Zeng does all his experiments with the weights

$$\omega_j = \text{Min}\left\{1, \frac{1}{|a_j|}\right\}, \quad (j = 0, \dots, n - 1) \tag{6.388}$$

which minimizes the relative backward error at every coefficient greater than one. Let  $\mathbf{J}(\mathbf{z})$  be the Jacobian of  $G_l(\mathbf{z})$ . To find a local minimum of  $F(\mathbf{z}) = \mathbf{W}[G_l(\mathbf{z}) - \mathbf{a}]$  with  $\hat{\mathbf{J}}(\mathbf{z}) = \mathbf{W}\mathbf{J}(\mathbf{z})$ , we look for  $\tilde{\mathbf{z}} \in C^m$  such that

$$\begin{aligned} \hat{\mathbf{J}}(\tilde{\mathbf{z}})^H F(\tilde{\mathbf{z}}) &= [\mathbf{W}\mathbf{J}(\tilde{\mathbf{z}})]^H \mathbf{W}[G_l(\tilde{\mathbf{z}}) - \mathbf{a}] = \\ \mathbf{J}(\tilde{\mathbf{z}})^H \mathbf{W}^2[G_l(\tilde{\mathbf{z}}) - \mathbf{a}] &= \mathbf{0} \end{aligned} \tag{6.389}$$

**Definition 3.2.** Let  $p \sim \mathbf{a}$  be a polynomial of degree  $n$ . For any  $l$  also of degree  $n$ , the vector  $\tilde{\mathbf{z}}$  satisfying 6.389 is called a **pejorative root** of  $p$  corresponding to  $l$  and  $\mathbf{W}$ .

**Theorem 3.3** Let  $G_l : C^m \rightarrow C^n$  be the coefficient operator associated with a multiplicity structure  $l = [l_1, \dots, l_m]$ . Then the Jacobian  $\mathbf{J}(\mathbf{z})$  of  $G_l(\mathbf{z})$  is of full rank if and only if the components of  $\mathbf{z} = [\zeta_1, \dots, \zeta_m]^T$  are distinct.

**Proof.** Let  $\zeta_1, \dots, \zeta_m$  be distinct. We seek to show that the columns of  $\mathbf{J}(\mathbf{z})$  are linearly independent. Write the  $j$ 'th column as

$$J_j = \left[ \frac{\partial g_{n-1}(\mathbf{z})}{\partial \zeta_j}, \dots, \frac{\partial g_0(\mathbf{z})}{\partial \zeta_j} \right]^T \tag{6.390}$$

For  $j = 1, \dots, m$  let  $q_j(x)$ , a polynomial in  $x$ , be defined by

$$q_j(x) = \left(\frac{\partial g_{n-1}(\mathbf{z})}{\partial \zeta_j}\right)x^{n-1} + \dots + \left(\frac{\partial g_1(\mathbf{z})}{\partial \zeta_j}\right)x + \left(\frac{\partial g_0(\mathbf{z})}{\partial \zeta_j}\right) \tag{6.391}$$

$$= \frac{\partial}{\partial \zeta_j} [x^n + g_{n-1}(\mathbf{z})x^{n-1} + \dots + g_0(\mathbf{z})] \quad (6.392)$$

$$= \frac{\partial}{\partial \zeta_j} [(x - \zeta_1)^{l_1} \dots (x - \zeta_m)^{l_m}] \quad (6.393)$$

$$= -l_j (x - \zeta_j)^{l_j-1} \left[ \prod_{k \neq j} (x - \zeta_k)^{l_k} \right] \quad (6.394)$$

Assume that

$$c_1 J_1 + \dots + c_m J_m = 0 \quad (6.395)$$

for constants  $c_1, \dots, c_m$ . Then

$$\begin{aligned} q(x) &\equiv c_1 q_1(x) + \dots + c_m q_m(x) = - \sum_{j=1}^m \{c_j l_j (x - \zeta_j)^{l_j-1} \prod_{k \neq j} (x - \zeta_k)^{l_k}\} \\ &= - \left[ \prod_{\sigma=1}^m (x - \zeta_\sigma)^{l_\sigma-1} \right] \sum_{j=1}^m [c_j l_j \prod_{k \neq j} (x - \zeta_k)] \end{aligned} \quad (6.396)$$

is a zero polynomial (e.g. coefficient of  $x^{n-1} = c_1 \frac{\partial g_{n-1}}{\partial \zeta_1} + \dots + c_m \frac{\partial g_{n-1}}{\partial \zeta_m} =$  first element of  $c_1 J_1 + \dots + c_m J_m$ ). Hence

$$r(x) = \sum_{j=1}^m c_j l_j \prod_{k \neq j} (x - \zeta_k) = 0 \quad (6.397)$$

Hence for  $t = 1, \dots, m$ ,

$$r(\zeta_t) = c_t [l_t \prod_{k \neq t} (\zeta_t - \zeta_k)] = 0 \quad (6.398)$$

implies  $c_t = 0$  since the  $l_t s$  are positive and  $\zeta_k s$  are distinct. Hence the  $J_j s$  are independent as claimed. On the other hand suppose that  $\zeta_1, \dots, \zeta_m$  are not distinct, e.g.  $\zeta_1 = \zeta_2$ . Then the first two columns of  $\mathbf{J}(\mathbf{z})$  are coefficients of polynomials

$$h_1 = -l_1 (x - \zeta_1)^{l_1-1} (x - \zeta_2)^{l_2} \prod_{k=3}^m (x - \zeta_k)^{l_k} \quad (6.399)$$

and

$$h_2 = -l_2 (x - \zeta_1)^{l_1} (x - \zeta_2)^{l_2-1} \prod_{k=3}^m (x - \zeta_k)^{l_k} \quad (6.400)$$

Since  $\zeta_1 = \zeta_2$ , these differ only by constant multiples  $l_1$  and  $l_2$ . Hence  $\mathbf{J}(\mathbf{z})$  is rank deficient ( $rank \leq m - 1$ ).

With the system 6.386 being non-singular by the above theorem, the Gauss-Newton iteration

$$\mathbf{z}_{k+1} = \mathbf{z}_k - [\mathbf{J}(\mathbf{z}_k)_W^+][G_l(\mathbf{z}_k) - \mathbf{a}] \quad (k = 0, 1, \dots) \tag{6.401}$$

on  $\Pi_l$  is well-defined. Here

$$\mathbf{J}(\mathbf{z}_k)_W^+ = [\mathbf{J}(\mathbf{z}_k)^H \mathbf{W}^2 \mathbf{J}(\mathbf{z}_k)]^{-1} \mathbf{J}(\mathbf{z}_k)^H \mathbf{W}^2 \tag{6.402}$$

**Theorem 3.4.** Let  $\tilde{\mathbf{z}} = (\tilde{\zeta}_1, \dots, \tilde{\zeta}_m)$  be a pejorative root of  $p \sim \mathbf{a}$  associated with multiplicity structure  $l$  and weight  $\mathbf{W}$ . Assume  $\tilde{\zeta}_1, \tilde{\zeta}_2, \dots, \tilde{\zeta}_m$  are distinct. Then there is a number  $\epsilon (> 0)$  such that, if  $\|\mathbf{a} - G_l(\tilde{\mathbf{z}})\|_W < \epsilon$  and  $\|\mathbf{z}_0 - \tilde{\mathbf{z}}\|_2 < \epsilon$ , then iteration 6.401 is well-defined and converges to the pejorative root  $\tilde{\mathbf{z}}$  with at least a linear rate. If further  $\mathbf{a} = G_l(\tilde{\mathbf{z}})$ , then the convergence is quadratic.

**Proof.** Let  $F(\mathbf{z}) = \mathbf{W}[G_l(\mathbf{z}) - \mathbf{a}]$  and  $\hat{\mathbf{J}}(\mathbf{z})$  be its Jacobian.  $F(\mathbf{z})$  is obviously analytic. From Theorem 3.3, the smallest singular value of  $\hat{\mathbf{J}}(\mathbf{z})$  is strictly positive. If  $\mathbf{a}$  is sufficiently close to  $G_l(\tilde{\mathbf{z}})$ , then

$$\|F(\tilde{\mathbf{z}})\|_2 = \|G_l(\tilde{\mathbf{z}}) - \mathbf{a}\|_W \tag{6.403}$$

will be small enough so that 6.378 holds with  $\delta < \sigma^2$ . Thus all the conditions of Lemma 2.8 are satisfied and there is a neighborhood  $\Omega$  of  $\tilde{\mathbf{z}}$  such that if  $\mathbf{z}_0 \in \Omega$ , the iteration 6.401 converges and satisfies 6.381. If in addition  $\mathbf{a} = G_l(\tilde{\mathbf{z}})$ , then  $F(\tilde{\mathbf{z}}) = \mathbf{0}$  and so  $\delta = 0$  in 6.378 and 6.381. Thus the convergence becomes quadratic.

As a special case for  $l = [1, 1, \dots, 1]$ , the equations 6.386 form Vieta’s nonlinear system. Solving this by Newton’s n-dimensional method is equivalent to the WDK algorithm. When a polynomial has multiple roots Vieta’s system becomes singular at the (nondistinct) root vector. This appears to be the reason that causes ill-conditioning of conventional root-finders: a wrong pejorative manifold is used.

Zeng next discusses a “structure-preserving” condition number. In general, a condition number is the smallest number satisfying

$$[\text{forward error}] \leq [\text{condition number}] \times [\text{backward error}] + \text{h.o.t.} \tag{6.404}$$

where h.o.t means higher-order terms in the backward error. In our context forward error means error in the roots, and backward error means the errors in the coefficients which would produce that root error. For a polynomial with multiple roots, under **unrestricted** perturbations, the only condition number satisfying 6.404 is infinity. For example, consider  $p(x) = x^2$  (roots 0,0). A backward error  $\epsilon$  gives a perturbed polynomial  $x^2 + \epsilon$ , which has roots  $\pm\sqrt{\epsilon}i$ , i.e. forward error of magnitude  $\sqrt{\epsilon}$ . The only constant  $c$  which accounts for  $\sqrt{\epsilon} \leq c\epsilon$  for **all**  $\epsilon > 0$  must be infinity (for if  $\epsilon < 1$ ,  $\sqrt{\epsilon} > \epsilon$ ). However by changing our objective from solving  $p(x) = 0$  to solving the non-linear least squares problem in the form 6.387, the

structure-altering noise is filtered out, and the multiplicity structure is preserved, leading usually to far less sensitivity in the roots.

Consider the root vector  $\mathbf{z}$  of  $p \sim \mathbf{a} = G_l(\mathbf{z})$ . The polynomial  $p$  is perturbed, with multiplicity structure  $l$  preserved, to  $\hat{p} \sim \hat{\mathbf{a}} = G_l(\hat{\mathbf{z}})$ . That is, both  $p$  and  $\hat{p}$  are on the same pejorative manifold  $\Pi_l$ . Then

$$\hat{\mathbf{a}} - \mathbf{a} = G_l(\hat{\mathbf{z}}) - G_l(\mathbf{z}) = \mathbf{J}(\mathbf{z})(\hat{\mathbf{z}} - \mathbf{z}) + O(\|\hat{\mathbf{z}} - \mathbf{z}\|^2) \quad (6.405)$$

where  $\mathbf{J}(\mathbf{z})$  is the Jacobian of  $G_l(\mathbf{z})$ . If the elements of  $\mathbf{z}$  are distinct, then by Theorem 3.3  $\mathbf{J}(\mathbf{z})$  is of full rank. Hence

$$\|\mathbf{W}(\hat{\mathbf{a}} - \mathbf{a})\|_2 = \|[\mathbf{W}\mathbf{J}(\mathbf{z})](\hat{\mathbf{z}} - \mathbf{z})\|_2 + h.o.t \quad (6.406)$$

i.e.

$$\|\hat{\mathbf{a}} - \mathbf{a}\|_W \geq \sigma_{min}\|\hat{\mathbf{z}} - \mathbf{z}\|_2 + h.o.t \quad (6.407)$$

or

$$\|\hat{\mathbf{z}} - \mathbf{z}\|_2 \leq \frac{1}{\sigma_{min}}\|\hat{\mathbf{a}} - \mathbf{a}\|_W + h.o.t. \quad (6.408)$$

where  $\sigma_{min}$ , the smallest singular value of  $\mathbf{W}\mathbf{J}(\mathbf{z})$ , is  $> 0$  since  $\mathbf{W}$  and  $\mathbf{J}(\mathbf{z})$  are of full rank. 6.408 is in the form of 6.404, with forward error =  $\|\hat{\mathbf{z}} - \mathbf{z}\|_2$ , backward error =  $\|\hat{\mathbf{a}} - \mathbf{a}\|_W$ , and condition number =  $\frac{1}{\sigma_{min}}$ . Thus in the present sense (of multiplicity-preserving perturbations) the sensitivity of multiple roots is finite. The above condition number, which depends on the multiplicity structure  $l$  and the weight  $\mathbf{W}$ , is called  $\kappa_{l,w}(\mathbf{z})$ . Note that the array  $l = [l_1, \dots, l_m]$  may or may not be the actual multiplicity structure. Thus a polynomial has different condition numbers corresponding to different pejorative roots on various pejorative manifolds.

Now suppose a polynomial  $p$  is perturbed to give a slightly different polynomial  $\hat{p}$ , with both polynomials near a pejorative manifold  $\Pi_l$ . It is possible that neither polynomial possesses the structure  $l$  exactly, so that both polynomials may be ill-conditioned in the conventional sense. So the exact roots of  $p$  and  $\hat{p}$  may be far apart. However the following theorem ensures that their **pejorative roots** (unlike their exact ones) may still be insensitive to perturbations.

**Theorem 3.6.** For a fixed  $l = [l_1, \dots, l_m]$ , let the polynomial  $\hat{p} \sim \hat{\mathbf{b}}$  be an approximation to  $p \sim \mathbf{b}$  with pejorative roots  $\hat{\mathbf{z}}$  and  $\mathbf{z}$ , respectively, that correspond to the multiplicity structure  $l$  and weight  $\mathbf{W}$ . Assume the components of  $\mathbf{z}$  are distinct while  $\|G_l(\mathbf{z}) - \hat{\mathbf{b}}\|_W$  reaches a local minimum at  $\hat{\mathbf{z}}$ . If  $\|\mathbf{b} - \hat{\mathbf{b}}\|_W$  and  $\|G_l(\mathbf{z}) - \mathbf{b}\|_W$  are sufficiently small, then

$$\|\mathbf{z} - \hat{\mathbf{z}}\|_2 \leq 2\kappa_{l,w}(\mathbf{z}) \cdot (\|G_l(\mathbf{z}) - \mathbf{b}\|_W + \|\mathbf{b} - \hat{\mathbf{b}}\|_W) + h.o.t \quad (6.409)$$

**Proof.** From 6.408

$$\|\mathbf{z} - \hat{\mathbf{z}}\|_2 \leq \kappa_{l,w}(\mathbf{z})\|G_l(\mathbf{z}) - G_l(\hat{\mathbf{z}})\|_W + h.o.t$$



$$\leq \kappa_{l,w}(\mathbf{z})(\|G_l(\mathbf{z}) - \mathbf{b}\|_W + \|\mathbf{b} - \hat{\mathbf{b}}\|_W + \|G_l(\hat{\mathbf{z}}) - \hat{\mathbf{b}}\|_W) + h.o.t \quad (6.410)$$

Since  $\|G_l(\hat{\mathbf{z}}) - \hat{\mathbf{b}}\|_W$  is a local minimum, we have

$$\|G_l(\hat{\mathbf{z}}) - \hat{\mathbf{b}}\|_W \leq \|G_l(\mathbf{z}) - \hat{\mathbf{b}}\|_W \leq \|G_l(\mathbf{z}) - \mathbf{b}\|_W + \|\mathbf{b} - \hat{\mathbf{b}}\|_W \quad (6.411)$$

and the theorem follows.

This means that even if the exact roots are hypersensitive, the pejorative roots are stable if  $\kappa_{l,w}(\mathbf{z})$  is not too large. For a polynomial  $p$  having a multiplicity structure  $l$ , we can now estimate the error of its multiple roots computed from its approximation  $\hat{p}$ . The **exact** roots of  $\hat{p}$  are in general all simple and far from the multiple roots of  $p$ . However by the following corollary the **pejorative** roots  $\hat{\mathbf{z}}$  of  $\hat{p}$  w.r.t.  $l$  can be an accurate approximation to the multiple roots  $\mathbf{z}$  of  $p$ .

**Corollary 3.7.** Under the conditions of Theorem 3.6, if  $\mathbf{z}$  is the exact root vector of  $p$  with multiplicity structure  $l$ , then

$$\|\mathbf{z} - \hat{\mathbf{z}}\|_2 \leq 2\kappa_{l,w}(\mathbf{z})\|\mathbf{b} - \hat{\mathbf{b}}\|_W + h.o.t. \quad (6.412)$$

**Proof.** Since  $\mathbf{z}$  is exact,  $\|G_l(\mathbf{z}) - \mathbf{b}\|_W = \mathbf{0}$  in 6.411.

The “attainable accuracy” barrier suggests that when multiplicity increases, so does the root sensitivity. But apparently this does not apply to the structure-constrained sensitivity. For example, consider the set of polynomials

$$p_l(x) = (x + 1)^{l_1}(x - 1)^{l_2}(x - 2)^{l_3} \quad (6.413)$$

with different sets  $l = [l_1, l_2, l_3]$ . For the weight  $\mathbf{W}$  defined in 6.388 the condition number is 2.0 for  $l = [1, 2, 3]$ , .07 for  $l = [10, 20, 30]$ , and only .01 for  $l = [100, 200, 300]$ . We get the surprising result that the root error may be **less** than the data error in such cases. Thus multiprecision arithmetic may not be a necessity, and the attainable accuracy barrier may not apply. The condition number can be calculated with relatively little cost, for the QR decomposition of  $\mathbf{WJ}(\mathbf{z})$  is required by the Gauss-Newton iteration 6.401, and can be re-used to calculate  $\kappa_{l,w}$ . That is, inverse iteration as in Lemma 2.6 can be used to find  $\sigma_{min}$ . Iteration 6.401 requires calculation of the vector value of  $G_l(\mathbf{z}_k)$  and the matrix  $\mathbf{J}(\mathbf{z}_k)$ , where the components of  $G_l(\mathbf{z})$  are defined in 6.384 and 6.385 as coefficients of the polynomial

$$p(x) = (x - z_1)^{l_1} \dots (x - z_m)^{l_m} \quad (6.414)$$

Zeng suggests doing this numerically by constructing  $p(x)$  recursively by multiplication with  $(x - z_j)$ , which is equivalent to convolution with vectors  $(1, -z_j)^T$ . We do this  $l_j$  times for  $j = 1, \dots, m$ . The following, Algorithm EVALG, summarizes this calculation in pseudo-code. It takes  $n^2 + O(n)$  flops.

#### Algorithm EVALG

input:  $m, n, \mathbf{z} = (z_1, \dots, z_m)^T, l = [l_1, \dots, l_m]$ .

output: vector  $G_l(\mathbf{z}) \in C^m$ .

Calculation:

```

 $\mathbf{s} = (1)$ 
for  $i = 1, 2, \dots, m$  do
  for  $k = 1, 2, \dots, l_i$  do
     $\mathbf{s} = \text{conv}(\mathbf{s}, (1, -z_i))$ 
  end do
end do
 $g_{n-j}(\mathbf{z}) = (j+1)\text{th component of } \mathbf{s}$  for  $j=1, \dots, n$ 

```

The  $j$ 'th column of the Jacobian  $\mathbf{J}(\mathbf{z})$ , as shown in the proof of Theorem 3.3, can be considered as the coefficients of  $q_j(x)$  defined in 6.394. See the pseudo-code for EVALJ shown below:

#### Algorithm EVALJ

input:  $m, n, \mathbf{z} = (z_1, \dots, z_m)^T, l = [l_1, \dots, l_m]$ .

output: Jacobian  $\mathbf{J}(\mathbf{z}) \in C^{n \times m}$ .

Calculation:

```

 $\mathbf{u} = \prod (x - z_j)^{l_j - 1}$  by EVALG
for  $j = 1, 2, \dots, m$  do
   $\mathbf{s} = -l_j \mathbf{u}$ 
  for  $l = 1, \dots, m, l \neq j$  do
     $\mathbf{s} = \text{conv}(\mathbf{s}, (1, -z_l))$ 
  end do
   $j$ 'th column of  $\mathbf{J}(\mathbf{z}) = \mathbf{s}$ 
end do

```

Zeng states that this takes  $mn^2 + O(n)$  flops. Each step of the Gauss-Newton iteration takes  $O(nm^2)$  flops, for a total of  $O(m^2n + mn^2)$ . The complete pseudo-code for Algorithm I is shown below:

#### Pseudo-code PEJROOT (Algorithm I)

input:  $m, n, \mathbf{a} \in C^n$ , weight matrix  $\mathbf{W}$ , initial iterate  $\mathbf{z}_0$ ,  
multiplicity structure  $l$ , error tolerance  $\tau$ .

output: roots  $\mathbf{z} = (\zeta_1, \dots, \zeta_m)$ , or a message of failure.

Calculation:

for  $k = 0, 1, \dots$  do

Calculate  $G_l(\mathbf{z}_k)$  and  $\mathbf{J}(\mathbf{z}_k)$  with EVALG and EVALJ

Compute the least squares solution  $\Delta \mathbf{z}_k$  to the linear system

$$[\mathbf{W}\mathbf{J}(\mathbf{z}_k)](\Delta \mathbf{z}_k) = \mathbf{W}[G_l(\mathbf{z}_k) - \mathbf{a}]$$

Set  $\mathbf{z}_{k+1} = \mathbf{z}_k - \Delta \mathbf{z}_k$  and  $\delta_k = \|\Delta \mathbf{z}_k\|_2$

if  $k \geq 1$  then

if  $\delta_k \geq \delta_{k-1}$  then stop, output failure message

```

    else if  $\frac{\delta_k^2}{\delta_{k-1}-\delta_k} < \tau$  then stop, output  $\mathbf{z} = \mathbf{z}_{k+1}$ 
    end if
  end if
end do

```

Zeng performed several tests of his Algorithm I, implemented as a Matlab code PEJROOT. The tests were performed strictly with IEEE double precision arithmetic. His method was compared with the third-order method of Farmer and Loizou (1977) which is subject to the “attainable accuracy” barrier. Both methods were applied to

$$p_1(x) = (x-1)^4(x-2)^3(x-3)^2(x-4) \quad (6.415)$$

starting with  $\mathbf{z}_0 = (1.1, 1.9, 3.1, 3.9)$ . Farmer-Loizou bounces around, for example giving 3.3 for the second root after 100 iterations. In contrast Zeng’s method converges to 14 digits after 8 iterations. For the next case the multiplicities in 6.415 are changed to 40,30,20, and 10 respectively. Now the Farmer-Loizou program uses 1000-digit arithmetic and yet still fails dismally, while Zeng attains 14-digit accuracy in 6 iterations. The accuracy barrier in Algorithm I is  $\kappa_{l,w}(\mathbf{z})$ , which is 29.3 in this case. PEJROOT calculated the coefficients with a relative error of  $4.56 \times 10^{-16}$ . The actual root error is about  $1 \times 10^{-14}$ , which is  $<$  the error bound  $2 \times (29.3) \times (4.56 \times 10^{-16}) = 2.67 \times 10^{-14}$  given by 6.412. Root-finding packages which use multiprecision, such as MPSOLVE of Bini (1999) and EIGENSOLVE of Fortune (2002), can accurately solve polynomials with exact coefficients, but for inexact coefficients and multiple roots the accuracy is very limited. For example the polynomial  $p(x) = (x - \sqrt{2})^{20}(x - \sqrt{3})^{10}$  with coefficients calculated to 100 digits has “attainable accuracy” of 5 and 10 digits for the  $\sqrt{2}$  and  $\sqrt{3}$  roots respectively. MPSOLVE and EIGENSOLVE reach this accuracy but no more, whereas PEJROOT gives roots and multiplicity to 15-digit accuracy using only 16-digit precision in the coefficients and standard machine precision (also 16 digits). Even clustered roots can be dealt with, e.g.

$$f(x) = (x - .9)^{18}(x - 1)^{10}(x - 1.1)^{16}$$

was solved by the MATLAB function ROOTS giving 44 alleged roots in a  $2 \times 2$  box (i.e. some roots have an imaginary part  $> 1i$ ). On the other hand PEJROOT obtains all roots to 14 digit accuracy, starting with the multiplicity structure and initial approximations provided by Algorithm II. The condition number is 60.4, and coefficients are perturbed in the 16th digit, so that 14 digits accuracy in the roots is the best that can be expected. Zeng also considers a case with multiplicities up to 400. He perturbs the coefficients in the 6th digits, and PEJROOT obtains all roots correct to 7 digits; whereas ROOTS gives many totally incorrect estimates.

We now describe “Algorithm II” which calculates the multiplicity structure of a given polynomial as well as an initial root approximation, both to be used by

Algorithm I (note that a little strangely Algorithm II must be applied BEFORE Algorithm I). Now for a polynomial  $p$  with  $u = GCD(p, p')$ ,  $v = \frac{p}{u}$  has the same **distinct** roots as  $p$ , but all roots of  $v$  are simple. If we can find  $v$  its simple roots can be found by some “standard” root-finder such as those described in earlier sections of this Chapter. Thus the following process, also described in Chapter 2, will in principle give us the factorization of  $p$ :

```

 $u_0 = p$ 
for  $j = 1, 2, \dots$  while  $deg(u_{j-1}) > 0$  do
    calculate
         $u_j = GCD(u_{j-1}, u'_{j-1}); v_j = \frac{u_{j-1}}{u_j}$ 
        calculate the (simple) roots of  $v_j(x)$ 
end do

```

The usual GCD calculations are often numerically unstable. Zeng avoids this problem as follows: he factors a polynomial  $p$  and its derivative  $p'$  with a GCD triplet  $(u, v, w)$  such that

$$u(x)v(x) = p(x) \quad (6.417)$$

$$u(x)w(x) = p'(x) \quad (6.418)$$

where  $u(x)$  is monic while  $v(x)$  and  $w(x)$  are coprime. He uses a successive updating process that calculates only the smallest singular values of the Sylvester matrices  $S_j(p)$ ,  $j = 1, 2, \dots$  and stops at the first rank-deficient matrix  $S_m(p)$ . With this  $S_m(p)$  we can find the degrees of  $u$ ,  $v$ ,  $w$ , and obtain coefficients of  $v$  and  $w$  from the right singular vector (see later). Using a more stable least squares division we can generate an approximation to the GCD triplet, and obtain an initial iterate. The key part of Algorithm II is the following GCD-finder:

**STEP 1.** Find the degree  $k$  of  $GCD(p, p')$  ( $= u$ ).

**STEP 2.** Set up the system 6.417-6.418 according to the degree  $k$ .

**STEP 3.** Find an initial approximation to  $u$ ,  $v$ ,  $w$ .

**STEP 4.** Use the Gauss-Newton iteration to refine the GCD triplet  $(u, v, w)$ .

We shall now describe each step in detail, starting with **STEP 1**. Let  $p$  be a polynomial of degree  $n$ . By Lemma 2.4, the degree of  $u = GCD(p, p')$  is  $k = n - m$  iff the  $m$ 'th Sylvester matrix is the first one being rank-deficient. Hence  $k = deg(u)$  can be found by calculating the sequence of the smallest singular values  $\sigma_j$  of  $S_j(p)$ , ( $j = 1, 2, \dots$ ) until reaching  $\sigma_m$  that is *approximately* zero. Since only one singular pair (i.e. the singular value and the right singular vector) is needed, the inverse iteration of Lemma 2.6 can be used. Moreover we can further

reduce the cost by recycling and updating the QR decomposition of the  $S_j(p)$ 's. For let

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_0 \tag{6.419}$$

$$\text{and } p'(x) = b_{n-1} x^{n-1} + b_{n-2} x^{n-2} + \dots + b_0 \tag{6.420}$$

We rotate the columns of  $S_j(p)$  to form  $\hat{S}_j(p)$  so that the odd and even columns of  $\hat{S}_j(p)$  consist of the coefficients of  $p'$  and  $p$  respectively, i.e.

$$\begin{bmatrix} b_{n-1} & 0 & \dots & a_n & 0 & \dots \\ b_{n-2} & \dots & \dots & a_{n-1} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & b_{n-1} & \dots & \dots & a_n \\ b_0 & \dots & b_{n-2} & \dots & \dots & \dots \\ \dots & \dots & \dots & a_0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & b_0 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & a_0 \end{bmatrix} \tag{6.421}$$

(with  $j+1$  columns of the  $b_i$ , and  $j$  columns of the  $a_i$ ) becomes:

$$\begin{bmatrix} b_{n-1} & a_n & \dots & \dots & \dots & \dots & \dots & \dots \\ b_{n-2} & a_{n-1} & b_{n-1} & a_n & \dots & \dots & \dots & \dots \\ \dots & \dots & b_{n-2} & a_{n-1} & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ b_0 & \dots & \dots & \dots & \dots & b_{n-1} & a_n & \dots \\ \dots & a_0 & b_0 & \dots & \dots & \dots & \dots & b_{n-1} \\ \dots & \dots & \dots & a_0 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & b_0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & a_0 & b_0 \end{bmatrix} \tag{6.422}$$

Thus the new matrix  $\hat{S}_{j+1}(p)$  is formed from  $\hat{S}_j(p)$  by adding a zero row at the bottom and then two columns at the right. Updating the QR decomposition of successive  $\hat{S}_j(p)$ 's requires only  $O(n)$  flops. The inverse iteration 6.365-6.367 requires  $O(j^2)$  flops for each  $\hat{S}_j(p)$ . Let  $\theta$  be a given *zero singular value threshold* (see later), then the algorithm for finding the degrees of  $u, v, w$  can be summarized as follows:

Calculate the QR decomposition of the  $(n + 1) \times 3$  matrix  $\hat{S}_1(f) = Q_1 R_1$ .

for  $j = 1, 2, \dots$  do

use the inverse iteration 6.365-6.367 to find the smallest singular value  $\sigma_j$  of  $\hat{S}_j(p)$  and the corresponding right singular vector  $\mathbf{y}_j$

If  $\sigma_j \leq \theta \|\mathbf{p}\|_2$ , then  $m = j$ ,  $k = n - m$ , extract  $\mathbf{v}$  and  $\mathbf{w}$  from  $\mathbf{y}_j$  (see Lemma 2.5), exit

else update  $\hat{S}_j(p)$  to  $\hat{S}_{j+1}(p) = Q_{j+1}R_{j+1}$

end if

end do

**STEP 2.** Let  $k = n - m$  be the degree of  $u$  calculated in Step 1. We express the GCD system 6.417-6.418 in vector form with unknown vectors  $\mathbf{u}$ ,  $\mathbf{v}$  and  $\mathbf{w}$ :

$$\begin{bmatrix} u_k \\ conv(\mathbf{u}, \mathbf{v}) \\ conv(\mathbf{u}, \mathbf{w}) \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{p} \\ \mathbf{p}' \end{bmatrix} \tag{6.423}$$

for  $\mathbf{u} \in C^{k+1}$ ,  $\mathbf{v} \in C^{m+1}$ ,  $\mathbf{w} \in C^m$ .

**Lemma 4.1.** The Jacobian of 6.423 is

$$\mathbf{J}(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \begin{bmatrix} \mathbf{e}_1^T & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_m(\mathbf{v}) & \mathbf{C}_k(\mathbf{u}) & \mathbf{0} \\ \mathbf{C}_m(\mathbf{w}) & \mathbf{0} & \mathbf{C}_{k-1}(\mathbf{u}) \end{bmatrix} \tag{6.424}$$

If  $u = GCD(p, p')$  with  $(\mathbf{u}, \mathbf{v}, \mathbf{w})$  satisfying 6.423, then  $\mathbf{J}(\mathbf{u}, \mathbf{v}, \mathbf{w})$  is of full rank.

**Proof.** 6.424 follows from Lemma 2.2. To prove that  $\mathbf{J}(\mathbf{u}, \mathbf{v}, \mathbf{w})$  is of full rank, assume that there exist polynomials  $q(x)$ ,  $r(x)$  and  $s(x)$  of degrees  $\leq k$ ,  $\leq m$ ,  $\leq m - 1$  respectively such that

$$\mathbf{J}(\mathbf{u}, \mathbf{v}, \mathbf{w}) \begin{bmatrix} \mathbf{q} \\ \mathbf{r} \\ \mathbf{s} \end{bmatrix} = \mathbf{0} \text{ or } \left\{ \begin{array}{l} q_k = 0 \\ vq + ur = 0 \\ wq + us = 0 \end{array} \right\} \tag{6.425}$$

where  $\mathbf{q}$ ,  $\mathbf{r}$ ,  $\mathbf{s}$  are the coefficient vectors of  $q(x)$ ,  $r(x)$  and  $s(x)$ . From 6.425 we have  $vq = -ur$  and  $wq = -us$ . Hence  $wvq - vwq = -uwr + uvs = 0$ , i.e.  $-ur + vs = 0$  (since  $u \not\equiv 0$ ) or  $wr = vs$ . Since  $w$  and  $v$  are coprime, the factors of  $v =$  factors of  $r$ , hence  $r = t_1v$ . Similarly  $s = t_2w$ , so that  $wr = t_1vw = t_2vw$  and finally  $t_1 = t_2 = t$ . Hence  $vq = -ur = -utv$  leads to  $q = -tu$ . But  $deg(q) = deg(tu) \leq k$ ,  $deg(u) = k \geq 0$  and  $u_k = 1$ , so  $deg(t) = 0$  i.e.  $t = \text{const}$ . Finally by the first equation in 6.425 and  $u_k = 1$  we have  $q_k = -tu_k = -t = 0$ , so that  $q = -tu = 0$ ,  $r = tv = 0$ , and  $s = tw = 0$ . Thus  $\mathbf{J}(\mathbf{u}, \mathbf{v}, \mathbf{w})$  must be of full rank.

Zeng next states

**Theorem 4.2.** Let  $\tilde{u} = GCD(p, p')$  with  $\tilde{v}$  and  $\tilde{w}$  satisfying 6.423, and let  $\mathbf{W}$  be

a weight matrix. Then there exists  $\epsilon > 0$  such that for all  $\mathbf{u}_0, \mathbf{v}_0, \mathbf{w}_0$  satisfying  $\|\mathbf{u}_0 - \tilde{\mathbf{u}}\|_2 < \epsilon, \|\mathbf{v}_0 - \tilde{\mathbf{v}}\|_2 < \epsilon,$  and  $\|\mathbf{w}_0 - \tilde{\mathbf{w}}\| < \epsilon,$  the Gauss-Newton iteration

$$\begin{bmatrix} \mathbf{u}_{j+1} \\ \mathbf{v}_{j+1} \\ \mathbf{w}_{j+1} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_j \\ \mathbf{v}_j \\ \mathbf{w}_j \end{bmatrix} - \mathbf{J}(\mathbf{u}_j, \mathbf{v}_j, \mathbf{w}_j)_W^+ \begin{bmatrix} \mathbf{e}_1^T \mathbf{u}_j - 1 \\ \text{conv}(\mathbf{u}_j, \mathbf{v}_j) - \mathbf{f} \\ \text{conv}(\mathbf{u}_j, \mathbf{w}_j) - \mathbf{f}' \end{bmatrix}$$

( $j = 0, 1, \dots$ ) (6.426)

converges to  $[\tilde{\mathbf{u}}, \tilde{\mathbf{v}}, \tilde{\mathbf{w}}]^T$  quadratically. Here

$$\mathbf{J}(\cdot)_W^+ = [\mathbf{J}(\cdot)^H \mathbf{W}^2 \mathbf{J}(\cdot)]^{-1} \mathbf{J}(\cdot)^H \mathbf{W}^2$$

(6.427)

is the weighted pseudo-inverse of the Jacobian  $\mathbf{J}(\cdot)$  defined in 6.424.

**Proof.** Zeng refers to Lemmas 2.8 and 4.1

**STEP 3.** Initial iterates  $\mathbf{v}_0, \mathbf{w}_0$  can be obtained from Step 1 i.e. when the singular value  $\sigma_m$  is calculated, the associated singular vector  $\mathbf{y}_m$  consists of  $\mathbf{v}_0$  and  $\mathbf{w}_0,$  which are approximations to  $\mathbf{v}$  and  $\mathbf{w}$  in 6.423 (see Lemma 2.5 (c)). Because of the column permutation in 6.422, the odd and even entries of  $\mathbf{y}_m$  form  $\mathbf{v}_0$  and  $\mathbf{w}_0$  respectively.  $\mathbf{u}_0$  is not found by conventional long division

$$p(x) = v_0(x)q(x) + r(x) \text{ with } u_0 = q, r = 0$$

(6.428)

(which may not be numerically stable). Rather we solve the linear system (see Lemma 2.5(d)):

$$\mathbf{C}_k(\mathbf{v}_0)\mathbf{u}_0 = \mathbf{p}$$

(6.429)

by least squares so as to minimize

$$\|\text{conv}(\mathbf{u}_0, \mathbf{v}_0) - \mathbf{p}\|_2$$

(6.430)

This “least squares division” is more accurate than 6.428, which is equivalent to solving the  $(n + 1) \times (n + 1)$  lower triangular linear system

$$\mathbf{L}_k(\mathbf{v}_0) \begin{bmatrix} \mathbf{q} \\ \mathbf{r} \end{bmatrix} = \mathbf{p}, \text{ with } \mathbf{L}_k(\mathbf{v}_0) = \begin{bmatrix} C_k(\mathbf{v}_0) & O_{(k+1) \times (n-k)} \\ \dots & I_{(n-k) \times (n-k)} \end{bmatrix}$$

(6.431)

In Theorem 4.3 Zeng proves that the condition number of  $\mathbf{C}_k(\mathbf{v})$  is  $\leq$  the condition number of  $\mathbf{L}_k(\mathbf{v})$ . In an example with  $v(x) = x + 25, m = 20,$  we have  $\kappa(\mathbf{C}_k(\mathbf{v})) = 1.08, \kappa(\mathbf{L}_k(\mathbf{v})) = 9 \times 10^{27}.$  In another example, the coefficients of  $\frac{p(x)}{v(x)}$  were obtained correct to at least 8 decimal digits by the “least squares division” 6.429-6.430, but some were wrong by a factor of 30 in the case of “long division” 6.428. After extracting  $\mathbf{v}_0$  and  $\mathbf{w}_0$  from  $\mathbf{y}_m$  and solving 6.429 for  $\mathbf{u}_0,$  we use them as initial iterates for the Gauss-Newton process 6.426 that refines the GCD triplet. The system 6.429 is banded, with band-width  $k+1,$  so the cost of solving it is relatively low.

**STEP 4.** The Gauss-Newton iteration is used to reduce the residual

$$\left\| \begin{pmatrix} conv(\mathbf{u}_j, \mathbf{v}_j) \\ conv(\mathbf{u}_j, \mathbf{w}_j) \end{pmatrix} - \begin{pmatrix} \mathbf{p} \\ \mathbf{p}' \end{pmatrix} \right\|_W \tag{6.432}$$

at each step. We stop when this residual no longer decreases.  $\mathbf{W}$  is used to scale the GCD system 6.423 so that the entries of  $\mathbf{W} \begin{bmatrix} \mathbf{p} \\ \mathbf{p}' \end{bmatrix}$  are of similar magnitude. Each step of Gauss-Newton requires solving an overdetermined linear system

$$[\mathbf{WJ}(\mathbf{u}_j, \mathbf{v}_j, \mathbf{w}_j)]\mathbf{z} = \mathbf{W} \begin{bmatrix} \mathbf{e}_1^T \mathbf{u}_j - 1 \\ conv(\mathbf{u}_j, \mathbf{v}_j) - \mathbf{p} \\ conv(\mathbf{u}_j, \mathbf{w}_j) - \mathbf{p}' \end{bmatrix} \tag{6.433}$$

for its least squares solution  $\mathbf{z}$ , and requires a QR factorization of the Jacobian  $\mathbf{WJ}$  and a backward substitution for an upper triangular linear system. This Jacobian has a special sparsity structure that can largely be preserved during the process. Taking this sparsity into account, the cost of the sparse QR factorization is  $O(mk^2 + m^2k + m^3)$  where  $m$  = number of distinct roots.

We next discuss the computation of the multiplicity structure. The procedure 6.416 generates a sequence of square-free polynomials  $v_1, v_2, \dots, v_s$  of degrees  $d_1 \geq d_2 \geq \dots \geq d_s$  respectively, such that

$$p = v_1 v_2 \dots v_s = \frac{u_0}{u_1} \frac{u_1}{u_2} \dots \frac{u_{s-1}}{u_s} \tag{6.434}$$

where  $u_0 = p$  and  $u_s = 1$ . Furthermore

$$\{\text{roots of } v_1\} \supseteq \{\text{roots of } v_2\} \supseteq \dots \supseteq \{\text{roots of } v_s\} \tag{6.435}$$

All  $v_j$  are simple; roots of  $v_1$  consist of all distinct roots of  $p$ ; roots of  $v_2$  consist of all distinct roots of  $\frac{p}{v_1}$ , etc. Then the multiplicity structure is determined by the degrees  $d_1, d_2, \dots, d_s$ . For example, considering

$$p(x) = (x - a)(x - b)^3(x - c)^4$$

we have the following:

	$v_j$	$deg(v_j)$	$roots$
$v_1(x)$	$= (x - a) (x - b) (x - c)$	$d_1 = 3$	$a, b, c$
$v_2(x)$	$= (x - b) (x - c)$	$d_2 = 2$	$b, c$
$v_3(x)$	$= (x - b) (x - c)$	$d_3 = 2$	$b, c$
$v_4(x)$	$= (x - c)$	$d_4 = 1$	$c$
<i>multiplicity structure</i> $\rightarrow$			1, 3, 4



Without finding the roots  $a, b, c$  the multiplicity structure  $[l_1, l_2, l_3] = [1, 3, 4]$  (with the  $l_i$  in increasing order) is solely determined by the degrees  $d_1, \dots, d_4$ . Thus

$$\begin{aligned} l_1 = 1 & \text{ since } d_1 \geq 3 = (d_1 + 1) - 1 \\ l_2 = 3 & \text{ since } d_1, d_2, d_3 \geq 2 = (d_1 + 1) - 2 \\ l_3 = 4 & \text{ since } d_1, d_2, d_3, d_4 \geq 1 = (d_1 + 1) - 3 \end{aligned}$$

In general we have

**Theorem 4.4.** With  $v_i$  and  $d_i$  as above let  $m = d_1 = \deg(v_1)$ . Then the multiplicity structure  $l$  consists of components

$$l_j = \max\{t | d_t \geq (d_1 + 1) - j\}, \quad j = 1, 2, \dots, m \quad (6.436)$$

**Proof.** Each  $u_i$  contains the factors of  $p(x)$  to degree one less than in  $u_{i-1}$  (except of course that those linear in  $u_{i-1}$  have disappeared in  $u_i$ ). Hence  $v_i = \frac{u_{i-1}}{u_i}$  contains only those factors of  $p(x)$  of degree at least  $i$ . So if  $d_1 = d_2 = \dots = d_r > d_{r+1}$ , then the factor of lowest degree in  $p(x)$  has degree  $r$ . Likewise if  $d_{r+1} = d_{r+2} = \dots = d_t > d_{t+1}$ , then the factor of next lowest degree has degree  $t$ , and so on.

The location of the roots is not needed in determining the multiplicity structure. The initial root approximation is determined based on the fact that an  $l$ -fold root of  $p(x)$  appears  $l$  times as a simple root among  $v_1, \dots, v_l$ . After calculating the roots of each  $v_j$  with a standard root-finder, numerically “identical” roots of the  $v_j$ ’s are grouped, according to the multiplicity structure  $[l_1, \dots, l_m]$ , to form the initial approximation  $(z_1, \dots, z_m)$  that is needed by Algorithm I.

We use three control parameters for the above. First is the *zero singular value threshold*  $\theta$  for identifying a numerically zero  $\sigma_m$ . The default value used is  $10^{-8}$ . When the smallest  $\sigma_l$  of  $\hat{S}_l(\mathbf{u}_{j-1})$  is  $< \theta \|\mathbf{u}_{j-1}\|_2$ , it will be tentatively considered 0. Then the Gauss-Newton iteration is used until

$$\rho_j = \left\| \begin{pmatrix} \text{conv}(\mathbf{u}_j, \mathbf{v}_j) - \mathbf{u}_{j-1} \\ \text{conv}(\mathbf{u}_j, \mathbf{w}_j) - \mathbf{u}'_{j-1} \end{pmatrix} \right\|_W \leq \rho \|\mathbf{u}_{j-1}\|_2 \quad (6.437)$$

where  $\rho$ , the *initial residual tolerance*, defaults to  $10^{-10}$ . If 6.437 is not yet satisfied we continue to update  $\hat{S}_l(\mathbf{u}_{j-1})$  to  $\hat{S}_{l+1}(\mathbf{u}_{j-1})$  and check  $\sigma_{l+1}$ . For the third control parameter (the residual growth factor) see the cited paper, p 896. A pseudo-code for Algorithm II is shown below. It is called GCDROOT and is included with PEJROOT in the overall package MULTROOT.

**Pseudocode GCDROOT (Algorithm II)**

**input:** Polynomial  $p$  of degree  $n$ , singular threshold  $\theta$ ,

residual tolerance  $\rho$ , residual growth factor  $\phi$

(If only  $p$  is provided, set  $\theta = 10^{-8}$ ,  $\rho = 10^{-10}$ ,  $\phi = 100$ )

**output:** the root estimates  $(z_1, \dots, z_m)^T$  and

multiplicity structure  $(l_1, \dots, l_m)$

Initialize  $u_0 = p$

for  $j=1,2,\dots,s$ , until  $\text{deg}(u_s) = 0$  do

  for  $l = 1,2,\dots$  until residual  $< \rho \|\mathbf{u}_{j-1}\|_2$  do

    calculate the singular pair  $(\sigma_l, \mathbf{y}_l)$  of  $\hat{S}_l(u_{j-1})$  by iteration 6.365-6.367

    if  $\sigma_l < \theta \|u_{j-1}\|_2$  then

      set up the GCD system 6.423 with  $p = u_{j-1}$

      extract  $v_j^{(0)}, w_j^{(0)}$  from  $\mathbf{y}_l$  and calculate  $u_j^{(0)}$

      apply the Gauss-Newton iteration 6.426 from

$u_j^{(0)}, v_j^{(0)}, w_j^{(0)}$  to obtain  $u_j, v_j, w_j$

      extract the residual  $\rho_j$  as in 6.437

    end if

  end do

  adjust the residual tolerance  $\rho$  to be  $\max(\rho, \phi\rho_j)$

  and set  $d_j = \text{deg}(v_j)$

end do

set  $m = d_1, l_j = \max\{t | d_t \geq m - j + 1\}, (j = 1, \dots, m)$

match the roots of  $v_i(x)$ ,  $(i = 1, \dots, s)$  according to the multiplicities  $l_j$ .

Convergence of Algorithm II with respect to inverse iteration is guaranteed by Lemma 2.6 (unless  $\mathbf{x}_0$  is orthogonal to  $\mathbf{y}$ , in which unlikely event orthogonality will in any case be destroyed by roundoff). The Gauss-Newton iteration could in theory cause trouble if the polynomial is perturbed to a place equidistant from two or more pejorative manifolds, but in extensive tests the algorithm always converged in practise.

Numerical tests included the following: consider

$$p(x) = (x-1)^{20}(x-2)^{15}(x-3)^{10}(x-4)^5$$

with coefficients rounded to 16 digits. GCDROOTS correctly finds the multiplicity structure, while the roots are approximated to 10 digits or better. Inputting these results to PEJROOT, Zeng obtained all roots correct to at least 14 digits. In contrast MPSOLVE (although using multiprecision) obtained spurious imaginary parts up to  $\pm 2.5i$ . Zeng's program is believed to be the only one to date which works at all accurately in such difficult cases. Another comparison was made with Uhlig's program PZERO based on the Euclidean division, for the polynomial

$$p_k(x) = (x-1)^{4k}(x-2)^{3k}(x-3)^{2k}(x-4)^k$$

with  $k = 1,2,\dots,8$ . PZERO fails to identify the multiplicity structure beyond  $k = 2$ , whereas GCDROOT finds the correct multiplicities up to  $k = 7$ , and the roots

correct to 11 digits up to this value of k. Also the effect of inexact coefficients was tested on the combined method using

$$p(x) = \left(x - \frac{10}{11}\right)^5 \left(x - \frac{20}{11}\right)^5 \left(x - \frac{30}{11}\right)^5$$

with coefficients rounded to k digits (k = 10,9,...). GCDROOT gives the correct multiplicity structure down to k = 7. When multiplicities are given manually PEJROOT converges for data correct only to 3 digits. Finally the author Zeng generates a polynomial  $f(x)$  of degree 20 based on known exact roots, and rounds the coefficients to 10 digits. He constructs multiple roots by repeated squaring, i.e.  $g_k(x) = [f(x)]^{2^k}$  for k = 1,2,3,4,5. Thus  $g_5$  has 20 complex roots each of multiplicity 32. The polynomials  $g_k(x)$  have inexact coefficients. MULTROOT finds accurate multiplicities and roots correct to at least 11 digits. Thus ends our description of Zeng’s paper and method.

Niu and Sakurai (2003) describe a somewhat different approach to finding multiple roots. They propose a new companion matrix which is efficient in finding multiple zeros and their multiplicities, and (perhaps more usefully) the mean of a cluster and the number of zeros in that cluster. They make use of a theorem of Smith (1970), which we quote in a form suitable when all zeros are simple: i.e. “For a monic polynomial  $p(z)$  of degree n, suppose that n distinct approximate zeros  $z_1, \dots, z_n$  are given, then the zeros of  $p(z)$  are the eigenvalues of the matrix

$$\mathbf{R} = \begin{bmatrix} z_1 - \frac{p(z_1)}{q'(z_1)} & -\frac{p(z_1)}{q'(z_2)} & \dots & -\frac{p(z_1)}{q'(z_n)} \\ -\frac{p(z_2)}{q'(z_1)} & z_2 - \frac{p(z_2)}{q'(z_2)} & \dots & -\frac{p(z_2)}{q'(z_n)} \\ \dots & \dots & \dots & \dots \\ -\frac{p(z_n)}{q'(z_1)} & \dots & \dots & z_n - \frac{p(z_n)}{q'(z_n)} \end{bmatrix} \tag{6.438}$$

(This is not how Niu and Sakurai write it, but this author suspects a misprint in their text).

In the above  $q(z) = \prod_{i=1}^n (z - z_i)$ . An iterative application of Smith’s method (find eigenvalues of  $\mathbf{R}$  and use those for  $z_i$  in the next iteration) does not work well for multiple roots, and this is true for other methods, such as Fiedler’s (unless perhaps multiple precision is used). As stated, Niu and Sakurai describe a new companion matrix which works well for multiple roots. Unlike most “classical” methods they compute the distinct zeros and their multiplicities separately (as Zeng does). Let

$$p(z) = c_n \prod_{k=1}^m (z - \zeta_k)^{l_k}, \quad \sum_{k=1}^m l_k = n, \quad c_n \neq 0 \tag{6.439}$$

have all n zeros located inside the circle  $\Gamma : \{z : |z - \gamma| < \rho\}$ , and let  $\zeta_1, \dots, \zeta_m$  be mutually distinct zeros of  $p(z)$  with multiplicities  $l_1, \dots, l_m$ . By a change of origin

and scale we can ensure that all zeros are inside the unit circle  $C$ . Hence  $l_1, \dots, l_m$  are residues of  $\frac{p'(z)}{p(z)}$  at  $\zeta_1, \dots, \zeta_m$ . Let

$$\mu_s = \frac{1}{2\pi i} \int_C z^s \frac{p'(z)}{p(z)} dz, \quad (s = 0, 1, 2, \dots) \tag{6.440}$$

Then by the residue Theorem (see e.g. Henrici(1974))

$$\mu_s = \sum_{k=1}^m l_k \zeta_k^s, \quad (s = 0, 1, 2, \dots) \tag{6.441}$$

Let  $H_m, H_m^<$  be the  $m \times m$  Hankel matrices

$$H_m = [\mu_{s+q}]_{s,q=0}^{m-1} = \begin{bmatrix} \mu_0 & \mu_1 & \dots & \mu_{m-1} \\ \mu_1 & \dots & \dots & \mu_m \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \mu_{2m-3} \\ \mu_{m-1} & \dots & \dots & \mu_{2m-2} \end{bmatrix} \tag{6.442}$$

$$H_m^< = [\mu_{s+q}]_{s,q=1}^m = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_m \\ \mu_2 & \dots & \dots & \mu_{m+1} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \mu_{2m-2} \\ \mu_m & \dots & \dots & \mu_{2m-1} \end{bmatrix} \tag{6.443}$$

N.B. We label subsequent lemmas etc according to the numbering of Niu and Sakurai.

**Lemma 3.1.** If  $\zeta_1, \dots, \zeta_m$  are mutually distinct, then  $H_m$  is non-singular. **Proof** Let

$$\mathbf{V}_m = \begin{bmatrix} 1 & \dots & 1 \\ \zeta_1 & \dots & \zeta_m \\ \dots & \dots & \dots \\ \zeta_1^{m-1} & \dots & \zeta_m^{m-1} \end{bmatrix} \tag{6.444}$$

and let

$$\mathbf{D}_m = \begin{bmatrix} l_1 & 0 & \dots & \dots \\ 0 & l_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & l_m \end{bmatrix} \tag{6.445}$$

Then by 6.441,

$$\mathbf{H}_m = \mathbf{V}_m \mathbf{D}_m \mathbf{V}_m^T \tag{6.446}$$

$$\begin{aligned}
 \text{(e.g. for } m = 3, \mathbf{V}_3 \mathbf{D}_3 \mathbf{V}_3^T &= \begin{bmatrix} 1 & 1 & 1 \\ \zeta_1 & \zeta_2 & \zeta_3 \\ \zeta_1^2 & \zeta_2^2 & \zeta_3^2 \end{bmatrix} \begin{bmatrix} l_1 & 0 & 0 \\ 0 & l_2 & 0 \\ 0 & 0 & l_3 \end{bmatrix} \begin{bmatrix} 1 & \zeta_1 & \zeta_1^2 \\ 1 & \zeta_2 & \zeta_2^2 \\ 1 & \zeta_3 & \zeta_3^2 \end{bmatrix} \\
 &= \begin{bmatrix} l_1 & l_2 & l_3 \\ l_1 \zeta_1 & l_2 \zeta_2 & l_3 \zeta_3 \\ l_1 \zeta_1^2 & l_2 \zeta_2^2 & l_3 \zeta_3^2 \end{bmatrix} \begin{bmatrix} 1 & \zeta_1 & \zeta_1^2 \\ 1 & \zeta_2 & \zeta_2^2 \\ 1 & \zeta_3 & \zeta_3^2 \end{bmatrix} = \begin{bmatrix} \sum l_i & \sum l_i \zeta_i & \sum l_i \zeta_i^2 \\ \sum l_i \zeta_i & \sum l_i \zeta_i^2 & \sum l_i \zeta_i^3 \\ \sum l_i \zeta_i^2 & \sum l_i \zeta_i^3 & \sum l_i \zeta_i^4 \end{bmatrix} \\
 &= \begin{bmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{bmatrix} = \mathbf{H}_3)
 \end{aligned}$$

Since  $\zeta_1, \dots, \zeta_m$  are distinct and  $l_1, \dots, l_m \neq 0$ ,  $\mathbf{V}_m$  and  $\mathbf{D}_m$  are non-singular; hence  $\mathbf{H}_m$  also is non-singular. Let  $\phi_m$  be the polynomial

$$\phi_m(z) = z^m + b_{m-1}z^{m-1} + \dots + b_0 = \prod_{k=1}^m (z - \zeta_k) \tag{6.447}$$

then the problem of finding the  $n$  zeros of  $p(z)$  reduces to finding the  $m$  zeros of  $\phi_m(z)$  and then computing their multiplicities.  $\phi_m(z)$  is often ill-conditioned, so instead of computing its coefficients, the authors use a type of companion matrix whose eigenvalues are the zeros of  $\phi_m(z)$ .

**Theorem 3.2.** Let  $\zeta_1, \dots, \zeta_m$  be the  $m$  distinct zeros of  $p(z)$  and  $\mathbf{C}_m$  be the Frobenius companion matrix of  $\phi_m(z)$ , then

$$\mathbf{H}_m^{-1} \mathbf{H}_m^< = \mathbf{C}_m \tag{6.448}$$

**Proof.** Let

$$\begin{aligned}
 I_k &= \mu_{k+m} + b_{m-1} \mu_{k+m-1} + \dots + b_0 \mu_k = \\
 &\mu_{k+m} + \sum_{i=0}^{m-1} b_i \mu_{k+i} \quad (k = 0, \dots, m-1)
 \end{aligned} \tag{6.449}$$

Then by 6.441

$$\begin{aligned}
 I_k &= \sum_{i=1}^m l_i \zeta_i^k (\zeta_i^m + b_{m-1} \zeta_i^{m-1} + \dots + b_0) = \\
 &\sum_{i=1}^m l_i \zeta_i^k \phi_m(\zeta_i) = 0 \quad (k = 0, 1, \dots, m-1)
 \end{aligned} \tag{6.450}$$

Hence

$$\mu_{k+m} = - \sum_{i=0}^{m-1} b_i \mu_{k+i} \quad (k = 0, 1, \dots, m-1) \tag{6.451}$$

Thus, from 6.441 and 6.451 we have

$$\mathbf{H}_m \mathbf{C}_m = \begin{bmatrix} \mu_0 & \mu_1 & \dots & \mu_{m-1} \\ \mu_1 & \dots & \dots & \mu_m \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \mu_{2m-3} \\ \mu_{m-1} & \dots & \dots & \mu_{2m-2} \end{bmatrix} \begin{bmatrix} 0 & \dots & \dots & -b_0 \\ 1 & 0 & \dots & -b_1 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & -b_{m-2} \\ 0 & \dots & 1 & -b_{m-1} \end{bmatrix} \tag{6.452}$$

$$= \begin{bmatrix} \mu_1 & \mu_2 & \dots & -\sum_{i=0}^{m-1} \mu_i b_i \\ \mu_2 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \mu_m & \dots & \dots & -\sum_{i=0}^{m-1} \mu_{m+i-1} b_i \end{bmatrix} = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_m \\ \mu_2 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \mu_{2m-2} \\ \mu_m & \dots & \dots & \mu_{2m-1} \end{bmatrix} = \mathbf{H}_m^< \tag{6.453}$$

Since  $\mathbf{H}_m$  is non-singular, the theorem follows.

**Theorem 3.3.** Let  $z_1, \dots, z_m$  be distinct **approximate** zeros of  $\phi_m(z)$ , and define

$$p_m(z) = \det(\mathbf{H}_m^< - z\mathbf{H}_m) / \det(\mathbf{H}_m) \tag{6.454}$$

$$q_m(z) = \prod_{k=1}^m (z - z_k) \tag{6.455}$$

Then the  $m$  distinct zeros of  $p(z)$  are the eigenvalues of

$$\mathbf{A} = \begin{bmatrix} z_1 - \frac{p_m(z_1)}{q'_m(z_1)} & -\frac{p_m(z_1)}{q'_m(z_2)} & \dots & -\frac{p_m(z_1)}{q'_m(z_m)} \\ -\frac{p_m(z_2)}{q'_m(z_1)} & z_2 - \frac{p_m(z_2)}{q'_m(z_2)} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ -\frac{p_m(z_m)}{q'_m(z_1)} & \dots & \dots & z_m - \frac{p_m(z_m)}{q'_m(z_m)} \end{bmatrix} \tag{6.456}$$

**Proof.** Since  $p_m(z) = \det(\mathbf{H}_m^< - z\mathbf{H}_m) / \det(\mathbf{H}_m) = \det(\mathbf{H}_m^{-1}) \det(\mathbf{H}_m^< - z\mathbf{H}_m) = \det(\mathbf{H}_m^{-1} \mathbf{H}_m^< - z\mathbf{I})$ , then by Theorem 3.2

$$p_m(z) = \det(\mathbf{C}_m - z\mathbf{I}) = (-1)^m \phi_m(z) \tag{6.457}$$

Hence the zeros of  $p_m(z)$  are given by  $\zeta_1, \dots, \zeta_m$  (the zeros of  $\phi_m(z)$ ). But by 6.438 the distinct zeros of  $\phi_m(z)$  are the eigenvalues of

$$\mathbf{S} = \begin{bmatrix} z_1 - \frac{\phi_m(z_1)}{q'_m(z_1)} & -\frac{\phi_m(z_1)}{q'_m(z_2)} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ -\frac{\phi_m(z_m)}{q'_m(z_1)} & \dots & \dots & z_m - \frac{\phi_m(z_m)}{q'_m(z_m)} \end{bmatrix} \tag{6.458}$$

(The authors do not make clear what happens to the minus sign when  $m$  is odd) By 6.457  $\mathbf{S}$  is the same as 6.456. Thus the theorem is proved. This matrix is a new companion matrix (for  $\phi$ ) and all the eigenvalues of  $\mathbf{A}$  are simple. They give the distinct zeros  $\zeta_1, \dots, \zeta_m$  of  $p(z)$ .

The integrals  $\mu_s$  (see 6.440) can be approximated via numerical integration, e.g. we can use the K-point Trapezoidal rule on the unit circle. Let  $\omega_j$  be the  $K$ 'th roots of unity, i.e.

$$\omega_j = \exp\left(\frac{2\pi}{K}ij\right), \quad (j = 0, 1, \dots, K - 1) \tag{6.459}$$

Then setting  $z = e^{i\theta}$  in 6.440 gives

$$\mu_s = \frac{1}{2\pi} \int_0^{2\pi} e^{is\theta} e^{i\theta} \frac{p'(e^{i\theta})}{p(e^{i\theta})} d\theta, \quad (s = 0, 1, \dots) \tag{6.460}$$

The Trapezoidal rule approximation of this is, with  $\theta_j = \frac{2\pi}{K}j$ ,

$$\hat{\mu}_s = \frac{1}{K} \sum_{j=0}^{K-1} \frac{p'(\omega_j)}{p(\omega_j)} \omega_j^{s+1}, \quad (s = 0, 1, \dots) \tag{6.461}$$

Let

$$\hat{\mathbf{H}}_m = [\hat{\mu}_{k+i}]_{k,i=0}^{m-1}, \quad \hat{\mathbf{H}}_m^< = [\hat{\mu}_{1+k+i}]_{k,i=0}^{m-1} \tag{6.462}$$

Then  $\hat{\mu}_s$ ,  $\hat{\mathbf{H}}_m$  and  $\hat{\mathbf{H}}_m^<$  are approximations to  $\mu_s$ ,  $\mathbf{H}_m$  and  $\mathbf{H}_m^<$  respectively.

**Theorem 3.4.** Let  $\zeta_1, \dots, \zeta_m$  be the  $m$  distinct zeros of  $p(z)$ , then the corresponding multiplicities  $l_1, \dots, l_m$  are the solutions of the linear system:

$$\sum_{k=1}^m \left(\frac{\zeta_k^s}{1 - \zeta_k^K}\right) l_k = \hat{\mu}_s, \quad (s = 0, 1, \dots, m - 1) \tag{6.463}$$

**Proof.** Since

$$\frac{p'(z)}{p(z)} = \sum_{k=1}^m \frac{l_k}{z - \zeta_k} = \frac{\mu_0}{z} + \frac{\mu_1}{z^2} + \dots \tag{6.464}$$

then by 6.461

$$\hat{\mu}_s = \frac{1}{K} \sum_{j=0}^{K-1} \omega_j^{s+1} \left(\sum_{i=0}^{\infty} \frac{\mu_i}{\omega_j^{i+1}}\right) \tag{6.465}$$

$$= \sum_{i=0}^{\infty} \mu_i \left(\frac{1}{K} \sum_{j=0}^{K-1} \omega_j^{s-i}\right) \tag{6.466}$$

But

$$\frac{1}{K} \sum_{j=0}^{K-1} \omega_j^{s-i} = \begin{cases} 1 & \text{if } s-i = rK \text{ (} r \text{ integer)} \\ 0 & \text{otherwise} \end{cases} \quad (6.467)$$

Hence

$$\hat{\mu}_s = \sum_{r=0}^{\infty} \mu_{s+rK} = \sum_{k=1}^m l_k \zeta_k^s (1 + \zeta_k^K + \zeta_k^{2K} + \dots) \quad (6.468)$$

$$= \sum_{k=1}^m \left( \frac{l_k}{1 - \zeta_k^K} \right) \zeta_k^s \quad (s = 0, 1, \dots, m-1) \quad (6.469)$$

i.e. the theorem is proved.

**Lemma 3.5.** If  $\zeta_1, \dots, \zeta_m$  are distinct zeros of  $p(z)$ , then  $\hat{\mathbf{H}}_m$  is non-singular.

**Proof.** Let

$$\mathbf{U}_m = \begin{bmatrix} \frac{1}{1-\zeta_1^K} & \cdots & \frac{1}{1-\zeta_m^K} \\ \frac{\zeta_1}{1-\zeta_1^K} & \cdots & \frac{\zeta_m}{1-\zeta_m^K} \\ \vdots & \ddots & \vdots \\ \frac{\zeta_1^{m-1}}{1-\zeta_1^K} & \cdots & \frac{\zeta_m^{m-1}}{1-\zeta_m^K} \end{bmatrix} \quad (6.470)$$

Then by 6.463

$$\hat{\mathbf{H}}_m = \mathbf{U}_m \mathbf{D}_m \mathbf{V}_m^T \quad (6.471)$$

where  $\mathbf{V}_m$  and  $\mathbf{D}_m$  are defined in 6.444 and 6.445. (For example with  $m = 2$ ,

$$\begin{aligned} \mathbf{U}_2 \mathbf{D}_2 \mathbf{V}_2^T &= \begin{bmatrix} \frac{1}{1-\zeta_1^K} & \frac{1}{1-\zeta_2^K} \\ \frac{\zeta_1}{1-\zeta_1^K} & \frac{\zeta_2}{1-\zeta_2^K} \end{bmatrix} \begin{bmatrix} l_1 & 0 \\ 0 & l_2 \end{bmatrix} \begin{bmatrix} 1 & \zeta_1 \\ 1 & \zeta_2 \end{bmatrix} \\ &= \begin{bmatrix} \hat{\mu}_0 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\mu}_2 \end{bmatrix} = \hat{\mathbf{H}}_2) \end{aligned}$$

Since  $\zeta_1, \dots, \zeta_m$  are distinct and inside the unit circle,  $\mathbf{U}_m$ ,  $\mathbf{D}_m$  and  $\mathbf{V}_m$  are non-singular, and then so is  $\hat{\mathbf{H}}_m$ .

**Theorem 3.6.** Let  $\zeta_1, \dots, \zeta_m$  be distinct zeros of  $p(z)$ , and let  $\mathbf{C}_m$  be the Frobenius companion matrix of  $\phi_m(z)$ . Let  $\hat{\mathbf{H}}_m$  and  $\hat{\mathbf{H}}_m^<$  be defined as in 6.462, then

$$\hat{\mathbf{H}}_m^{-1} \hat{\mathbf{H}}_m^< = \mathbf{C}_m \quad (6.472)$$

**Proof.** Let  $\hat{I}_k = \hat{\mu}_{k+m} + b_{m-1} \hat{\mu}_{k+m-1} + \dots + b_0 \hat{\mu}_k$ , ( $k = 0, 1, \dots, m-1$ )

Then by 6.463

$$\hat{I}_k = \sum_{i=1}^m \frac{l_i \zeta_i^k}{1 - \zeta_i^K} (\zeta_i^m + b_{m-1} \zeta_i^{m-1} + \dots + b_0) \quad (6.473)$$



$$\sum_{i=1}^m \frac{l_i \zeta_i^k}{1 - \zeta_i^K} \phi_m(\zeta_i) = 0, \quad (k = 0, 1, \dots, m - 1) \tag{6.474}$$

Hence

$$\hat{\mu}_{k+m} = - \sum_{i=0}^{m-1} b_i \hat{\mu}_{k+i}, \quad (k = 0, 1, \dots, m - 1) \tag{6.475}$$

Then similarly to the proof of Theorem 3.2, we have

$$\hat{\mathbf{H}}_m \mathbf{C}_m = \hat{\mathbf{H}}_m^< \tag{6.476}$$

Since  $\hat{\mathbf{H}}_m$  is non-singular, the result follows.

The above theorem means that the error of numerical integration of  $\hat{\mu}_s$  does not affect the result.

In the case that the polynomial has one or more clusters of zeros, suppose that  $\zeta_1, \dots, \zeta_\nu$  form such a cluster. Let  $\zeta_G$  be the arithmetic mean, then  $\zeta_j = \zeta_G + \epsilon_j$ , ( $j = 1, \dots, \nu$ ). If we set

$$\epsilon = \max_{1 \leq j \leq \nu} |\epsilon_j| \tag{6.477}$$

then

$$\begin{aligned} \sum_{j=1}^{\nu} \frac{\zeta_j^s}{1 - \zeta_j^K} &= \sum_{j=1}^{\nu} \frac{(\zeta_G + \epsilon_j)^s}{1 - (\zeta_G + \epsilon_j)^K} = \sum_{j=1}^{\nu} \frac{\zeta_G^s + s\epsilon_j \zeta_G^{s-1} + O(\epsilon_j^2)}{1 - \zeta_G^K - K\epsilon_j \zeta_G^{K-1} + O(\epsilon_j^2)} \\ &= \sum_{j=1}^{\nu} \frac{\zeta_G^s}{1 - \zeta_G^K} + C \sum_{j=1}^{\nu} \epsilon_j + O(\sum \epsilon_j^2) \end{aligned}$$

(for some constant C)

$$= \nu \frac{\zeta_G^s}{1 - \zeta_G^K} + O(\epsilon^2) \tag{6.478}$$

(since  $\sum \epsilon_j = 0$ , as  $\zeta_G$  is ‘‘centre of gravity’’). If the size of the cluster is small enough, we can take its centre as one multiple zero and the number of zeros in the cluster as its multiplicity. Thus we can calculate the centre and the number in the cluster by the new method.

In practise the authors suggest taking  $K = 2m$ , and to apply the algorithm to the general case where the zeros are inside  $\Gamma : \{z : |z - \gamma| < \rho\}$  we set  $P(z) = p(\gamma + z\rho)$  and use  $P(e^{i\theta})$  in place of  $p(e^{i\theta})$ . Let  $\lambda_k$  and  $\zeta_k$  be the zeros of  $P(z)$  and  $p(z)$ , then if the eigenvalues of the companion matrix  $\mathbf{A}$  associated with  $p_m(z)$  are  $\lambda_1, \dots, \lambda_m$ , then the  $\zeta_j = \gamma + \rho\lambda_j$ , ( $j = 1, 2, \dots, m$ ). Since

$$\hat{\mu}_s = \frac{1}{2m} \sum_{j=0}^{2m-1} \frac{P'(\omega_j)}{P(\omega_j)} \omega_j^{s+1} = \frac{1}{2m} \sum_{j=0}^{2m-1} \frac{\rho p'(\gamma + \rho\omega_j)}{p(\gamma + \rho\omega_j)} \omega_j^{s+1} \tag{6.479}$$

and the zeros of  $P(z)$  have the same multiplicities as  $p(z)$ , the multiplicities of  $\zeta_k$  can be calculated from

$$\sum_{k=1}^m \left( \frac{\lambda_k^s}{1 - \lambda_k^{2m}} \right) l_k = \hat{\mu}_s, \quad (s = 0, 1, \dots, m-1) \quad (6.480)$$

The entire algorithm is summarized below:

**Input:**  $c_0, \dots, c_n$ , the polynomial coefficients

$m$ , the number of distinct zeros.

$\gamma, \rho$ , the centre and radius of the circle which contains all the zeros.

$z_1, \dots, z_m$ , mutually distinct approximate zeros.

**Output:** distinct zeros and their multiplicities.

### Calculation

(1) set  $\omega_j = e^{\frac{2\pi j i}{2m}}$ , ( $j = 0, 1, \dots, 2m-1$ )

(2) set  $\hat{\mu}_k = \frac{1}{2m} \sum_{j=0}^{2m-1} \rho \frac{p'(\gamma + \rho\omega_j)}{p(\gamma + \rho\omega_j)} \omega_j^{k+1}$ , ( $k = 0, \dots, 2m-1$ )

(3) set  $\hat{\mathbf{H}}_m = [\hat{\mu}_{s+q}]_{s,q=0}^{m-1}$ ,  $\hat{\mathbf{H}}_m^< = [\hat{\mu}_{s+q}]_{s,q=1}^m$

(4) compute the companion matrix  $\mathbf{A}$  by 6.456

(5) compute  $\lambda_1, \dots, \lambda_m$ , the eigenvalues of  $\mathbf{A}$

(6) set  $\zeta_j = \gamma + \rho\lambda_j$ , ( $j = 1, \dots, m$ )

(7) compute  $l_1, \dots, l_m$  by solving 6.480

In step (4), to get the  $p_m(z_i)$  we need to calculate some Hankel determinants. Fast methods for that problem of  $O(n^2)$  are discussed in Phillips (1971) and Trench (1965).

Some numerical experiments were carried out using MATLAB and double precision arithmetic. In one test case involving a root of multiplicity 4, Fiedler's method (and Smith's) gave errors about  $10^{-4}$  for that root, whereas the new method gave about 14 digit accuracy and correct multiplicities. Also some tests were performed on a case with clusters of roots; in the first the members of the cluster were separated by about  $10^{-8}$ , and the hoped-for result was the centre of the cluster. Again, Fiedler's method gave errors in the 4th place while the new method was correct to about 14. Similar results were obtained with separations of  $10^{-4}$  and  $10^{-6}$ .

The algorithm above assumes that the number of distinct roots is known. The authors suggest that one ascertain this by initially applying some other method, such as Fiedler's, and observing the distances between approximate roots. Those that agree within some tolerance will be regarded as a multiple root, so that we may then apply the new method.

Kravanja, Sakurai, and Van Barel (1999) and Kravanja et al (2000) in related papers describe a rather similar method to that of Niu and Sakurai. Our description

will be based on the second paper mentioned above, and for further details we refer the reader to the first one. The method is developed for general analytic functions, which include polynomials. Let  $W$  be a rectangular region in  $C$ ,  $f : W \rightarrow C$  analytic in the closure of  $W$  and  $\Gamma$  the (positively oriented) boundary of  $W$ . Suppose that  $\Gamma$  does not pass through any of the zeros of  $f$ , and that the edges of  $\Gamma$  are parallel to the coordinate axes. (By finding an upper bound on the magnitude of the roots we can ensure that all the roots of a given polynomial are inside  $W$ ). The authors present a Fortran 90 program, called ZEAL, which will find all the zeros inside  $\Gamma$ , together with their multiplicities. Let  $N$  denote the total number of zeros inside  $\Gamma$ . This is given by

$$N = \frac{1}{2\pi i} \int_{\Gamma} \frac{f'(z)}{f(z)} dz \quad (6.481)$$

Initially, when  $\Gamma$  includes all the zeros,  $N = n$ , the degree of the polynomial (if  $f$  is in fact a polynomial); but later we will sub-divide the region and so 6.481 will be needed. Similarly

$$s_p = \frac{1}{2\pi i} \int_{\Gamma} z^p \frac{f'(z)}{f(z)} dz, \quad (p = 0, 1, \dots) \quad (6.482)$$

Again, for the initial  $\Gamma$ ,  $s_p$  is identical with the  $\mu_s$  of Niu and Sakurai. In general it

$$= \zeta_1^p + \dots + \zeta_N^p, \quad (p = 0, 1, 2, \dots) \quad (6.483)$$

where  $\zeta_1, \dots, \zeta_N$  are the zeros of  $f$  inside  $\Gamma$ . It is known as the Newton sum. We consider the polynomial

$$P_N(z) = \prod_{k=1}^N (z - \zeta_k) \quad (6.484)$$

If  $N$  is large, one needs to calculate the  $s_p$  very accurately, i.e. with multiple precision. To avoid this, and to reduce the number of integrals required, the authors suggest to construct and solve the  $P_N(z)$  only if its degree is  $<$  some given number  $M$  of moderate size such as 5. Otherwise  $\Gamma$  is subdivided and each smaller region treated separately in turn. Kravanja et al, like other authors referred to in this section, consider the distinct roots and their multiplicities separately. Numerical approximations for the integrals in 6.482 and elsewhere are evaluated by QUADPACK (see Piessens et al (1983)). Their method will now be summarized.

As usual, let  $\zeta_1, \dots, \zeta_m$  be the distinct roots of  $f$  inside  $\Gamma$ , and  $l_1, \dots, l_m$  their multiplicities. The authors define, for any two polynomials  $\phi$  and  $\psi$ , "inner products"

$$\langle \phi, \psi \rangle = \frac{1}{2\pi i} \int_{\Gamma} \phi(z) \psi(z) \frac{f'(z)}{f(z)} dz \quad (6.485)$$

$$= \sum_{k=1}^m l_k \phi(\zeta_k) \psi(\zeta_k) \tag{6.486}$$

(this follows because  $\frac{f'}{f}$  has a simple pole at  $\zeta_k$  with residue  $l_k$ ). These  $\langle \phi, \psi \rangle$  can be evaluated by numerical integration. Then

$$s_p = \langle 1, z^p \rangle = \sum_{k=1}^m l_k \zeta_k^p \tag{6.487}$$

In particular  $s_0 = l_1 + \dots + l_m = N$ , the total number of zeros inside  $\Gamma$ . Let  $\mathbf{H}_k$  (as before) be the Hankel matrix

$$\mathbf{H}_k = \begin{bmatrix} s_0 & s_1 & \dots & s_{k-1} \\ s_1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ s_{k-1} & \dots & \dots & s_{2k-2} \end{bmatrix}, \quad (k = 1, 2, \dots) \tag{6.488}$$

A monic polynomial  $\phi_t$  of degree  $t$  that satisfies

$$\langle z^p, \phi_t(z) \rangle = 0, \quad (p = 0, 1, \dots, t - 1) \tag{6.489}$$

is known as a *formal orthogonal polynomial* (FOP). The adjective “formal” comes from the fact that in general the form 6.485 does not necessarily define a true inner product. Consequently FOP’s need not exist or be unique. But if 6.489 is satisfied and  $\phi_t$  is unique, then  $\phi_t$  is called a *regular FOP* and  $t$  a *regular index*. If we set

$$\phi_t(z) = u_{0,t} + u_{1,t}z + \dots + u_{t-1,t}z^{t-1} + z^t \tag{6.490}$$

then 6.489 becomes

$$\begin{bmatrix} s_0 & s_1 & \dots & s_{t-1} \\ s_1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ s_{t-1} & \dots & \dots & s_{2t-2} \end{bmatrix} \begin{bmatrix} u_{0,t} \\ u_{1,t} \\ \dots \\ u_{t-1,t} \end{bmatrix} = - \begin{bmatrix} s_t \\ s_{t+1} \\ \dots \\ s_{2t-1} \end{bmatrix} \tag{6.491}$$

Thus, the regular FOP of degree  $t \geq 1$  exists iff  $\mathbf{H}_t$  is non-singular. We can find  $m$ , the number of distinct zeros, by the following theorem (at least in theory): “ $m = \text{rank}(\mathbf{H}_{m+p})$  for every integer  $p \geq 0$ .” In particular,  $m = \text{rank}(\mathbf{H}_m)$ . So  $\mathbf{H}_m$  is non-singular but  $\mathbf{H}_t$  is singular if  $t > m$ .  $\mathbf{H}_1 = [s_0]$  is non-singular by assumption. The regular FOP of degree 1 exists and is given by:  $\phi_1(z) = z - \mu$  where

$$\mu = \frac{s_1}{s_0} = \frac{\sum_{k=1}^m l_k \zeta_k}{\sum_{k=1}^m l_k} \tag{6.492}$$

is the arithmetic mean of the zeros. For 6.491 becomes

$$[s_0][u_{0,1}] = -s_1 \tag{6.493}$$

The above theorem implies that the regular FOP of degree  $m$  exists, but for degree  $> m$  they do not exist. We may show that

$$\phi_m(z) = (z - \zeta_1)(z - \zeta_2)\dots(z - \zeta_m) \tag{6.494}$$

This follows because 6.489 and 6.486 give

$$\sum_{k=1}^m l_k \phi(\zeta_k) \zeta_k^p = 0, \quad (p = 0, \dots, m - 1) \tag{6.495}$$

i.e. we have a set of  $m$  homogeneous equations in the  $m$  unknowns  $l_k \phi(\zeta_k)$  whose coefficients form a Vandermonde matrix. Since the  $\zeta_k$  are distinct this matrix is non-singular so  $l_k \phi(\zeta_k) = 0$  all  $k$ ; but  $l_k \geq 1$ , so  $\phi(\zeta_k) = 0$  and the result is proved. Once  $m$  is known  $\zeta_1, \dots, \zeta_m$  can be found by solving a generalized eigenvalue problem. For if

$$\mathbf{H}_m^< = \begin{bmatrix} s_1 & \dots & s_m \\ s_2 & \dots & \dots \\ \dots & \dots & \dots \\ s_m & \dots & s_{2m-1} \end{bmatrix} \tag{6.496}$$

then the eigenvalues of  $\mathbf{H}_m^< - \lambda \mathbf{H}_m$  are given by  $\zeta_1, \dots, \zeta_m$  (see 6.448). Once the  $\zeta_i$  have been found the multiplicities  $l_i$  can be found by solving the Vandermonde system

$$\begin{bmatrix} 1 & \dots & 1 \\ \zeta_1 & \dots & \zeta_m \\ \dots & \dots & \dots \\ \zeta_1^{m-1} & \dots & \zeta_m^{m-1} \end{bmatrix} \begin{bmatrix} l_1 \\ \dots \\ l_m \end{bmatrix} = \begin{bmatrix} s_0 \\ s_1 \\ \dots \\ s_{m-1} \end{bmatrix} \tag{6.497}$$

which is 6.487 written for  $p = 0, \dots, m - 1$ . Such systems are often ill-conditioned, but the solutions  $l_i$  are known to be integers, and so can be found accurately as long as the errors in their computed values are  $< .5$  in absolute value. The authors obtain accurate zeros if  $\mathbf{H}_m$  and  $\mathbf{H}_m^<$  are replaced by

$$\mathbf{G}_m = [ < \phi_p, \phi_q > ]_{p,q=0}^{m-1} \text{ and } \mathbf{G}_m^{(1)} = [ < \phi_k, \phi_1 \phi_q > ]_{p,q=0}^{m-1} \tag{6.498}$$

where the  $\phi_i$  are FOP's or related polynomials (see the cited papers for details). It is proved that the eigenvalues of  $\mathbf{G}_m^{(1)} - \lambda \mathbf{G}_m$  are given by  $\zeta_1 - \mu, \dots, \zeta_m - \mu$  where as before  $\mu = \frac{s_1}{s_0}$ . We also refer to the cited papers for details of how we may determine  $m$ . The (approximate)  $\zeta_i$  found by the above method are refined by the modified Newton's method, i.e.

$$z_k^{(i+1)} = z_k^{(i)} - l_k \frac{f'(z_k^{(i)})}{f(z_k^{(i)})}, \quad (k = 1, \dots, m; \quad i = 0, 1, \dots) \tag{6.499}$$

with  $l_k$  the now known multiplicity of  $z_k^{(0)} = \zeta_k$  as determined by the above method. 6.499 is quadratically convergent.

In a test with a function having a triple zero as well as a double one, results were accurate to about 14 digits.

## 6.5 Methods for a Few Roots

Quite often users require only a few zeros of a given polynomial, say the  $k$ -th largest or the  $k$ -th smallest. The inverse power method (mentioned in a previous section) is one method which serves this purpose. Others will be discussed here; for example Gemignani (1998) describes a method which is relatively efficient in this case. In fact a factor of degree  $k$  is found ( $k \ll n$ ), which is usually a better-conditioned problem than that of finding  $k$  separate zeros. He considers a lower Hessenberg matrix of the form:

$$\mathbf{A}_1 = \begin{bmatrix} a_{1,1}^{(1)} & 1 & 0 & \dots & \dots & 0 \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n-1,1}^{(1)} & a_{n-1,2}^{(1)} & \dots & \dots & \dots & 1 \\ a_{n,1}^{(1)} & a_{n,2}^{(1)} & \dots & \dots & \dots & a_{n,n}^{(1)} \end{bmatrix} \quad (6.500)$$

such that

$$\det(t\mathbf{I} - \mathbf{A}_1) = p(t) \quad (6.501)$$

where  $p(t)$  is the polynomial whose zeros are being sought. That is,  $\mathbf{A}_1$  is a companion matrix of  $p(t)$ , such as the Frobenius one. Now the standard LR algorithm with shift defines a sequence of similar matrices by

$$\mathbf{A}_s - \sigma_s \mathbf{I} = \mathbf{L}_s \mathbf{R}_s \quad (6.502)$$

$$\mathbf{A}_{s+1} = \sigma_s \mathbf{I} + \mathbf{R}_s \mathbf{L}_s \quad (s \geq 1) \quad (6.503)$$

where  $\mathbf{L}_s$  is unit lower triangular and  $\mathbf{R}_s$  is upper triangular. As in the QR method  $\mathbf{A}_s$  tends to a triangular or block-triangular matrix whose eigenvalues are the same as those of  $\mathbf{A}_1$ , and can easily be found. The LR method preceded the QR method historically. The above method was generalized by Watkins and Elsner (1991) to

$$p_s(\mathbf{A}_s) = \mathbf{L}_s \mathbf{R}_s \quad (6.504)$$

$$\mathbf{A}_{s+1} = \mathbf{L}_s^{-1} \mathbf{A}_s \mathbf{L}_s \quad (6.505)$$

where  $p_s(t)$  is a monic polynomial of degree  $k_s < n$  (indeed usually  $k_s \ll n$ ). Writing  $p_s(t)$  in factored form (e.g.  $(t - \alpha)(t - \beta)$ ), we find that 6.504 and 6.505 correspond to  $k_s$  steps of 6.502-6.503, where the shifts  $\sigma_s$  are the zeros of  $p_s(t)$ .

Gemignani then defines the polynomials

$$\psi_i^{(s)}(t) = \det(t\mathbf{I} - \hat{\mathbf{A}}_{s,i}) \quad (i = 1, \dots, n-1) \quad (6.506)$$

where  $\hat{\mathbf{A}}_{s,i}$  is the submatrix formed by the first  $i$  rows and columns of  $\mathbf{A}_s$ . The polynomial vector

$$[\psi_0^{(s)}(t), \dots, \psi_{n-1}^{(s)}(t)]^T \tag{6.507}$$

with

$$\psi_0^{(s)}(t) = 1 \tag{6.508}$$

satisfies

$$t \begin{bmatrix} \psi_0^{(s)}(t) \\ \dots \\ \psi_{n-1}^{(s)}(t) \end{bmatrix} = \mathbf{A}_s \begin{bmatrix} \psi_0^{(s)}(t) \\ \dots \\ \psi_{n-1}^{(s)}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \dots \\ 0 \\ p(t) \end{bmatrix} \tag{6.509}$$

**Proof.** Consider the  $(i+1)$ -th element on either side of 6.509 above (ignoring the superscripts on the  $a_{ij}$ ). This gives, using the definition 6.506:

$$\begin{aligned} & t \begin{vmatrix} t - a_{11} & -1 & 0 & \dots & \dots \\ -a_{21} & t - a_{22} & -1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ -a_{i1} & -a_{i2} & \dots & \dots & t - a_{ii} \end{vmatrix} = \\ & a_{i+1,1} + a_{i+1,2}(t - a_{11}) + a_{i+1,3} \begin{vmatrix} t - a_{11} & -1 \\ -a_{21} & t - a_{22} \end{vmatrix} \\ & + \dots + a_{i+1,i+1} \begin{vmatrix} t - a_{11} & -1 & 0 & \dots & \dots \\ -a_{21} & t - a_{22} & -1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ -a_{i1} & -a_{i2} & \dots & \dots & t - a_{ii} \end{vmatrix} \\ & + 1 \begin{vmatrix} t - a_{11} & -1 & 0 & \dots & \dots \\ -a_{21} & t - a_{22} & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & -1 \\ -a_{i+1,1} & -a_{i+1,2} & \dots & \dots & t - a_{i+1,i+1} \end{vmatrix} \end{aligned} \tag{6.510}$$

Expanding the last determinant on the right-hand-side by its last row we find, after much cancellation, that the left-hand-side = the right-hand-side. Applying 6.509 with  $s$  replaced by  $s+1$ , and using 6.505, we get:

$$t \begin{bmatrix} \psi_0^{(s+1)}(t) \\ \dots \\ \psi_{n-1}^{(s+1)}(t) \end{bmatrix} = \mathbf{L}_s^{-1} \mathbf{A}_s \mathbf{L}_s \begin{bmatrix} \psi_0^{(s+1)}(t) \\ \dots \\ \psi_{n-1}^{(s+1)}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \dots \\ \dots \\ p(t) \end{bmatrix} \tag{6.511}$$

It then follows that

$$\begin{bmatrix} \psi_0^{(s+1)}(t) \\ \dots \\ \psi_{n-1}^{(s+1)}(t) \end{bmatrix} = \mathbf{L}_s^{-1} \begin{bmatrix} \psi_0^{(s)}(t) \\ \dots \\ \psi_{n-1}^{(s)}(t) \end{bmatrix} \tag{6.512}$$

Replacing  $\mathbf{A}_s$  in 6.509 by  $\sigma_s \mathbf{I} + \mathbf{L}_s \mathbf{R}_s$  (from 6.502) gives:

$$(t - \sigma_s) \begin{bmatrix} \psi_0^{(s)}(t) \\ \dots \\ \psi_{n-1}^{(s)}(t) \end{bmatrix} = \mathbf{L}_s \mathbf{R}_s \begin{bmatrix} \psi_0^{(s)}(t) \\ \dots \\ \psi_{n-1}^{(s)}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \dots \\ p(t) \end{bmatrix} \tag{6.513}$$

and substituting 6.512 into 6.513 gives

$$(t - \sigma_s) \begin{bmatrix} \psi_0^{(s+1)}(t) \\ \dots \\ \psi_{n-1}^{(s+1)}(t) \end{bmatrix} = \mathbf{R}_s \begin{bmatrix} \psi_0^{(s)}(t) \\ \dots \\ \psi_{n-1}^{(s)}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \dots \\ p(t) \end{bmatrix} \tag{6.514}$$

Since  $\mathbf{A}_s$  is Hessenberg,  $\mathbf{R}_s$  takes the form

$$\mathbf{R}_s = \begin{bmatrix} r_1^{(s)} & 1 & 0 & \dots & \dots \\ 0 & r_2^{(s)} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & r_{n-1}^{(s)} & 1 \\ 0 & \dots & \dots & 0 & r_n^{(s)} \end{bmatrix} \tag{6.515}$$

so that 6.514 may be written as

$$\left\{ \begin{array}{l} (t - \sigma_s) \psi_0^{(s+1)}(t) = r_1^{(s)} \psi_0^{(s)}(t) + \psi_1^{(s)}(t) \\ \dots \\ (t - \sigma_s) \psi_{n-1}^{(s+1)}(t) = r_n^{(s)} \psi_{n-1}^{(s)}(t) + p(t) \end{array} \right\} \tag{6.516}$$

Putting  $t = \sigma_s$  here gives

$$r_1^{(s)} = -\frac{\psi_1^{(s)}(\sigma_s)}{\psi_0^{(s)}(\sigma_s)} \tag{6.517}$$

etc., so that 6.516 becomes:

$$\left\{ \begin{array}{l} (t - \sigma_s) \psi_0^{(s+1)}(t) = -(\psi_1^{(s)}(\sigma_s) / \psi_0^{(s)}(\sigma_s)) \psi_0^{(s)}(t) + \\ \psi_1^{(s)}(t) \dots \dots \\ \dots \\ (t - \sigma_s) \psi_{n-1}^{(s+1)}(t) = -(p(\sigma_s) / \psi_{n-1}^{(s)}(\sigma_s)) \psi_{n-1}^{(s)}(t) + \\ p(t) \dots \dots \end{array} \right\} \tag{6.518}$$



By performing  $k_s$  steps of 6.516 (or 6.518) we may see that 6.504 and 6.505 are equivalent to

$$\left\{ \begin{array}{lcl} p_s(t)\psi_0^{(s+1)}(t) & = & \sum_{i=0}^{k_s} b_{0,i}^{(s)}\psi_i^{(s-k_s+1)}(t) \\ \dots & \dots & \dots \\ p_s(t)\psi_{n-2}^{(s+1)}(t) & = & b_{n-2,n-2}^{(s)}\psi_{n-2}^{(s-k_s+1)}(t) + \\ b_{n-2,n-1}^{(s)}\psi_{n-1}^{(s-k_s+1)}(t) & + & g_{n-2}^{(s)}(t)p(t) \\ p_s(t)\psi_{n-1}^{(s+1)}(t) & = & b_{n-1,n-1}^{(s)}\psi_{n-1}^{(s-k_s+1)}(t) \\ +g_{n-1}^{(s)}(t)p(t) & \dots & \dots \end{array} \right\} \tag{6.519}$$

where the  $b_{i,j}^{(s)}$  are suitable scalars and  $g_{n-i}^{(s)}(t)$  ( $i = 1, 2$ ) are monic polynomials of degree  $k_s - i$ . N.B. the reader should not confuse  $p_s(t)$ , of degree  $k_s$ , with  $p(t)$ , the original polynomial of degree  $n$  (for which we are seeking some of the roots).

Initially the author suggests taking

$$p_1(t) = (t - a)^k \tag{6.520}$$

where we seek the  $k$  zeros of  $p(t)$  closest to  $a$ , and setting

$$p_s(t) = p_1(t) \quad (s = 1, 2, \dots) \tag{6.521}$$

Suppose that  $p_1(t)$  separates  $h$  zeros of  $p(t)$  from the others, in the sense that

$$\begin{aligned} |p_1(t_1)| &\geq |p_1(t_2)| \geq \dots \geq |p_1(t_h)| > |p_1(t_{h+1})| \geq \dots \geq \\ |p_1(t_n)| &> 0 \end{aligned} \tag{6.522}$$

where

$$\frac{|p_1(t_{h+1})|}{|p_1(t_h)|} \ll \frac{|p_1(t_{h+j+1})|}{|p_1(t_{h+j})|} \quad (j = 1, \dots, n - h - 1) \tag{6.523}$$

and the  $t_i$  are the zeros of  $p(t)$ . For example, if  $k = 4$  and  $t_n = t_{n-1} = t_{n-2} = t_{n-3} = .1$ ,  $t_{n-4} = 1$ , and  $p_1(t) = t^4$  then 6.523 will be satisfied for  $h = n - 4$ . Then, let  $\psi_{n-j}^{(0)}(t)$  ( $j = 1, \dots, n - h = k$ ) be arbitrary monic polynomials of degree  $n - j$  respectively (in his numerical experiments Gemignani takes them as the derivatives of  $p(t)$  of order  $j$ ), and for  $s = 1, 2, \dots$  compute  $\psi_{n-j}^{(s)}$  ( $j = 1, \dots, n - h$ ) by means of the last  $n - h$  equalities of 6.519 (using  $p_s(t) = p_1(t)$  for all  $s$ —this being known as a stationary iteration). Theoretically the process could break down due to  $\mathbf{A}_s - \sigma_s \mathbf{I}$  having a leading principal minor equal to zero. But the author shows that the probability of this happening is very low. Moreover he proves that

$$\|\psi_h^{(s+1)}(t) - \prod_{i=1}^h (t - t_i)\|_\infty = O\left(\left|\frac{p_1(t_{h+1})}{p_1(t_h)}\right|^{s+1}\right) \tag{6.524}$$

(Here  $\|q(t)\|_\infty = \max_{0 \leq i \leq n} |q_i|$ ,  $q_n = 1$  for a polynomial  $q(t) = \sum_{i=0}^n q_i(t)$ ). For more details see the cited paper. Now suppose  $\{\psi_{n-k}^{(s)}(t)\}_{s=1,2,\dots}$  is a sequence of polynomials of degree  $n-k$  generated by the stationary iteration, and let the polynomials  $\eta_k^{(s)}(t)$  be defined by

$$p(t) = \eta_k^{(s)}(t)\psi_{n-k}^{(s)}(t) + \theta^{(s)} \tag{6.525}$$

where  $\deg(\theta^{(s)}(t)) < \deg(\psi_{n-k}^{(s)}(t))$ . When  $\psi_{n-k}^{(s)}(t)$  approaches  $\prod_{i=1}^{n-k} (t-t_i)$ , which is usually the case, then  $\eta_k^{(s)}(t)$  should converge to

$$\prod_{i=n-k+1}^n (t-t_i) \tag{6.526}$$

Thus we add the following feature to the stationary iteration, after computing a new set  $\psi_{n-j}^{(s)}(t)$  ( $j = 1, \dots, k$ ):

Compute  $\eta_k^{(s)}(t)$  and check for convergence, i.e. test whether

$$\|\eta_k^{(s+1)}(t) - \eta_k^{(s)}(t)\|_\infty \leq u(n+k)\|p(t)\|_\infty \tag{6.527}$$

where  $u$  is the machine precision. The algorithm is halted if 6.527 is satisfied or if  $s >$  some predefined value **imax**. In the latter case the procedure reports failure. The author shows that convergence of this modified stationary iteration is linear, but by varying  $p_s(t)$  after a certain value of  $s$  (i.e. computing  $\eta_k^{(s)}(t)$ , checking for convergence, and setting  $p^{(s+1)}(t) = \eta_k^{(s)}(t)$ ) we may achieve quadratic convergence (again, see the cited paper for proof). We shift to the variable  $p_s(t)$  when

$$\| \|\eta_k^{(s+1)}(t)\|_\infty - \|\eta_k^{(s)}(t)\|_\infty \| \leq a \text{ constant } \eta \text{ (say } .01) \tag{6.528}$$

It is stated that the process may be ill-conditioned, but poor results at one step may be corrected by later iterations at the cost of taking more iterations. Also, it is stated that each iteration will cost

$$O(k^4 + nk^3) \tag{6.529}$$

operations.

Some numerical tests involving multiple roots or clusters were unsuccessful. The author also considered some random polynomials of degree 15 with real and imaginary parts of the coefficients in the range  $[0,1]$ ,  $p_1(t) = (t-1)^4$  and  $k = 4$  (i.e. he was seeking zeros close to 1). These tests were performed quite successfully, although increasing the degree to 20 resulted in failure in 10% of cases. In a further test, a cluster of degree  $k$  ( $k = 5$  or  $10$ ) was multiplied by 100 random polynomials of degree  $n-k$ , for  $n$  in steps of 50 up to 200. The factorization of the resulting polynomial into factors of degree  $k$  and  $n-k$  was performed successfully in all cases,

at a cost of about 6 or 7 iterations per polynomial.

Gemignani (private communication) points out that the above LR-based method is not perfect, but that a similar treatment by the QR method (for small  $k$ ) may perform much better. Such a treatment has not been worked out yet (June 2006), but is expected in the near future.

### 6.6 Errors and Sensitivity

It is well known that the two problems of finding polynomial zeros and finding eigenvalues of non-symmetric matrices may be highly sensitive to perturbations. And of course the two problems are related, for the zeros of a polynomial are the same as the eigenvalues of any of its companion matrices. Toh and Trefethen (1994), at a time when interest in matrix methods for polynomial zeros was increasing, considered the relationship between the sensitivities of these two problems, and showed that under certain circumstances the sensitivities are very close. They proceed as follows: for a monic polynomial  $p(z)$ , let  $Z(p)$  denote the set of zeros of  $p(z)$ , and for any  $\epsilon \geq 0$ , define the  $\epsilon$ -pseudozero set of  $p(z)$  by

$$Z_\epsilon(p) = \{z \in C : z \in Z(\hat{p}) \text{ for some } \hat{p}\} \tag{6.530}$$

where  $\hat{p}$  ranges over all polynomials whose coefficients are those of  $p$  modified by perturbations of size  $\leq \epsilon$ . Similarly, for a matrix  $\mathbf{A}$ , let  $\Lambda(\mathbf{A})$  denote the set of eigenvalues of  $\mathbf{A}$  (i.e. its spectrum), and define the  $\epsilon$ -pseudospectrum of  $\mathbf{A}$  by

$$\Lambda_\epsilon(\mathbf{A}) = \{z \in C : z \in \Lambda(\mathbf{A} + \mathbf{E}) \text{ for some } \mathbf{E} \text{ with } \|\mathbf{E}\| \leq \epsilon\} \tag{6.531}$$

The authors report numerical experiments which show that  $Z_{\epsilon\|\mathbf{p}\|}(p)$  and  $\Lambda_{\epsilon\|\mathbf{A}\|}(\mathbf{A})$  are generally quite close to each other when  $\mathbf{A}$  is a companion matrix of  $p$  that has been “balanced” in the sense first described by Parlett and Reinsch (1969). It follows that the zero-finding and balanced eigenvalue problems are comparable in conditioning, and so finding roots via eigenvalues of companion matrices is a stable algorithm. Note that Toh and Trefethen consider only the classical or Frobenius companion matrix.

We define some notation:

$P$  is the set of monic polynomials of degree  $n$ .

$p(z)$  is the polynomial  $z^n + c_{n-1}z^{n-1} + \dots + c_0$ . Sometimes we denote the vector of coefficients  $(c_0, \dots, c_{n-1})^T$  by  $\mathbf{p}$ .

$p^*$  is the reciprocal polynomial of  $p$  i.e.

$$p^*(z) = z^n p\left(\frac{1}{z}\right) \tag{6.532}$$

$\mathbf{D}$  is an  $n \times n$  diagonal matrix with diagonal vector  $\mathbf{d} = (d_0, \dots, d_{n-1})^T$ , or we may write  $\mathbf{D} = \text{diag}(\mathbf{d})$ .  $\mathbf{d}^{-1}$  denotes  $(d_0^{-1}, \dots, d_{n-1}^{-1})^T$ , and similarly  $\mathbf{p}^{-1}$  denotes

$(c_0^{-1}, \dots, c_{n-1}^{-1})^T$  etc.

$$\|\mathbf{x}\|_d = \|\mathbf{D}\mathbf{x}\|_2 \tag{6.533}$$

(provided  $\mathbf{D}$  is non-singular).

For given  $z$ ,  $\tilde{\mathbf{z}}$  = the vector  $(1, z, \dots, z^{n-1})^T$ .

For  $i = 1, \dots, n$   $\mathbf{e}_i$  has 1 in the  $i$ 'th position and 0's elsewhere.

Then

$$\|\mathbf{p} - \hat{\mathbf{p}}\|_d = \left[ \sum_{i=0}^{n-1} |d_i|^2 |c_i - \hat{c}_i|^2 \right]^{\frac{1}{2}} \tag{6.534}$$

measures the perturbations in the coefficients of  $p$  relative to the weights given by  $\mathbf{d}$ . Also we have

$$\|\mathbf{A}\|_d = \|\mathbf{DAD}^{-1}\|_2 \tag{6.535}$$

Now we define the  $\epsilon$ -pseudozero set of  $p$  a little more formally as follows:

$$Z_\epsilon(p; d) = \{z \in C : z \in Z(\hat{p}) \text{ for some } \hat{p} \in P$$

$$\text{with } \|\hat{\mathbf{p}} - \mathbf{p}\|_d \leq \epsilon\} \tag{6.536}$$

These sets quantify the conditioning of the zerofinding problem: for a zerofinding algorithm to be stable the computed zeros of  $p$  should lie in a region  $Z_{Cu}(p; d)$  where  $C = O(\|p\|_d)$  and  $u$  is the machine precision.  $\mathbf{d}$ , for example, may be

$$\|p\|_2 \mathbf{p}^{-1} \tag{6.537}$$

for coefficientwise perturbations, or

$$\sqrt{n}(1, 1, \dots, 1)^T \tag{6.538}$$

for normwise perturbations.

**Proposition 6.1**

$$Z_\epsilon(p; d) = \{z \in C : \frac{|p(z)|}{\|\tilde{\mathbf{z}}\|_{d^{-1}}} \leq \epsilon\} \tag{6.539}$$

**Proof.** If  $z \in Z_\epsilon(p; d)$ , then  $\exists \hat{p} \in P$  with  $\hat{p}(z) = 0$  and  $\|\hat{\mathbf{p}} - \mathbf{p}\|_d \leq \epsilon$  Now the Holder inequality (see e.g Hazewinkel (1989) or Hardy (1934)) states that

$$\sum_i |x_i y_i| \leq \left(\sum_i |x_i|^p\right)^{\frac{1}{p}} \left(\sum_i |y_i|^q\right)^{\frac{1}{q}}; \frac{1}{p} + \frac{1}{q} = 1 \tag{6.540}$$

Setting  $p = q = 2$ ,  $x_i = (\hat{c}_i - c_i)d_i$ ,  $y_i = \frac{z^i}{d_i}$  in 6.540 gives  $|p(z)| = |\hat{p}(z) - p(z)| = \left| \sum_{i=0}^{n-1} (\hat{c}_i - c_i)z^i \right| = \left| \sum_{i=0}^{n-1} (\hat{c}_i - c_i)d_i \frac{z^i}{d_i} \right| \leq \|\hat{\mathbf{p}} - \mathbf{p}\|_d \|\tilde{\mathbf{z}}\|_{d^{-1}}$

$$\leq \epsilon \|\tilde{\mathbf{z}}\|_{d^{-1}} \tag{6.541}$$

$$i.e. \frac{|p(z)|}{\|\tilde{\mathbf{z}}\|_{d-1}} \leq \epsilon \tag{6.542}$$

Conversely, suppose  $z \in C$  is such that 6.541 is satisfied. Let  $\theta = arg(z)$ , and consider the polynomial  $r$  of degree  $n-1$  defined by

$$r(w) = \sum_{k=0}^{n-1} r_k w^k \tag{6.543}$$

where

$$r_k = |z^k| e^{-ik\theta} d_k^{-2} \tag{6.544}$$

Then the authors state that

$$r(z) = \|\mathbf{r}\|_d \|\tilde{\mathbf{z}}\|_{d-1} \tag{6.545}$$

Then the polynomial  $\hat{p}$  defined by

$$\hat{p}(w) = p(w) - \frac{p(z)}{r(z)} r(w) \tag{6.546}$$

satisfies  $\hat{p}(z) = 0$  and

$$\|\hat{\mathbf{p}} - \mathbf{p}\|_d = \left| \frac{p(z)}{r(z)} \right| \|\mathbf{r}\|_d \leq \epsilon \frac{\|\tilde{\mathbf{z}}\|_{d-1} \|\mathbf{r}\|_d}{|r(z)|} = \epsilon \tag{6.547}$$

Thus  $z \in Z_\epsilon(p; d)$ .

The authors define a condition number in the case of infinitesimal perturbations, i.e. the condition number of the root  $\zeta$  of  $p$  is given by

$$\kappa(\zeta, p; d) = \lim_{\|\hat{\mathbf{p}} - \mathbf{p}\|_d \rightarrow 0} \sup_{\hat{\mathbf{p}}} \frac{|\hat{\zeta} - \zeta|}{\|\hat{\mathbf{p}} - \mathbf{p}\|_d / \|\mathbf{p}\|_d} \tag{6.548}$$

The authors state that the above =

$$\|\mathbf{p}\|_d \frac{\|\tilde{\zeta}\|_{d-1}}{|p'(\zeta)|} \tag{6.549}$$

We define the condition number of the entire zerofinding problem for  $p$  to be

$$\kappa(p; d) = \max_{\zeta} \kappa(\zeta, p; d) \tag{6.550}$$

We turn now to consider the pseudospectrum of the companion matrix

$$\mathbf{A}_p = \begin{bmatrix} 0 & \dots & 0 & -c_0 \\ 1 & 0 & \dots & -c_1 \\ \dots & \dots & \dots & \dots \\ \dots & 0 & 1 & -c_{n-1} \end{bmatrix} \tag{6.551}$$

For each  $p \in P$ , we define the  $\epsilon$ -pseudospectrum of  $\mathbf{A}_p$  (again, more formally) by

$$\Lambda_\epsilon(\mathbf{A}_p; d) = \{z \in C : z \in \Lambda(\hat{\mathbf{A}}) \text{ for some } \hat{\mathbf{A}} \text{ with} \\ \|\hat{\mathbf{A}} - \mathbf{A}_p\|_d \leq \epsilon\} \quad (6.552)$$

Note that

$$\|\hat{\mathbf{A}} - \mathbf{A}_p\|_d = \|\mathbf{D}(\hat{\mathbf{A}} - \mathbf{A}_p)\mathbf{D}^{-1}\|_2 \quad (6.553)$$

$\mathbf{D}$  is included because balancing and other transformations (see later) involve a diagonal similarity transformation. For an eigenvalue algorithm applied to a companion matrix to be stable, the computed eigenvalues of  $\mathbf{A}_p$  should lie in a region  $\Lambda_{Cu}(\mathbf{A}_p; d)$  for some  $C = O(\|\mathbf{A}(p)\|_d)$ .

We define the condition number of a simple eigenvalue  $\lambda$  of a matrix  $\mathbf{B}$  by

$$\kappa(\lambda, \mathbf{B}; d) = \lim_{\|\hat{\mathbf{B}} - \mathbf{B}\|_d \rightarrow 0} \sup_{\|\hat{\mathbf{B}}\|} \frac{|\hat{\lambda} - \lambda|}{\|\hat{\mathbf{B}} - \mathbf{B}\|_d / \|\mathbf{B}_d\|} \quad (6.554)$$

and this reduces to

$$\|\mathbf{B}\|_d \frac{\|\mathbf{x}\|_{d-1} \|\mathbf{y}\|_d}{|\bar{\mathbf{x}}^T \mathbf{y}|} \quad (6.555)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are left and right eigenvectors of  $\mathbf{B}$ , respectively, corresponding to  $\lambda$ . When  $\mathbf{B} = \mathbf{A}_p$  we have

$$\mathbf{x} = (1, \lambda, \dots, \lambda^{n-1})^T \quad (6.556)$$

and

$$\mathbf{y} = (b_0, b_1, \dots, b_{n-1})^T \quad (6.557)$$

where  $b_0, \dots, b_{n-1}$  (functions of  $\lambda$ ) are the coefficients of

$$\frac{p(z) - p(\lambda)}{z - \lambda} = \sum_{i=0}^{n-1} b_i z^i \quad (6.558)$$

Then the condition number reduces to

$$\kappa(\lambda, \mathbf{A}_p; d) = \|\mathbf{A}_p\|_d \frac{\|\tilde{\mathbf{b}}(\lambda)\|_d \|\tilde{\lambda}\|_{d-1}}{|p'(\lambda)|} \quad (6.559)$$

where

$$\tilde{\mathbf{b}}(\lambda) = (b_0, \dots, b_{n-1})^T \quad (6.560)$$

(same as  $\mathbf{y}$  above). If the eigenvalues of  $\mathbf{A}_p$  are simple, we can define a condition number for the entire problem of finding the eigenvalues of  $\mathbf{A}_p$  as

$$\kappa(\mathbf{A}_p; d) = \max_{\lambda} \kappa(\lambda, \mathbf{A}_p; d) \quad (6.561)$$

The conditioning of the eigenvalue problem for a companion matrix  $\mathbf{A}_p$  may be changed enormously by a diagonal similarity transformation. The best one could hope for would be a transformation which makes the eigenvalue problem no worse conditioned than the underlying zerofinding problem. The authors report experiments which show that up to a factor of about 10, balancing achieves this optimal result. They consider four possibilities of transforming to  $\mathbf{D}\mathbf{A}_p\mathbf{D}^{-1}$ :

1) **Pure companion matrix** i.e. no transformation, or equivalently

$$\mathbf{d} = \sqrt{n}(1, 1, \dots, 1)^T.$$

2) **Balancing**. This corresponds to finding a diagonal matrix  $\mathbf{T}$  such that  $\mathbf{T}\mathbf{A}_p\mathbf{T}^{-1}$  has the 2-norm of its  $i$ 'th row and  $i$ 'th column approximately equal for each  $i = 1, \dots, n$ . We denote this transformation by  $t$ . It is the standard option in EISPACK (see Smith (1976)) and the default in MATLAB (see MathWorks (1992)).

3) **Scaling**. For  $\alpha > 0$ , the scaled polynomial corresponding to  $p \in P$  is

$$p_\alpha(z) = \frac{1}{\alpha^n} p(\alpha z) = \sum_{i=0}^n \frac{c_i}{\alpha^{n-i}} z^i \quad (6.562)$$

The corresponding diagonal similarity transformation is said to be defined by  $\mathbf{D}_\alpha = \text{diag}(\mathbf{d}^{(\alpha)})$  where

$$\mathbf{d}^{(\alpha)} = \|(\alpha^n, \dots, \alpha^1)\|_2 (\alpha^{-n}, \dots, \alpha^{-1})^T \quad (6.563)$$

4) **Coefficientwise** (if coefficients are all non-zero). This is given by the diagonal matrix  $\mathbf{C} = \text{diag}(\mathbf{c})$  where

$$\mathbf{c} = \|\mathbf{p}\|_2 \mathbf{p}^{-1} \quad (6.564)$$

The authors found that balancing tends to achieve the best conditioned eigenvalue problem for  $\mathbf{A}_p$  among these four choices. That is, let us consider the ratios of the condition numbers for the other 3 choices to that of balancing, i.e.

$$\sigma(\mathbf{A}_p; d) = \frac{\kappa(\mathbf{A}_p; d)}{\kappa(\mathbf{A}_p; t)} \quad (6.565)$$

where  $d = c, d^{(\alpha)},$  or  $e,$  and  $t$  refers to balancing. In the case of scaling,  $\alpha$  is chosen to be optimal in the sense that  $\sigma(\mathbf{A}_p; d^{(\alpha)})$  is minimized. It was found that  $\sigma(\mathbf{A}_p; e)$  and  $\sigma(\mathbf{A}_p; d^{(\alpha)})$  are  $\gg 1,$  meaning that the use of the pure companion matrix or scaling lead to much worse conditioning than balancing does.  $\sigma(\mathbf{A}_p; c)$  is often close to 1, but coefficientwise transformations are not defined if some of the coefficients are zero.

Next the authors compare the condition number for the balanced eigenvalue problem with that of the coefficientwise perturbed zerofinding problem for  $p,$  i.e. they consider the ratio

$$\frac{\kappa(\mathbf{A}_p; t)}{\kappa(p; c)} \quad (6.566)$$

Their experiments indicate that this ratio is fairly close to 1, ranging from 2.6 to 21 when applied to 8 well-known polynomials such as Wilkinson's  $\prod_{i=1}^{20} (z - i)$ . They also performed tests on 100 random polynomials of degree 10. In this case the ratio varied from about 10 for well-conditioned polynomials to about  $10^3$  for ones having condition number near  $10^{10}$ . In the case of the eight polynomials referred to above, the authors graphically compared two pseudozero sets with the two pseudospectra of the corresponding balanced companion matrix. The two sets were derived by using two different values of  $\epsilon$  (which varied from case to case). A reasonably close agreement was observed in all cases. This suggests that it ought to be possible to compute zeros of polynomials stably via eigenvalues of companion matrices. To test this assumption, the authors compared three zerofinding methods:

- 1) J-T i.e. the Jenkins-Traub program CPOLY (see Jenkins and Traub (1970)). This is available from ACM TOMS via Netlib and also is in the IMSL library.
- 2) M-R i.e. the Madsen-Reid code PA16 from the Harwell library, see Madsen and Reid (1975). This is a Newton-based method coupled with line search. (At the time the article by Toh and Trefethen was written the above two programs were considered state-of-the-art).
- 3) ROOTS, the Matlab zerofinding code based on finding eigenvalues of the balanced companion matrix by standard methods, see Moler (1991).

In their experiments, the authors first find the “exact” roots of  $p$  by computing the eigenvalues of  $\mathbf{A}_p$  in quadruple precision via standard EISPACK routines. The rest of the calculations were carried out in double precision. For each of the eight polynomials referred to previously, they calculated the maximum absolute error of the roots as found by the above three methods, as well as the condition numbers of the coefficientwise perturbed zerofinding problem for  $p$ , and the balanced companion matrix eigenvalue problem for  $\mathbf{A}_p$ . The roots are always accurate (by all three methods) to at least 11 or 12 decimal places, except for the Wilkinson polynomial, which is notoriously ill-conditioned. The M-R code is a little more accurate than ROOTS, which in turn is a little more accurate than J-T. For the random degree-10 polynomials, it was found that M-R and ROOTS are always stable, while J-T is sometimes not (i.e. the errors are much greater than condition number\* $u$ ). In the case of multiple zeros, ROOTS is sometimes mildly unstable. The authors point out that their results are inexact and empirical, and do not necessarily apply to *all* polynomials. But they consider that a reasonable degree of confidence in zerofinding via companion matrix eigenvalues is justified.

Edelman and Murakami (1995) also consider the question of perturbations in companion matrix calculations. They ask the question “what does it mean to say that

$$\hat{p}(x) = x^n + \hat{c}_{n-1}x^{n-1} + \dots + \hat{c}_0 \quad (6.567)$$

is a slight perturbation of  $p(x) = x^n + c_{n-1}x^{n-1} + \dots + c_0$ ?” (or in other words, that the calculation is stable). Here the computed roots  $\hat{\zeta}_i$  ( $i = 1, \dots, n$ ) of  $p(x)$



(computed by the eigenvalue method) are the exact roots of  $\hat{p}(x)$ . They give four answers, of which we quote the fourth (said to be the best). It is that

$$\max_i \frac{|c_i - \hat{c}_i|}{|c_i|} = O(u) \tag{6.568}$$

where  $O(u)$  means a small multiple of machine precision. We call the above a small *coefficientwise* perturbation. If  $\mathbf{C}$  is the usual Frobenius companion matrix (so that  $P_C(z) \equiv \det(z\mathbf{I} - \mathbf{C}) = p(z)$ ), and  $\mathbf{E}$  is a perturbation matrix with “small” entries, we are interested in the computation of

$$P_{C+E}(z) - P_C(z) \equiv \delta c_{n-1}z^{n-1} + \dots + \delta c_1z + \delta c_0 \tag{6.569}$$

The authors state a theorem as follows: “To first order, the coefficient of  $z^{k-1}$  in  $P_{C+E}(z) - P_C(z)$  is

$$\sum_{m=0}^{k-1} c_m \sum_{i=k+1}^n E_{i,i+m-k} - \sum_{m=k}^n c_m \sum_{i=1}^k E_{i,i+m-k} \tag{6.570}$$

where  $c_n$  is defined as 1”.

This means that a small perturbation  $\mathbf{E}$  introduces errors in the coefficients that are linear in the  $E_{i,j}$ . Since standard eigenvalue procedures compute eigenvalues of matrices with a small backward error, we can claim that there is a polynomial near  $P_C(z)$  whose exact roots are computed by solving an eigenvalue problem. The result is stated in a matrix-vector format: let

$$f_{k,d} \equiv \sum_{i=1}^k E_{i,i+d} \text{ and } b_{k,d} = \sum_{i=k}^n E_{i,i+d} \tag{6.571}$$

then

$$\begin{bmatrix} \delta c_0 \\ \delta c_1 \\ \dots \\ \delta c_{n-1} \end{bmatrix} = \begin{bmatrix} b_{2,-1} & -f_{1,0} & -f_{1,1} & \dots & -f_{1,n-3} & -f_{1,n-2} & -f_{1,n-1} \\ b_{3,-2} & b_{3,-1} & -f_{2,0} & \dots & -f_{2,n-4} & -f_{2,n-3} & -f_{2,n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ b_{n,-(n-1)} & b_{n,-(n-2)} & \dots & \dots & b_{n,-1} & -f_{n-1,0} & -f_{n-1,1} \\ 0 & 0 & 0 & 0 & 0 & 0 & -f_{n,0} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \dots \\ c_{n-1} \\ c_n = 1 \end{bmatrix} \tag{6.572}$$

correct to first order. The last row states that perturbing the trace of a matrix perturbs the coefficient of  $z^{n-1}$  by the same amount. If  $\mathbf{E}$  is the backward error resulting from a standard eigenvalue routine, it is nearly upper triangular, i.e. there may also be up to two non-zero subdiagonals, but no more.

The authors use their theorem above to predict the componentwise backward error; and they perform numerical experiments to actually measure this error. Their results show that the theory always predicts a small backward error and is pessimistic by at most one or two (occasionally three) digits. They consider an error matrix  $\mathbf{E}$  with entries  $\epsilon = 2^{-52}$  in all elements  $(i, j)$  with  $j - i \geq -2$ . For example, when  $n = 6$

$$\mathbf{E} = \begin{bmatrix} \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\ 0 & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\ 0 & 0 & \epsilon & \epsilon & \epsilon & \epsilon \\ 0 & 0 & 0 & \epsilon & \epsilon & \epsilon \end{bmatrix} \quad (6.573)$$

This allows for the possibility of double shifting in the eigenvalue algorithm. The standard algorithms balance the matrix by finding a diagonal matrix  $\mathbf{T}$  such that  $\mathbf{B} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$  has a smaller norm than  $\mathbf{A}$ . The authors assume that the eigenvalue algorithm computes the exact eigenvalues of a matrix  $\mathbf{B} + \mathbf{E}'$  where  $|\mathbf{E}'| \leq \mathbf{E}$ . Thus we are computing the exact eigenvalues of  $\mathbf{A} + \mathbf{T}\mathbf{E}'\mathbf{T}^{-1}$ . So to first order the error in the coefficients is bounded by the absolute value of the matrix times the absolute value of the vector in the product given in 6.572 where the  $f_{i,j}$  and  $b_{i,j}$  are computed using  $\mathbf{T}\mathbf{E}\mathbf{T}^{-1}$ . Thus the  $\delta_i$  are predicted. Edelman and Murakami applied their tests to the same eight “well-known” polynomials as Toh and Trefethen. For each polynomial the coefficients were computed exactly or with 30- decimal precision. For each coefficient of each polynomial the predicted error according to 6.572 was calculated, i.e. the  $\delta c_i$ , and hence  $\log_{10}(\frac{\delta c_i}{c_i})$ . Next the eigenvalues were computed using MATLAB, then the exact polynomial  $\hat{p}$  using these computed roots, and hence  $\log_{10}(\frac{\hat{c}_i - c_i}{c_i})$  (N.B.  $\hat{c}_i - c_i$  is the backward error). Finally the authors computed a “pessimism index”, namely

$$\log_{10}\left(\frac{\hat{c}_i - c_i}{\delta c_i}\right) \quad (6.574)$$

Indices such as 0, -1, -2 indicate that we are pessimistic by at most two orders of magnitude. The results show that the predicted error is usually correct to about 13 decimal places; the actual computed error is correct to about 15 places (this is further indication that eigenvalue-based zerofinding calculations are reliable); and the pessimism index is usually between one and three. It is never positive, indicating that the predictions are “fail-safe”. Thus the analysis described in this paper would be a good way of confirming the accuracy of an eigenvalue-based zerofinder. The authors give a MATLAB program which was used in the experiments, and

presumably could also be used in the suggested accuracy confirmation.

Near the start of this section, and later, we mentioned the process of “balancing” a matrix. This is used in nearly all software for finding zeros via eigenvalues of companion matrices. It is useful because, as pointed out by Osborne (1960), most eigenvalue programs produce results with errors of order at least  $u\|\mathbf{A}\|_E$ , where as usual  $u$  is the machine precision and  $\|\mathbf{A}\|_E$  is the Euclidean norm (see later). Hence he recommends that we precede the calling of such a program by a diagonal similarity transformation of  $\mathbf{A}$  which will reduce its norm (while preserving its eigenvalues). Let

$$\|\mathbf{x}\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} \quad (6.575)$$

and

$$\|\mathbf{A}\|_p = \left( \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^p \right)^{\frac{1}{p}} \quad (6.576)$$

(the case  $p = 2$  gives the Euclidean norm). Parlett and Reinsch (1969) describe an algorithm which produces a sequence of matrices  $\mathbf{A}_k$  ( $k=1,2,\dots$ ), diagonally similar to  $\mathbf{A}$ , such that for an irreducible  $\mathbf{A}$ :

- (i)  $\mathbf{A}_f = \lim_{k \rightarrow \infty} \mathbf{A}_k$  exists and is diagonally similar to  $\mathbf{A}$ .
- (ii)

$$\|\mathbf{A}_f\|_p = \inf(\|\mathbf{D}^{-1}\mathbf{A}\mathbf{D}\|_p) \quad (6.577)$$

where  $\mathbf{D}$  ranges over the class of all non-singular diagonal matrices.

- (iii)  $\mathbf{A}_f$  is balanced, i.e.

$$\|\mathbf{a}_i\|_p = \|\mathbf{a}^i\|_p \quad (i = 1, \dots, n) \quad (6.578)$$

where  $\mathbf{a}_i$  and  $\mathbf{a}^i$  denote respectively the  $i$ -th column and  $i$ -th row of  $\mathbf{A}_f$ .

No rounding errors need occur in this process if the elements of the diagonal matrix are restricted to be exact powers of the radix base (usually 2). In more detail, let  $\mathbf{A}_0$  denote the off-diagonal part of  $\mathbf{A}$ . Then for any non-singular diagonal matrix  $\mathbf{D}$ , we have

$$\mathbf{D}^{-1}\mathbf{A}\mathbf{D} = \text{diag}(\mathbf{A}) + \mathbf{D}^{-1}\mathbf{A}_0\mathbf{D} \quad (6.579)$$

and only  $\mathbf{A}_0$  is affected. Assume that no row or column of  $\mathbf{A}_0$  vanishes identically. From  $\mathbf{A}_0$  a sequence  $\{\mathbf{A}_k\}$  is formed. The term  $\mathbf{A}_k$  differs from  $\mathbf{A}_{k-1}$  in only one row and the corresponding column. Let  $k = 1, 2, \dots$  and let  $i$  be the index of the row and column modified in the step from  $\mathbf{A}_{k-1}$  to  $\mathbf{A}_k$ . Then, if  $n$  is the order of  $\mathbf{A}_0$ ,  $i$  is given by

$$i - 1 \equiv k - 1 \pmod{n} \quad (6.580)$$

Thus the rows and columns are modified cyclically in natural order. The k-th step is as follows:

(a) Let  $R_k$  and  $C_k$  denote the  $\|\cdot\|_p$  norms of row i and column i of  $\mathbf{A}_{k-1}$ . According to the “no-null-row-or-column” assumption  $R_k C_k \neq 0$ . Hence, if  $\beta$  denotes the radix base, there is a unique (positive or negative) integer  $\sigma = \sigma_k$  such that

$$\beta^{2\sigma-1} < \frac{R_k}{C_k} \leq \beta^{2\sigma+1} \tag{6.581}$$

Define  $f = f_k$  by

$$f = \beta^\sigma \tag{6.582}$$

(b) For a constant  $\gamma \leq 1$  (taken as .95 in the authors’ program) take

$$\bar{\mathbf{D}}_k = \left\{ \begin{array}{ll} \mathbf{I} + (f-1)\mathbf{e}_i\mathbf{e}_i^T & \text{if } (C_k f)^p + (\frac{R_k}{f})^p < \gamma(C_k^p + R_k^p) \\ \mathbf{I} & \text{otherwise} \end{array} \right\} \tag{6.583}$$

where  $\mathbf{e}_i$  is the i-th column of the identity matrix  $\mathbf{I}$ .

(c) Form

$$\mathbf{D}_k = \bar{\mathbf{D}}_k \mathbf{D}_{k-1} \quad (\mathbf{D}_0 = \mathbf{I}) \tag{6.584}$$

$$\mathbf{A}_k = \bar{\mathbf{D}}_k^{-1} \mathbf{A}_{k-1} \bar{\mathbf{D}}_k \tag{6.585}$$

The authors claim that if  $\gamma = 1$  then in every step,  $f$  is that integer power of  $\beta$  which gives maximum reduction of the contribution of the i-th row and column to  $\|\mathbf{A}_k\|_p$ . If  $\gamma$  is slightly smaller than 1, a step is skipped if it would produce a very small reduction in  $\|\mathbf{A}_{k-1}\|_p$ . Iteration is terminated if, for a complete cycle ( $i = 1, \dots, n$ ),  $\bar{\mathbf{D}}_k = \mathbf{I}$ .

We will consider the example of a low-order matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 8 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix} . \text{ i.e. } \mathbf{A}_0 = \begin{bmatrix} 0 & 2 & 8 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix} , \|\mathbf{A}_0\|_E = \sqrt{76}$$

For  $i = k = 1$  we have  $R_1 = \sqrt{2^2 + 8^2} = \sqrt{68}$ ,  $C_1 = \sqrt{8}$ ,  $\frac{R_1}{C_1} = \sqrt{8.5} \approx 2.9$ , hence  $\sigma = 1$ ,  $f = 2^1 = 2$ .

$$(C_1 f)^2 + (\frac{R_1}{f})^2 = 32 + 17 = 49 < .95((C_1)^2 + (R_1)^2) = .95 \times 76$$

Hence we compute

$$\bar{\mathbf{D}}_1 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} ,$$

$$\mathbf{A}_1 = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 & 8 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 4 \\ 4 & 0 & 0 \\ 4 & 0 & 0 \end{bmatrix},$$

$$\|\mathbf{A}_1\|_E = \sqrt{49}$$

For  $i = k = 2$ , we have  $R_2 = 4, C_2 = 1, \frac{R_2}{C_2} = 4, \sigma = 1, f = 2, \overline{\mathbf{D}}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} 0 & 2 & 4 \\ 2 & 0 & 0 \\ 4 & 0 & 0 \end{bmatrix}$

For  $i = k = 3$ , we have  $R_3 = C_3 = 4, \frac{R_3}{C_3} = 1, \sigma = 0, f = 1, \overline{\mathbf{D}}_3 = \mathbf{I}, \mathbf{A}_3 = \mathbf{A}_2$

For  $k = 4, i = 1$  we have  $R_1 = \sqrt{20}, C_1 = \sqrt{20}, \frac{R_1}{C_1} = 1, \sigma = 0, f = 1, \overline{\mathbf{D}}_1 = \mathbf{I};$  as before  $\overline{\mathbf{D}}_2 = \overline{\mathbf{D}}_3 = \mathbf{I},$  i.e. no change for a complete cycle, so iteration is terminated.

If a row or column of  $\mathbf{A}_0$  is null then  $a_{ii}$  is an eigenvalue of  $\mathbf{A}$  and the calculation should proceed on the submatrix obtained by deleting row and column  $i$ . For more details see the cited paper.

The authors give an Algol program and report numerical experiments on three matrices of moderate order. The errors in the eigenvalues of the balanced matrices were reduced by a factor of at least  $10^4$  compared to the unbalanced case.

### 6.7 Miscellaneous Methods and Special Applications

Jonsson and Vavasis (2004) consider the case where the leading coefficient of the polynomial is much smaller than some of the other coefficients. This occurs for example in geometric applications where one often works with a fixed “toolbox” including cubic splines. An application might store a linear or quadratic polynomial as a cubic with leading coefficient of zero. Then transformations such as rotations might result in a leading coefficient which is small but no longer zero.

When we use eigenvalues of companion matrices to compute zeros, the translation from a polynomial to an eigenvalue problem should not cause the conditioning to become much worse. The authors concentrate on this issue. They consider a polynomial which is not necessarily monic, and its companion matrix in the form

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & 1 \\ -\frac{c_0}{c_n} & \dots & \dots & -\frac{c_{n-1}}{c_n} \end{bmatrix} \tag{6.586}$$

Suppose the eigenvalues of  $\mathbf{C}$  are computed using a backward stable algorithm, i.e. they are the exact eigenvalues of  $\mathbf{C} + \mathbf{E}$ , where  $\mathbf{E}$  has small entries. The computed

eigenvalues are also roots of a perturbed polynomial  $\tilde{p}$  with coefficients  $\tilde{c}_j = c_j + e_j$ . We recall that Edelman and Mirakami (1995) showed that

$$e_{j-1} = \sum_{m=0}^{j-1} c_m \sum_{i=j+1}^n E_{i+m-j,i} - \sum_{m=j}^n c_m \sum_{i=1}^j E_{i+m-j,i} \tag{6.587}$$

(The differences between this and 6.570 are due to the fact that  $\mathbf{C}$  above in 6.586 is the transpose of the form used by Edelman and Mirakami). Note that the leading coefficient  $c_n$  is not perturbed ( $e_n = 0$ ).

The Matlab routine **roots** solves the eigenvalue problem

$$\mathbf{C}\mathbf{x} = \lambda\mathbf{x} \tag{6.588}$$

by means of the QR-algorithm, and it is stated that

$$\|\mathbf{E}\| < k_1\|\mathbf{C}\|u \tag{6.589}$$

where  $u$  is the machine precision,  $k_1$  depends only on  $n$ , and for any matrix  $\mathbf{A}$ ,  $\|\mathbf{A}\| = \|\mathbf{A}\|_F$  is the Frobenius norm

$$= \left(\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2\right)^{\frac{1}{2}} \tag{6.590}$$

Let

$$\mathbf{c} = [c_0, \dots, c_n] \text{ and } \tilde{\mathbf{c}} = [\tilde{c}_0, \dots, \tilde{c}_n] \tag{6.591}$$

Now (very approximately)

$$\|\mathbf{C}\| \approx \left|\frac{c_{max}}{c_n}\right| \text{ where } |c_{max}| = \max_j |c_j| \tag{6.592}$$

So we get a backward error bound

$$\|\tilde{\mathbf{c}} - \mathbf{c}\| < k_2 \left|\frac{c_{max}}{c_n}\right| \|\mathbf{c}\|u + O(u^2) \tag{6.593}$$

where  $\|\mathbf{v}\| = \|\mathbf{v}\|_2$  for vectors  $\mathbf{v}$ . Henceforward we will omit the terms  $O(u^2)$ . The bound 6.593 is large when

$$|c_{max}| \gg |c_n|, \tag{6.594}$$

which is the case being considered in Jonsson and Vavasis' paper. Now consider the generalized eigenvalue problem

$$\mathbf{A} - \lambda\mathbf{B} = \begin{bmatrix} 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & 1 \\ -c_0 & \dots & \dots & -c_{n-1} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & 1 & 0 \\ \dots & \dots & 0 & c_n \end{bmatrix} = \mathbf{0} \tag{6.595}$$

If this is solved using the QZ-algorithm (see Golub and Van Loan (1996) Sec. 7.7) the computed eigenvalues are exact for a perturbed matrix pencil

$$(\mathbf{A} + \mathbf{E}) - \lambda(\mathbf{B} + \mathbf{F}) \quad (6.596)$$

with

$$\|\mathbf{E}\| < k_a \|\mathbf{A}\|u, \|\mathbf{F}\| < k_b \|\mathbf{B}\|u \quad (6.597)$$

where  $k_a, k_b$  depend only on  $n$ . Assume that the coefficients have been scaled so that  $\|\mathbf{c}\| = 1$ . The authors quote Van Dooren and Dewilde (1983) as showing that the computed roots are exact for a polynomial  $\tilde{p}$  with

$$\|\tilde{\mathbf{c}} - \mathbf{c}\| < k_3 \|\mathbf{c}\|u \quad (6.598)$$

where  $k_3$  depends only on  $n$ . For the type of polynomial being considered, this is a much better result than 6.593. The authors compare the accuracy of roots computed by 6.588 (solved by **roots**) and 6.595. It is hoped that the forward error is of order (condition number  $\times u$ ), or smaller. If we use 6.595, this is indeed the case. For polynomials with a small leading coefficient and roots of order 1 in magnitude or smaller, 6.595 does better than **roots**. For other cases, **roots** is sometimes better. In more detail, they generated 100 random test polynomials of degree 8 with coefficients of the form  $(\alpha + i\beta)10^\gamma$  with  $\alpha, \beta$  in the range  $[-1,1]$  and  $\gamma$  in the range  $[-10,10]$ . The leading coefficient was fixed at  $10^{-10}$ . The above was multiplied by  $(z - \frac{1}{2})^2$  in each case to give some ill-conditioning. The resulting degree 10 polynomials were solved by four methods (see below). Given computed roots

$$\hat{z}_1, \dots, \hat{z}_n, \text{ let } \hat{p}(z) = (z - \hat{z}_1)\dots(z - \hat{z}_n) \quad (6.599)$$

They compute the coefficients of  $\hat{p}$  using 40-decimal-digit precision. If we allow all the coefficients of  $p$  to be perturbed, the perturbation giving  $\hat{z}_1, \dots, \hat{z}_n$  as exact roots is not unique, since multiplying  $\hat{p}$  by a scalar does not change its roots. Unless otherwise stated, we assume that the backward error computed is minimal in a least squares sense, i.e. we find

$$\min_{\tau} \|\tau \hat{\mathbf{c}} - \mathbf{c}\| \quad (6.600)$$

This is obtained when

$$\tau = \frac{(\hat{\mathbf{c}}^H \mathbf{c})}{(\hat{\mathbf{c}}^H \hat{\mathbf{c}})} \quad (6.601)$$

The four methods used were:

- a) Matlab's **roots** with backward error computed by 6.600.
- b) Equation 6.595, with  $\|\mathbf{c}\| = 1$ , and error by 6.600.
- (c) Equation 6.595 but with  $c_n = 1$  instead of  $\|\mathbf{c}\| = 1$ , and error by 6.600.
- (d) Equation 6.595, with  $\|\mathbf{c}\| = 1$ , and backward error computed by  $\|\frac{c_n}{\hat{c}_n} \hat{\mathbf{c}} - \mathbf{c}\|$

(i.e.  $c_n$  not perturbed).

It was observed that (b) gives backward errors of order  $\|\mathbf{c}\|u$ , i.e. 6.598 holds. The other methods give much greater error, which shows that 6.588 is not good for these problems, and also that the normalization  $\|\mathbf{c}\| = 1$  and perturbing all coefficients including  $c_n$  are needed for 6.598 to hold.

Now we turn to forward error, i.e. the accuracy of the computed roots. Let  $z_1, \dots, z_n$  be the “exact” roots, found by Matlab in 40-decimal-digit arithmetic. The roots computed in double precision will be denoted by  $\hat{z}_1, \dots, \hat{z}_n$ . The absolute root error  $|z_j - \hat{z}_j|$  for each root of each polynomial is plotted as a function of its condition number, given by 6.549. It is found that for method (b), the error is nearly always well below (condition number  $\times u$ ), whereas for method (a) it is well above. Note that the results are only reported for roots  $|z| < 10$ , since in geometric computing we are usually only interested in the interval  $[-1,1]$ . The computation of larger roots is considered in section 4 of the cited paper, but will not be discussed here. The authors also considered dropping the leading term altogether, but concluded that this method is less accurate than 6.595. However if a step of Newton’s method is also performed, the forward errors are often improved, but the backward errors often get worse. The authors apply the above analysis to Bezier polynomials, which are widely used in geometric computing. For details see the cited paper. Again, the use of 6.595 followed by a fractional linear transformation is much more accurate than **roots**.

Good (1961) derives a matrix which serves the same purpose as the companion matrix when the polynomial is expressed as a series of Chebyshev polynomials. He calls it the **colleague** matrix. He makes use of

$$U_n(x) = (1 - x^2)^{-\frac{1}{2}} \sin\{(n + 1)\cos^{-1}x\} \quad (n = 0, 1, 2, \dots) \tag{6.602}$$

$$U_0 = 1$$

and

$$S_n(x) = U_n\left(\frac{x}{2}\right) \tag{6.603}$$

Then he supposes that the polynomial is given by an “S-series”, i.e.

$$p(x) = a_0 + a_1S_1(x) + a_2S_2(x) + \dots + a_nS_n(x) \tag{6.604}$$

We can express  $U_n(x)$  as a determinant

$$\begin{vmatrix} 2x & -1 & 0 & \dots & \dots & 0 \\ -1 & 2x & -1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 2x & -1 \\ 0 & \dots & \dots & 0 & -1 & 2x \end{vmatrix} \tag{6.605}$$



This can be re-expressed as:  $S_n(\lambda)$  is the characteristic polynomial  $|\lambda\mathbf{I} - \mathbf{K}|$  of the matrix

$$\mathbf{K} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{6.606}$$

Good then states that we may “easily” prove by induction that

$$U_n(x) + a_{n-1}U_{n-1} + \dots + a_0 = \begin{vmatrix} 2x & -1 & 0 & \dots & 0 & 0 \\ -1 & 2x & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2x & -1 \\ a_0 & a_1 & a_2 & \dots & -1 + a_{n-2} & 2x + a_{n-1} \end{vmatrix} \tag{6.607}$$

Again, this can be expressed as: the characteristic polynomial of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \dots & 1 - a_{n-2} & -a_{n-1} \end{bmatrix} \tag{6.608}$$

(which Good calls the colleague matrix) is

$$a(\lambda) = S_n(\lambda) + a_{n-1}S_{n-1}(\lambda) + \dots + a_0 \tag{6.609}$$

Now suppose we wish to approximate the zeros of a function in a finite interval, which can be normalized to  $[-1,1]$  The function may be approximated by means of an S-series such as 6.609. Then the roots of this can be found as the eigenvalues of  $\mathbf{A}$  (6.608).

Barnett (1975) treated the case, similar to the above, of polynomials expressed as a series of orthogonal polynomials. These are defined as  $\{p_i(x)\}$ , ( $i = 0, 1, 2, \dots$ ) where

$$p_0(x) = 1, \quad p_1(x) = \alpha_1x + \beta_1 \tag{6.610}$$

$$p_i(x) = (\alpha_i x + \beta_i)p_{i-1}(x) - \gamma_i p_{i-2}(x) \quad (i \geq 2) \tag{6.611}$$

where  $\alpha_i, \beta_i, \gamma_i$  are constants depending on  $i$ , and  $\alpha_i > 0, \gamma_i > 0$ . The  $p_i(x)$  can be assumed orthogonal. Any  $n$ -th degree polynomial can be expressed uniquely as:

$$a(x) = a_n p_n(x) + a_{n-1} p_{n-1}(x) + \dots + a_1 p_1(x) + a_0 \tag{6.612}$$

Assume that  $a_n = 1$ , and define the monic polynomial

$$\tilde{a}(x) = a(x)/(\alpha_1\alpha_2\dots\alpha_n) \tag{6.613}$$

$$= x^n + \tilde{a}_{n-1}x^{n-1} + \dots + \tilde{a}_1x + \tilde{a}_0 \tag{6.614}$$

**Theorem** The matrix

$$\mathbf{A} = \begin{bmatrix} -\frac{\beta_1}{\alpha_1} & \frac{\gamma_2^{\frac{1}{2}}}{\alpha_1} & 0 & \dots & 0 & 0 \\ \frac{\gamma_2^{\frac{1}{2}}}{\alpha_2} & -\frac{\beta_2}{\alpha_2} & \frac{\gamma_3^{\frac{1}{2}}}{\alpha_2} & \dots & 0 & 0 \\ 0 & \frac{\gamma_3^{\frac{1}{2}}}{\alpha_3} & -\frac{\beta_3}{\alpha_3} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\frac{\beta_{n-1}}{\alpha_{n-1}} & \frac{\gamma_n^{\frac{1}{2}}}{\alpha_{n-1}} \\ -\frac{a_0}{\alpha_n(\gamma_2\gamma_3\dots\gamma_n)^{\frac{1}{2}}} & -\frac{a_1}{\alpha_n(\gamma_3\dots\gamma_n)^{\frac{1}{2}}} & \dots & \dots & \frac{-a_{n-2}+\gamma_n}{\alpha_n\gamma_n^{\frac{1}{2}}} & \frac{-a_{n-1}-\beta_n}{\alpha_n} \end{bmatrix} \tag{6.615}$$

has  $\tilde{a}(x)$  as its characteristic polynomial, as we will show below. This generalizes Good’s result which follows if in 6.615 we take  $\alpha_i = \gamma_i = 1$ ,  $\beta_i = 0$ . Barnett suggests the term “comrade” matrix for 6.615. We will now prove the theorem above. For we have

$$p_i(x) = \begin{vmatrix} \alpha_1x + \beta_1 & -\gamma_2^{\frac{1}{2}} & 0 & \dots & 0 & 0 \\ -\gamma_2^{\frac{1}{2}} & \alpha_2x + \beta_2 & -\gamma_3^{\frac{1}{2}} & \dots & 0 & 0 \\ 0 & -\gamma_3^{\frac{1}{2}} & \alpha_3x + \beta_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_{i-1}x + \beta_{i-1} & -\gamma_i^{\frac{1}{2}} \\ \dots & \dots & \dots & \dots & -\gamma_i^{\frac{1}{2}} & \alpha_ix + \beta_i \end{vmatrix} \tag{6.616}$$

(as we may see by expanding the determinant by its last column and comparing with 6.611). Now define the diagonal matrix

$$\mathbf{D} = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n) \tag{6.617}$$

and consider

$$\det(x\mathbf{D} - \mathbf{DA}) = \det\mathbf{D}\det(x\mathbf{I} - \mathbf{A}) \tag{6.618}$$

$$= (\alpha_1\alpha_2\dots\alpha_n)\det(x\mathbf{I} - \mathbf{A}) \tag{6.619}$$

Expanding the determinant on the left side of 6.618 by the last row and using 6.616 gives

$$\begin{aligned}
 & a_0 + a_1 p_1(x) + a_2 p_2(x) + \dots + (a_{n-2} - \gamma_n) p_{n-2}(x) + \\
 & (a_{n-1} + \alpha_n x + \beta_n) p_{n-1}(x)
 \end{aligned} \tag{6.620}$$

Using 6.611 with  $i = n$  converts the above to  $a(x)$  in 6.612 (remembering that  $a_n = 1$ ); finally 6.619 and 6.613 give the desired result. We can show that

$$\mathbf{v}_i = \begin{bmatrix} 1 \\ \frac{p_1(\lambda_i)}{\gamma_2^{\frac{1}{2}}} \\ \frac{p_2(\lambda_i)}{(\gamma_2 \gamma_3)^{\frac{1}{2}}} \\ \dots \\ \dots \\ \frac{p_{n-1}(\lambda_i)}{(\gamma_2 \gamma_3 \dots \gamma_n)^{\frac{1}{2}}} \end{bmatrix} \tag{6.621}$$

is an eigenvector of  $\mathbf{A}$  corresponding to the eigenvalue  $\lambda_i$ . Barnett also proves that the comrade matrix

$$\mathbf{A} = \mathbf{TCT}^{-1} \tag{6.622}$$

where  $\mathbf{C}$  is the usual companion matrix i.e.

$$\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ -\tilde{a}_0 & -\tilde{a}_1 & -\tilde{a}_2 & \dots & -\tilde{a}_{n-1} \end{bmatrix} \tag{6.623}$$

and

$$\mathbf{T} = \mathbf{ES} \tag{6.624}$$

Here

$$\mathbf{E} = \text{diag}(1, \gamma_2^{-\frac{1}{2}}, (\gamma_2 \gamma_3)^{-\frac{1}{2}}, \dots, (\gamma_2 \dots \gamma_n)^{-\frac{1}{2}}) \tag{6.625}$$

and

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ p_{10} & p_{11} & 0 & \dots & 0 \\ p_{20} & p_{21} & p_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ p_{n-2,0} & p_{n-2,1} & p_{n-2,2} & \dots & 0 \\ p_{n-1,0} & p_{n-1,1} & p_{n-1,2} & \dots & p_{n-1,n-1} \end{bmatrix} \tag{6.626}$$

where

$$p_i(x) = \sum_{j=1}^i p_{ij} x^j \quad (i = 1, \dots, n-1) \quad (6.627)$$

For a proof see the cited paper.

In a later paper Barnett (1981) shows how the companion matrix  $\mathbf{C}$  can be used to find the GCD of two polynomials  $a(\lambda)$  (whose companion matrix is  $\mathbf{C}$ ) and

$$b(\lambda) = b_m \lambda^m + b_{m-1} \lambda^{m-1} + \dots + b_0 \quad (6.628)$$

with  $m < n$ . Let

$$b(\mathbf{C}) = b_m \mathbf{C}^m + b_{m-1} \mathbf{C}^{m-1} + \dots + b_1 \mathbf{C} + b_0 \mathbf{I} \quad (6.629)$$

Then it is claimed that the rows  $\mathbf{r}_1, \dots, \mathbf{r}_n$  of  $b(\mathbf{C})$  satisfy

$$\mathbf{r}_i = \mathbf{r}_{i-1} \mathbf{C} \quad (6.630)$$

and

$$\mathbf{r}_1 = [b_0, b_1, \dots, b_m, 0, \dots, 0] \quad (6.631)$$

Denote the columns of  $b(\mathbf{C})$  by  $\mathbf{c}_1, \dots, \mathbf{c}_n$ , and let the GCD of  $a(\lambda)$  and  $b(\lambda)$  be

$$d(\lambda) = \lambda^k + d_{k-1} \lambda^{k-1} + \dots + d_0 \quad (6.632)$$

Then Barnett quotes another paper of his (Barnett (1970)) as proving the following:

- (i)  $\det(b(\mathbf{C})) \neq 0$  iff  $d(\lambda) = 1$ .
- (ii)  $k = n - \text{rank}[b(\mathbf{C})]$ .
- (iii)  $\mathbf{c}_{k+1}, \dots, \mathbf{c}_n$  are linearly independent, and

$$\mathbf{c}_i = d_{i-1} \mathbf{c}_{k+1} + \sum_{j=k+2}^n x_{ij} \mathbf{c}_j \quad (6.633)$$

for some  $x_{ij}$ .

Ammar et al (2001) describe a zero-finding method based on Szego polynomials. It utilizes the fact that, after a change of variables, any polynomial can be considered as a member of a family of Szego polynomials. The zero-finder uses the recursion relations for these polynomials, defined as follows:

$$\phi_0(z) = \phi_0^*(z) = 1 \quad (6.634)$$

$$\sigma_{j+1} \phi_{j+1}(z) = z \phi_j(z) + \gamma_{j+1} \phi_j^*(z) \quad (6.635)$$

$$\sigma_{j+1} \phi_{j+1}^*(z) = \bar{\gamma}_{j+1} z \phi_j(z) + \phi_j^*(z) \quad (6.636)$$

where the  $\gamma_{j+1}$ ,  $\sigma_{j+1}$ , and  $\delta_{j+1}$  are given by

$$\gamma_{j+1} = -\frac{(z\phi_j, 1)}{\delta_j} \quad (6.637)$$

$$\sigma_{j+1} = \sigma_j(1 - |\gamma_{j+1}|^2) \quad (6.638)$$

$$\delta_{j-1} = \delta_j\sigma_{j+1}, \delta_0 = \sigma_0 = 1 \quad (6.639)$$

(the above all for  $j = 0, 1, 2, \dots$ ). In 6.637 the inner product

$$(f, g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(z)\overline{g(z)}d\omega(t) \quad (6.640)$$

where  $z = e^{it}$ . Moreover

$$\phi_j^*(z) = z^j \overline{\phi_j\left(\frac{1}{z}\right)} \quad (6.641)$$

The validity of these recurrence relations is partially proved in Ammar and Gragg (1987). The zeros of the Szego polynomials are strictly inside the unit circle and all  $\gamma_j$  have magnitude  $< 1$ . The leading coefficient of  $\phi_j$  is  $\frac{1}{\delta_j}$ .

Given a polynomial  $p_n(z)$  in the usual power-of-x form, we first transform  $p_n$  so that the average of its zeros vanishes. Then we determine a disk centered at the origin that contains all zeros of the transformed polynomial, and scale so that this becomes the unit disk. Thus the problem of finding the zeros of  $p_n(z)$  has been transformed into the problem of finding the zeros of a monic polynomial with all its zeros in the unit disk. We identify this with the monic Szego polynomial  $\Phi_n = \delta_n\phi_n$ . More details follow.

Let  $\{\zeta_j\}_{j=1}^n$  denote the zeros of  $p_n(z)$  and define their average:

$$\rho = \frac{1}{n} \sum_{i=1}^n \zeta_i \quad (6.642)$$

We compute this from

$$\rho = -\frac{c_{n-1}}{n} \quad (6.643)$$

and define

$$\hat{z} = z - \rho \quad (6.644)$$

Then

$$\hat{p}_n(\hat{z}) = p_n(z) = \hat{z}^n + \hat{c}_{n-2}\hat{z}^{n-2} + \dots + \hat{c}_1\hat{z} + \hat{c}_0 \quad (6.645)$$

The  $\hat{c}_j$  can be computed in  $O(n^2)$  operations. Now we use a theorem of Ostrowski (1969) which states that if a polynomial such as  $\hat{p}_n(\hat{z})$  in 6.645 has all  $|\hat{c}_i| \leq 1$ , then all its zeros lie in the disk  $\{z : |z| < \frac{1}{2}(1 + \sqrt{5})\}$  (for proof see the Ammar et al paper). After we make a change of variable  $\tilde{z} = \sigma\hat{z}$ , where  $\sigma > 0$  is chosen so that

$$\max_{2 \leq j \leq n} \sigma^j |\hat{c}_{n-j}| = 1 \tag{6.646}$$

the transformed polynomial  $\tilde{p}_n(\tilde{z}) = \sigma^n \hat{p}_n(\hat{z})$  satisfies the conditions of Ostrowski's theorem above. Finally we make another change of variables

$$\zeta = \tau\tilde{z} \tag{6.647}$$

where

$$\tau = \frac{2}{1 + \sqrt{5}} \tag{6.648}$$

to yield a monic polynomial

$$\Phi_n^{(\tau)}(\zeta) = \tau^n \tilde{p}_n(\tilde{z}) \tag{6.649}$$

with all zeros inside the unit circle. We identify  $\Phi_n^{(\tau)}$  with the monic Szego polynomial  $\delta_n \phi_n$  and wish to compute the recursion coefficients  $\{\gamma_j\}_{j=1}^n$  that determine polynomials of lower degree  $\{\phi_j\}_{j=0}^{n-1}$  in the same family of Szego polynomials. Given the coefficients of  $\phi_n$  we may compute those of  $\phi_n^*$  and apply 6.634-6.639 backwards to obtain  $\gamma_n$  and the  $\gamma_j$  and  $\phi_j$  for  $j=n-1, \dots, 1$ . Assuming that the  $\gamma_j$  and  $\sigma_j$  are available from the above, and eliminating  $\phi_j^*$  from 6.634-6.636, we may obtain an expression for  $\phi_{j+1}$  in terms of  $\phi_j, \phi_{j-1}, \dots, \phi_0$ . The Schur-Cohn algorithm (see Henrici (1974) Chapter 6) is an efficient way of doing this. Writing these expressions in matrix form yields

$$\begin{aligned} & [\phi_0(z), \phi_1(z), \dots, \phi_{n-1}(z)] H_n = \\ & z[\phi_0(z), \phi_1(z), \dots, \phi_{n-1}(z)] - [0, 0, \dots, \phi_n(z)] \end{aligned} \tag{6.650}$$

where

$$H_n = \begin{bmatrix} -\gamma_1 & -\sigma_1\gamma_2 & -\sigma_1\sigma_2\gamma_3 & \dots & \dots & -\sigma_1 \dots \sigma_{n-1} \gamma_n \\ \sigma_1 & -\bar{\gamma}_1\gamma_2 & -\bar{\gamma}_1\sigma_2\gamma_3 & \dots & \dots & -\bar{\gamma}_1\sigma_2 \dots \sigma_{n-1} \gamma_n \\ 0 & \sigma_2 & -\bar{\gamma}_2\gamma_3 & \dots & \dots & -\bar{\gamma}_2\sigma_3 \dots \sigma_{n-1} \gamma_n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \sigma_{n-2} & -\bar{\gamma}_{n-2}\gamma_{n-1} & -\bar{\gamma}_{n-2}\sigma_{n-1} \gamma_n \\ 0 & \dots & \dots & 0 & \sigma_{n-1} & -\bar{\gamma}_{n-1} \gamma_n \end{bmatrix} \tag{6.651}$$

this is called the Szego-Hessenberg matrix associated with the set  $\{\phi_j\}$ . 6.650 shows that the zeros of  $\phi_n(z)$  are the eigenvalues of  $H_n$ ; we can use this feature to find

the required zeros of  $p_n(z)$ . It is found that the zeros are calculated more accurately when  $\max_j |\zeta_j|$  is close to one. The authors describe an elaborate method of rescaling to achieve this situation, using the Schur-Cohn algorithm. When this has been done, we may re-calculate the  $\gamma_j$  etc, form the matrix  $H_n$ , balance it, and compute its eigenvalues by the QR algorithm.

The authors also describe briefly a continuation method, based on a paper by Ammar et al (1996). This method is often more accurate than that using  $H_n$  above, but on the other hand it quite often breaks down.

Numerous numerical test are reported comparing four methods, detailed below:

- 1) **CB**: the QR algorithm applied to the companion matrix of  $p_n(z)$  after balancing.
- 2) **CBS**: the QR algorithm applied to the companion matrix of the monic Szego polynomial  $\Phi_n$  after balancing.
- 3) **SHB**: the QR algorithm applied to  $H_n$  after balancing.
- 4) **CM**: the continuation method mentioned above.

In the great majority of cases SHB gave the lowest error (strictly speaking CM was often more accurate than SHB, but we suggest rejecting this method because of its large number of failures. On the other hand the authors suggest using CM but switching to SHB whenever CM fails).

## 6.8 Programs and Packages

In Section 2 of this Chapter we have described a method due to Fortune (2002). His algorithm has been implemented in C++, using EISPACK (see Smith (1976)) or LAPACK (see Anderson et al (1995)) for the eigenvalue computations, with GMP (see Granlund (1996)) for multiple-precision arithmetic. It accepts polynomials of arbitrary degree, with coefficients (real or complex) specified either as arbitrary precision integers or rationals. It is available from <http://cm.bell-labs.com/who/sjf/eigenolve.html>.

Then in Section 3 we described a method due to Bini, Gemignani, and Pan (2004a). Their algorithm has been implemented in Fortran 90 in the file **ips.tgz** which can be downloaded from [www.dm.unipi.it/~bini/software](http://www.dm.unipi.it/~bini/software)

Finally Zeng (2004a) devotes a whole paper to a Matlab implementation of his method as Algorithm 835: MULTROOT, which is available from ACM TOMS via Netlib. Using the Matlab representation for  $p$ , we can execute MULTROOT by

```
>>z = multroot(p);
```

For example, to find the roots of

$$p(x) = x^{10} - 17x^9 + 127x^8 - 549x^7 + 1521x^6 - 2823x^5 + 3557x^4 - 3007x^3 + 1634x^2 -$$

$516x + 72$

we need only two matlab commands:

```
>>p = [1 -17 127 -549 1521 -2823 3557 -3007 1634 -516 72];
>>z = multroot(p);
```

The following output appears on the screen:

THE CONDITION NUMBER	20.1463
THE BACKWARD ERROR	3.22e-016
THE ESTIMATED FORWARD ERROR	1.30e-014

<i>computed roots</i>	<i>multiplicities</i>
2.999999999999997	2
2.000000000000001	3
1.000000000000000	5

In addition there is output of a two-column matrix

$$\mathbf{z} = \begin{bmatrix} 2.999999999999997 & 2 \\ 2.000000000000001 & 3 \\ 1.000000000000000 & 5 \end{bmatrix}$$

The result shows an accurate factorization of

$$p(x) = (x - 1)^5(x - 2)^3(x - 3)^2$$

The full call sequence of MULTROOT is

```
>> [z,f_err,b_err,cond] = multroot(p,tol,thresh,growf)
```

where (besides  $\mathbf{z}$  and  $\mathbf{p}$  as previously described) other input/output items are optional as follows:

(1) INPUT:

tol: the backward error tolerance (default  $10^{-10}$ ).

thresh: the zero singular value threshold (default  $10^{-8}$ ).

growf: growth factor for the residual (default not mentioned but probably 100).

Most users will probably accept the default values, in which case they need not specify those options.

(ii) OUTPUT:

cond: the structure preserving condition number.

b\_err: the backward error.

f\_err: the estimated forward error based on the error estimate

$$\|\mathbf{z} - \hat{\mathbf{z}}\|_2 \leq 2\kappa_{l,W}(\hat{\mathbf{z}})\|\mathbf{p} - \hat{\mathbf{p}}\|_W$$

(see Section 4 of this Chapter).

The optional input/output items can be supplied/requested partially. For example

```
>> [z,f_err] = multroot(p,1.0e-8)
```

sets the backward error tolerance to  $10^{-8}$  and accepts as output  $\mathbf{z}$  and  $\mathbf{f\_err}$ .



If MULTROOT cannot find a non-trivial multiplicity structure (i.e. having one or more multiplicities  $> 1$ ) within the residual tolerance, the Matlab standard root-finder will automatically be called to calculate simple roots. We can force the continued use of MULTROOT in this case by calling the module MROOT in the first place. This returns  $\text{job} = 1$  if multiple roots are found, or  $\text{job} = 0$  otherwise.

Zeng states that the code may fail for several (unlikely) causes, such as:

- (1) High structure-preserving condition number.
- (2) The polynomial is near several polynomials with different multiplicity structures, e.g. the Wilkinson polynomial.
- (3) Multiplicity too high. Zeng does not state what is too high; one assumes it depends on the other roots as well.
- (4) Coefficients perturbed too much.

In an Appendix to the (2004a) paper Zeng lists a large number of polynomials which were used to test the package, all with great success. Each one has a label such as “jt01a”. The user may run a test case with a command such as

```
>> [p,z] = jt01a;
```

when roots and multiplicities will be displayed. The above may be followed by:

```
>> cond = spcond(z);
```

and then the structure-preserving condition number will be displayed.

## 6.9 References for Chapter 6

Ammar, G.S., Calvetti, D., and Reichel, L. (1996), Continuation methods for the computation of zeros of Szego polynomials, *Lin. Alg. Appls.* **249**, 125-155

——— and Gragg, W.B. (1987), The Generalized Schur Algorithm for the Superfast Solution of Toeplitz Systems, in *Rational Approximation and its Applications in Mathematics and Physics*, eds. J. Gilewicz et al, L.N.M 1237, Springer-Verlag, Berlin, 315-330; also updated at <http://www.math.niu.edu/~ammar/papers/gsa.pdf>

——— et al (2001), Polynomial zero-finders based on Szego polynomials, *J. Comput. Appl. Math.* **127**, 1-16

Anderson, E. et al (1995), *LAPACK User's Guide 2/E*, SIAM, Philadelphia, PA

Barnett, S. (1970), Greatest common divisor of two polynomials, *Lin. Alg. Appls.* **3**, 7-9

——— (1975), A Companion Matrix Analogue for Orthogonal Polynomials, *Lin.*

*Alg. Appls.* **12**, 197-208

——— (1981), Congenial Matrices, *Lin. Alg. Appls.* **41**, 277-298

Bini, D.A., Daddi, F. and Gemignani, L. (2004), On the shifted QR iteration applied to companion matrices, *Electr. Trans. Numer. Anal.* **18**, 137-152

——— and Fiorentino, G. (1999), Numerical computation of polynomial roots using MPSolve. Manuscript, Software at <ftp://ftp.dm.unipi.it/pub/mpsolve>

———, Gemignani, L., and Pan, V.Y. (2004a), Inverse Power and Durand- Kerner Iterations for Univariate Polynomial Root-Finding, *Comput. Math. Applic.* **47**, 447-459

———, ———, and ——— (2004b), Improved initialization of the accelerated and robust QR-like polynomial root-finding, *Electr. Trans. Numer. Anal.* **17**, 195-205

———, ———, and ——— (2005), Fast and stable QR eigenvalue algorithms for generalized companion matrices and secular equations, *Numer. Math* **100**, 373-408

Brand, L. (1964), The Companion Matrix and its Properties, *Amer. Math. Monthly* **71**, 629-634

Brugnano, L. (1995), Numerical Implementation of a New Algorithm for Polynomials with Multiple Roots, *J. Difference Equ. Appl.* **1**, 187-207

——— and Trigiantè, D. (1995), Polynomial Roots: The Ultimate Answer, *Lin. Alg. Appl.* **225**, 207-219

Carstensen, C. (1991), Linear construction of companion matrices, *Lin. Alg. Appl.* **149**, 191-214

Dennis, J.E. and Schnabel, R.B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, N.J.

Edelman, A. and Murakami, H. (1995), Polynomial Roots from Companion Matrix Eigenvalues, *Math. Comp.* **64**, 763-776

Elsner, L. (1973), A Remark on Simultaneous Inclusions of the Zeros of a Polynomial by Gershgorin's Theorem, *Numer. Math.* **21**, 425-427

Farmer, M.R. and Loizou, G. (1977), An algorithm for the total, or partial, factorization of a polynomial, *Math. Proc. Camb. Phil. Soc* **82**, 427-437

Fiedler, M. (1990), Expressing a Polynomial as the Characteristic Polynomial of a Symmetric Matrix, *Lin. Alg. Appl.* **141**, 265-270

Fortune, S. (2002), An Iterated Eigenvalue Algorithm for Approximating Roots of Univariate Polynomials, *J. Symb. Comput.* **33**, 627-646

Francis, J.G.F. (1961), The *QR* Transformation, a unitary analogue to the *LR* Transformation, I and II, *Computer J* **4**, 265-271 and 332-345

Gemignani, L. (1998), Computing a factor of a polynomial by means of multishift LR algorithms, *SIAM J. Matrix Anal. Appl.* **19**, 161-181

Gill, P.E. et al (1974), Methods for Modifying Matrix Factorizations, *Math. Comp.* **28**, 505-535

Goedecker, S. (1994), Remark on Algorithms to Find Roots of Polynomials, *SIAM J. Sci. Comput.* **15**, 1059-1063

Golub, G.H. and Van Loan, C.F. (1996), *Matrix Computations 3/E*, The Johns Hopkins University Press, Baltimore, MD.

Good, I.J. (1961), The Colleague Matrix, a Chebyshev Analogue of the Companion Matrix, *Quart. J. Math. Oxford* **12**, 61-68

Granlund, T. (1996), GNU MP: the GNU multiple precision arithmetic library, Ed. 2.0. Code available from [www.swox.com/gmp/](http://www.swox.com/gmp/).

Hammer, R. et al (1995), Chapter 9: Zeros of Complex Polynomials, in *C++ Toolbox for Verified Computing*, Springer-Verlag, Berlin, 164-185

Hardy, G.H. et al (1934), *Inequalities*, Cambridge University Press

Hazewinkel, M. (ed.) (1989), *Encyclopaedia of Mathematics* **4**, Kluwer Academic, Dordrecht, 438-439

Henrici, P. (1974), *Applied and Computational Complex Analysis*, Wiley, New York, 241-243

Jenkins, M.A. and Traub, J.F. (1970), A Three-Stage Variable-Shift Iteration for Polynomial Zeros and its Relation to Generalized Rayleigh Iteration, *Numer. Math.* **14**, 252-263

Jonsson, G.F. and Vavasis, S. (2004), Solving polynomials with small leading coef-

ficients, *SIAM J. Matrix Anal. Appls.* **26**, 400-414

Kahan, W. (1972), Conserving confluence curbs ill-condition, *Tech. Rep. 6*, Computer Science, University of California, Berkeley

Kravanja, P. et al (2000), ZEAL: A mathematical software package for computing zeros of analytic functions, *Comput. Phys. Comm.* **124**, 212-232

———, Sakurai, T., and Van Barel, M. (1999), On locating clusters of zeros of analytic functions, *BIT* **39**, 646-682

Krishnamurthy, E.V. (1960), Solving an algebraic equation by determining high powers of an associated matrix using the Cayley-Hamilton theorem, *Quart. J. Mech. Appl. Math.* **13**, 508-512

Laszlo, L. (1981), Matrix Methods for Polynomials, *Lin. Alg. Appl.* **36**, 129-131

Madsen. K. and Reid, J. (1975), Fortran subroutines for finding polynomial zeros, *Report HL. 75/1172 (c.13)*, Comp.Sci. and Systems Divn., A.E.R.E. Harwell, Oxford

Malek, F. and Vaillancourt, R. (1995A), Polynomial Zerofinding Iterative Matrix Algorithms, *Computers Math. Appl.* **29(1)**, 1-13

——— and ——— (1995B), A Composite Polynomial Zerofinding Matrix Algorithm, *Computers Math. Appl.* **30(2)**, 37-47

——— and ——— (1995C), On the conditioning of a composite polynomial zerofinding matrix algorithm, *C.R. Math. Rep. Acad. Sci. Canada* **17(2)**, 67-72

The Mathworks, Inc. (1992), *MATLAB User's Guide*

Mendelsohn, M.S. (1957), The Computation of Complex Proper Values and Vectors of a Real Matrix with Applications to Polynomials, *Math. Comp.* **11**, 91-94

Moler, C.B. (1991), "ROOTS-of polynomials, that is", *MathWorks Newsletter* **5**, 8-9

Niu, X.-M. and Sakurai, T. (2003), A Method for Finding the Zeros of Polynomials Using a Companion Matrix, *Japan J. Indust. Appl. Math.* **20**, 239-256

Osborne, E.E. (1960), On Pre-Conditioning of Matrices, *J. Assoc. Comput. Mach.* **7**, 338-345

Ostrowski, A.M. (1969), A Method for Automatic Solution of Algebraic Equations, in *Constructive Aspects of the Fundamental Theorem of Algebra*, eds. B. Dejon and P. Henrici, Wiley Interscience, London, 209-224

Parlett, B.N. and Reinsch, C. (1969), Balancing a Matrix for Calculation of Eigenvalues and Eigenvectors, *Numer. Math.* **13**, 293-304

Phillips, J.L. (1971), The triangular decomposition of Hankel matrices, *Math. Comp.* **25**, 599-602

Piessens, R. et al (1983), *QUADPACK: A Subroutine Package for Automatic Integration*, Springer, Berlin

Ralston, A. and Rabinowitz, P. (1978), *A First Course in Numerical Analysis*, McGraw-Hill, New York

Rupprecht, D. (1999), An algorithm for computing certified approximate GCD of  $n$  univariate polynomials, *J. Pure Appl. Alg.* **139**, 255-284

Schmeisser, G. (1993), A Real Symmetric Tridiagonal Matrix with a Given Characteristic Polynomial, *Lin. Alg. Appl.* **193**, 11-18

Smith, B.T. (1970), Error Bounds for Zeros of a Polynomial Based Upon Gershgorin's Theorems, *J. Assoc. Comput. Mach.* **17**, 661-674

——— et al (1976), *Matrix Eigensystem Routines-EISPACK GUIDE 2/E* (L.N.C.S Vol. 6), Springer, New York

Stewart., G.W. (1970), Short Notes: On the Convergence of Sebastiao E Silva's Method for Finding a Zero of a Polynomial, *SIAM Rev.* **12**, 458-460

Toh, K.-C. and Trefethen, L.N. (1994), Pseudozeros of polynomials and pseudospectra of companion matrices, *Numer. Math.* **68**, 403-425

Trench, W.F. (1965), An algorithm for the inversion of finite Hankel matrices, *J. Soc. Indust. Appl. Math.* **13**, 1102-1107

Uhlig, F. (1997), The DQR algorithm, basic theory, convergence, and conditional stability, *Numer. Math.* **76**, 515-553

——— (1999), General Polynomial Roots and their Multiplicities in  $O(n)$  Memory and  $O(n^2)$  Time, *Lin. Mult. Alg.* **46**, 327-359

- Van Barel, M. et al (2005), Orthogonal Rational Functions and Structured Matrices, *SIAM J. Matrix Anal. Appl.* **26**, 810-823
- Van Dooren, P. and Dewilde, P. (1983), The eigenstructure of an arbitrary polynomial matrix: computational aspects, *Lin. Alg. Appls.* **50**, 545-579
- Van Huffel, S. (1991), Iterative algorithms for computing the singular subspace of a matrix associated with its smallest singular values, *Lin. Alg. Appl.* **154-156**, 675-709
- Watkins, D.S. and Elsner, L. (1991), Convergence of algorithms of decomposition type for the eigenproblem, *Lin. Alg. Appls.* **143**, 19-47
- Wilkinson, J.H. (1963), *Rounding Errors in Algebraic Processes*, Her Majesty's Stationary Office, London
- Zeng, Z. (2003), A Method Computing Multiple Roots of Inexact Polynomials, *IS-SAC '03*, Philadelphia, PA, 266-272
- (2004a), Algorithm 835: MultRoot—A Matlab Package for Computing Polynomial Roots and Multiplicities, *ACM Trans. Math. Software* **30**, 218-236
- (2004b), Computing Multiple Roots of Inexact Polynomials, *Math. Comp.* **74**, 869-903

This page intentionally left blank

# Index

- Abel, xiii
- Aberth's method, 91, 95, 121
  - recursive interval version, 91
  - recursive version, 91
- accuracy, 131, 230, 232
  - level, 43, 44, 46, 47
  - required, xvi, 132
- algebraic methods
  - used by Newton, 135
- analytic function, 261
- annulus, 185
- approximate zero, 137, 281, 285
- approximation, xvii, 40, 83, 84
- argument, 184
  - principle, 50
- arithmetic
  - complex, 108
  - exact, 18
  - integer, 51
  - rational, 43
- asymptotic error constant, 132, 133
- asynchronous algorithm
  - on parallel processor, 118, 119
- attainable accuracy barrier, 268, 270
  
- Babylonians, xiii
- backward stable algorithm, 304
- Bairstow's method, 184, 187, 189
  - interval version, 169
- balancing
  - of matrix, 221, 294, 297, 298, 302
- banded system, 274
- base
  - of number system, 4, 160, 303
- Bernoulli's method, xiii
- Bernstein form, 186
- Bezier polynomials, 307
- bibliography, xiii
  - on roots of polynomials, xv, 26
- binary splitting, 10
- binomial coefficient, 187
- binomial theorem, 135, 136
- bisection, xvi, 26, 37, 39, 51, 131, 137, 174, 178, 184, 185
  - hybrid with Newton method, 187
  - two-dimensional, 50
- block diagonal matrix, 217, 232
- Borsch-Supan method, 115
  - disk version of, 89
  - generalization of, 91
- Borsch-Supan-Ehrlich-Aberth method
  - see Ehrlich-Aberth method, 70
- bound, 26
  - accurate, 26
  - actual, 26
  - Deutsch's, 27
  - estimated, 26
  - Kalantari's, 27
  - list of formulas, 28
  - on errors, 84, 116
  - on number of real roots, 53
  - on roots, 26, 111, 286
  - rational, 55
  - verified, 211
- bracketing method, 149, 168, 188
- breakdown, 215, 220, 221, 242



- complete, 220, 222
  - partial, 220, 221
  - step, 220
- Budan's theorem, 54
- calculus, 135, 136
- Cardano, xiii
- Carstensen's method, 84
- cartesian coordinates, 141
- Cauchy index, 38
- Cauchy's rule, 62
- Cauchy-Riemann equation, 262
- Cayley-Hamilton theorem, 210
- center, 172
  - of disk, 85
- centroid, 194
- characteristic polynomial, xviii, 175, 180,
  - 207, 208, 210, 214, 220, 223,
  - 231, 308, 309
- Chebyshev polynomial, 180, 187, 257, 307
- circle
  - equation of, 86
- cluster, 43, 47, 83, 94, 121, 156, 158, 258,
  - 270, 284, 285, 293
  - detection by Yakoubsohn's method,
  - 156
  - k-cluster, 43, 46
  - mean of, 278
- coefficient
  - leading, 48
- coefficients, xiii, 45, 47, 49
  - integral, 47, 49
- cofactor, 208
- combined method, 117
  - point-interval, 116
- common factor, 51
- common factors (of  $f$  and  $f'$ ), 40
- communication time
  - on parallel processor, 118
- companion matrix, 218, 278, 280, 284,
  - 289, 294, 296, 304, 311
  - balanced, 299
  - classical, 207, 218
  - Frobenius, 207, 225, 280, 283, 300
  - generalized, 242, 245, 246, 248
  - generalized (Fortune's version), 225
  - generalized( tridiagonal), 231
  - modified, 235
- complex
  - argument, 3
  - coefficients, 3
  - conjugate, 169
  - disk arithmetic, 84
- complexity
  - of Vincent's method, 60
- computed value, 4
- computer-aided geometric design, 186
- computing
  - parallel, 67
- comrade matrix
  - Chebyshev, 257
- condition number, 222, 224, 263, 267,
  - 268, 270, 274, 296, 297, 299,
  - 307
  - structure-preserving, 266
- conditioning, 295
- conjugate pairs, xiii
- continuation method, 314
- continued fraction, 56, 62, 63
  - infinite, 63, 64
  - roots by, 53
- control theory, xiv
- convergence, 74, 169
  - 4th order, 93
  - acceleration (Aitken's method), 154
  - acceleration (Levin's method), 154
  - acceleration of, 226
  - check for, 293
  - cubic, 95, 101
  - cubic (for Aberth's method), 71
  - cubic (for Nourein's method), 71
  - fourth order (of Nourein's improved method), 71
  - global, 180, 193
  - global (interval methods for), 161
  - global for real roots, 150
  - guaranteed, 95, 96, 100, 120, 149,
  - 187

- guaranteed (conditions for), 72
- guaranteed (order 4), 96
- not guaranteed, 131
- of order 4, 97, 98
- of order 5, 96, 97
- of order 6, 96, 97
- of order 7, 97
- of order  $k+2$ , 95
- of WDK method, 68
- order of, 23–25, 91, 116
- Ostrowski's condition, 165
- Ostrowski's condition (interval version), 166
- quadratic, 101, 172
- quadratic for WDK method, 68
- rapid, 165
- to the mean is quadratic, 81, 82
- convergent
  - to continued fraction, 63
- converging, 39
- conversions
  - between bases, 187
- convolution, 8, 268
- correction, 83, 98
- cost bound, 175
- Cramer's rule, 191
- cubic, xiii
- cubic spline, 304
- cycle, 140
- De Moivre's theorem, 191
- decoupling algorithm
  - for parallel evaluation, 10
- deflated polynomial, 211
- deflation, 165, 183, 186, 241, 245, 256
  - of known root, 67
  - problems with, 224
- derivative, 7, 114, 120, 132, 218, 259, 271, 292
  - Aho's method, 7
  - efficient methods for, 7
  - evaluation of, 2, 24
  - first, 3
  - logarithmic, 79
  - second, 120
  - use of in iteration function, 25
  - value of, 133
- Descartes' theorem, 53–55, 106
- determinant, 49, 208, 290, 307, 309
  - Hankel, 285
- diagonal matrix, 113
- diagonal similarity transformation, 302
- difference quotient, 168
- differences
  - backward, 14
- differential equation, 208
- disk, 85, 158, 161, 312
  - boundary of, 85
  - in interval methods, 84
  - inclusion, 89
  - initial, 96
  - iteration, 87, 95, 172
  - method (square root), 96
- distinct roots, 217
- divide and conquer method, 108
- divided differences, 107, 120
- double precision, 217, 270, 285, 307
- Durand-Kerner method, xiii
  - derivation of, 69
  - improved, 72, 115
- efficiency, 23, 25, 91, 95–100, 107, 131, 141, 144–147, 149, 168, 171, 172, 188, 195, 210
  - of Newton's method, 135
  - bound on, 26
  - comparison of, 114
  - computational, 175
  - index, 88
  - measure of, 114
  - of Bini's program, 121
  - of multiple root method, 156
  - on parallel processor, 119
- Ehrlich-Aberth method, 115, 117
  - disk version, 88, 100, 116, 117
  - Gauss-Seidel version, 104
  - improved, 72, 116, 117
  - improved (Gauss-Seidel version), 107

- eigenproblem, 211, 214
- eigenvalue, xviii, 106, 175–178, 208, 209, 215–217, 223, 225, 226, 232, 235, 245, 248, 250, 256, 278, 280, 281, 284, 289, 294, 304, 308, 310
  - method, 186, 300
  - problem, 207, 300
  - problem-generalized, 288, 305
  - simple, 217
- eigenvector, 209, 223, 310
  - initial, 245
  - left and right, 297
- eigenvector/eigenvalue pair, 243
- EISPACK, 298, 299, 314
- elimination
  - of elements, 234
- equation, xiii
  - non-linear, 111
- equations
  - homogeneous, 288
- error
  - actual, 270
  - backward, 45, 264, 266, 300, 301, 306, 307
  - bound (backward), 305
  - bounds, 46, 83, 84, 168, 175, 177
  - bounds (guaranteed), 161
  - criterion, 39
  - estimate, 87
  - forward, 266, 267, 306, 307
  - of approximation, 25
  - of multiple root, 268
  - relative, 270
- errors
  - avoiding, 47
  - in floating arithmetic, 47
- Euclidean algorithm, 44, 47, 62, 180, 182, 185, 215, 218, 231, 232
  - extended, 179
  - Schmeisser's modification, 214
  - stabilized, 47
- Euclidean norm, 302
- Euclidean polynomial remainder
  - sequence, 48
- Euler-like method, 99
- evaluation, 1, 22, 100, 101, 114, 141, 163, 247
  - accurate, 17
  - at equidistant points, 14
  - Clenshaw's algorithm, 186
  - cost of (per iteration), 114
  - instability of, 16
  - matrix-vector product, 12
  - multipoint, 11, 13
  - of Bernstein form, 187
  - of derivative, 97, 100, 188
  - parallel, 8
- evaluations, 116, 149, 168, 188, 189
  - new per iteration, 25, 88
  - per root, 91
- extension
  - circular, 86
- Farmer and Loizou's method, 84, 117
- Fast Fourier Transform, 8, 11
- Fibonacci numbers, 56
- Fiedler's method, 222, 223
- financial calculations, xiv
- fixed points, 132, 140
- fixed-point theorem
  - Schauder's, 214
- floating-point, 17, 18, 22, 43, 45, 46, 55, 100, 246
- folding method, 10
- formal orthogonal polynomial, 287
- Fourier, 136
- Fourier's theorem, 53–55
- fractional linear transformation, 307
- Frobenius norm, 305
- Fundamental Theorem of Algebra, 67
- g.c.d., 40, 43, 44, 47, 184, 186, 215–217, 224, 258–260, 271, 311
  - exact, 46, 47
  - near, 44, 46
  - triplet, 271, 274

- Gauss, xiii
- Gauss-Newton iteration, 258, 261, 262, 266, 268, 269, 271, 274–277
- Gauss-Seidel and SOR variations
  - of simultaneous methods, 102
- Gauss-Seidel variation, 99
- Gaussian elimination, 225
- geometric applications, 304
- Gershgorin's theorem, 175, 184
- Givens rotation, 232, 233, 237, 238, 240, 254, 256
  - complex, 235
  - parameters, 239
- Givens' method, 225
- globally convergent method, 120, 137
- GMP
  - multiprecision package, 122, 314
- Graeffe's method, xiii, xvii, 192
- greatest common divisor
  - (see g.c.d.), 39
- Halley's method, 97
  - high-order variations, 97
- Halley-like method, 99, 172
  - disk version, 97, 102, 116
- Hansen and Patrick's family of methods, 99, 100
- Hermite information, 25
- Hermite interpolatory method, 25
- Hessenberg form
  - upper, 236, 256
- high order method
  - similar to Newton's, 132
- Holder inequality, 295
- homotopy method, 157
  - of Pomentale, 193
  - of Wu, 192
- Horner's method, 1, 2, 6, 7, 9, 18, 22, 23, 67, 159, 169, 246
  - parallel implementation of, 8
  - repeated, 14
- Householder's method, 225
- Hull-Mathon procedure, 224
- hybrid method
  - modified and regular Laguerre's method, 93
- hybrid method(point plus disk), 87
- hybrid version
  - of Nourein's method, 89
- hypercube, 119
- ill-conditioning, 266, 280, 293, 306
- immediate basin, 139
- IMSL Library, 299
- inclusion property
  - of disks, 86
- initial conditions, 100
- initial guess, 68, 73, 77, 112, 117, 120, 131, 135, 139, 149, 173, 175, 188, 190, 191, 194, 213, 246, 258, 270, 276
  - for WDK method, 72
- inner product, 286, 312
- integrals
  - approximations to, 286
- integration
  - numerical, 282, 284, 287
- interior point, 165
- interlace theorem
  - Cauchy's, 176
- interlacing sets, 177
- interpolation, xvii, 11
  - Hermite, 188
  - inverse cubic, 149
  - inverse quadratic, 194
  - Lagrange, 76, 79, 247
  - linear, 174, 187
  - linear inverse, xvii
  - quadratic, xvii, 149
- interval, 37, 90
  - arithmetic, 84, 165
  - arithmetic (extended), 163, 164
  - circular, 172
  - complex, 173
  - extension, 162, 167
  - isolating, 61
  - matrix, 214
  - method, 161

- method (combined), 170
- method (disk- $k$ 'th root), 94
- method (for real roots), 90
- midpoint of, 162
- narrow, 166, 170
- rectangular, 89, 102, 161, 169
- reducing size of, 166
- inverse
  - of companion matrix, 209
  - of disk, 89
- inverse function, 144
- inverse iteration, 268, 272
  - guaranteed convergence of, 277
- inversion
  - centered, 96
- iteration, xvi, 4, 23, 165
  - function, 23
  - multipoint, 24–26
  - one-point, 24, 25
- iterative method, xvi, 17, 84, 132
  - point, 168
- Jacobian, 109, 112, 261, 264, 266, 267, 269, 273–275
  - inverse of, 212
- Jarratt's method, 141
- Jenkins-Traub method, xiii, xvii, 186, 195, 230, 299
- $k$ 'th root method, 95
- Kanno's SOR-GS method, 104
- King's 5th order method, 144
- Laguerre's method, xvii, 2, 93, 230
  - (modified) serial version, 107
  - modified, 93, 94
  - simultaneous version, 92
- Laguerre-like method, 99, 173
- LAPACK, 314
- Laplace transform, xiv
- Larkin's method, xiii
- latitude and longitude, 141
- leading coefficient
  - small, 304, 306
- leading principal minor, 292
- least common multiplier, 55
- least-squares problem, 258, 261, 263, 266
- least-squares solution, 275
- Leibnitz' theorem, 2
- linear equation, 208
- linear independence
  - of columns, 264
- linear system, 260, 274, 282
- linearly independent, 209
- linux, 122
- Lipschitz constant, 262
- LR algorithm, 289
- machine precision, xvi, 17, 20, 154, 246, 270, 293, 295, 300, 305
- Madsen-Reid code, 299
- mantissa, 4, 22, 160
- MAPLE, 222
- MATLAB, 186, 270, 285, 298, 299, 301, 305, 307, 314
  - function ROOTS, 270
- MATLAB program, 217
- matrices
  - similar, 289
- matrix
  - block diagonal, 176
  - colleague, 307
  - companion, xviii, 310
  - comrade, 309, 310
  - conjugate transpose, 236
  - convolution, 259
  - diagonal, 243, 302, 309
  - diagonal plus rank-one, 250
  - generalized semi-diagonal, 252
  - Hankel, 279, 287
  - Hermitian, 256
  - Hermitian diagonal+semiseparable, 257
  - Hessenberg, 240, 289, 291
  - Hessenberg form, 225
  - identity, 237, 303
  - method
    - for a few roots, 289

- methods, xiii, 207
  - errors and sensitivity, 294
  - programs, 314
- methods (of order  $N$ -squared operations), 231
- orthogonal, 225
- pencil
  - perturbed, 306
- product, 228
- rank-deficient, 271
- rank-one, 242
- real symmetric tridiagonal, 175
- real tridiagonal, 176
- right triangular, 227, 240, 254
- rotation or Givens, 226, 229
- similar, 220
- Sylvester, 258, 259, 271
- symmetric, 223
- symmetric tridiagonal, 225
- Szego-Hessenberg, 313
- trace of, 301
- triangular, 113
- tridiagonal, 214, 216, 231, 232
- unit lower triangular, 219, 220, 289
- upper triangular, 237, 260, 289
- Vandermonde, 288
- mean, 287
  - of a cluster, 284
  - of error, 160
- minimization problem, 264
- minimum
  - local, 268
- minor, 209
- Mobius transformation, 250
- modular method, 49
- Monsi's symmetric single-step method, 103
- MPsolve program, 121
- Muller's method, xiii, xvii, 141, 186, 195
- multiple
  - precision, 17, 47, 121, 168, 186, 193, 223, 224, 257, 258, 268, 270, 286
- root, 40, 43, 81, 84, 88, 95, 97–99, 107, 116, 121, 139, 155, 164, 168, 170, 173, 187, 209, 210, 216, 224, 230, 232, 263, 266, 267, 270, 278, 293, 299
  - matrix methods for, 257
  - root (approximations to), 70, 81
  - root (attainable accuracy of), 258
  - roots
    - simultaneous methods for, 80
- multiple root method, 175
- multiplication
  - vector-matrix, 240
- multiplicity, 38, 40–43, 80, 83, 84, 88, 97, 171, 184, 209, 216, 217, 222–224, 259, 270, 278, 282, 285, 286, 288
  - Chanabasappa's method, 153
  - Dong's method, 156
  - estimation by Schroeder's method, 153
  - estimation of, 152, 232
  - extrapolation method's, 155
  - finding by Lagouanelle's method, 154, 171
  - Hansen and Patrick's method, 153
  - Madsen's estimation, 153
  - of cluster, 83
  - Ostrowski's method of estimation, 153
  - Traub's method, 154
  - Ypma's treatment, 155
- multiplicity structure, 257, 258, 263, 264, 266–268, 270, 275–277, 316
- $n$ 'th roots of unity, 11, 247
- nested multiplication, 1
- Neta's 6th order method, 145
- Neta's order 16 method, 147
- Netlib, 299
- Newton sum, 286
- Newton's correction, 115
- Newton's identities, 150
- Newton's integral theorem, 144

- Newton's method, xiii, xvi, 7, 23–25, 40, 61, 68, 70, 109, 116, 120, 131–133, 137, 139, 144, 155, 158, 161, 174, 178, 183, 184, 186–188, 190, 307
- Chun's variation, 195
- cluster-adapted variation, 194
- computable convergence conditions, 137
- convergence in real root case, 141
- convergence of, 131, 138
- corrected, 152
- damped, 111
- derivation of, 131
- discrete mp version, 193
- earliest description, 135
- for system, 69, 212, 266
- Fourier's condition for convergence, 134
- Fourier's variation, 134
- generalizations, 141
- history of, 135
- homotopy version, 192
- Hubbard's set of initial points, 139
- hybrid with bisection, 188
- hybrid with other methods, 175
- hybrid with Steffensen's, 193
- interval version, 214
- Lagrange gave first modern, 136
- linear convergence to multiple roots, 80
- m-cycle improved, 188
- methods related to, 190
- modified, 288
- order of, 132
- Ostrowski's convergence condition, 133, 137
- parallel extended interval version, 174
- programs using, 189
- Raphson's version, 136
- Smale's convergence condition, 137
- versions for multiple roots, 151
- Newton-Fourier method, 168
- Newton-like method
  - parallel versions, 173
  - simultaneous, 173
- Newton-Moore-like methods, 173
- nonlinear system
  - Vieta's, 266
- Nourein's method, 71, 91, 115
  - convergence of, 79
  - disk version of, 89
- null-space, 260
- number
  - double length, 20
- numbers
  - normalized, 22
- numerical methods, xiii
- Omar Khayyam, xiii
- order
  - of iteration with  $n$  evaluations, 146
- orthogonal polynomial, 308
- Ostrowski's method, 95
- Ostrowski's square root method
  - disk version, 101
- Ostrowski-like method, 99
- over-determined system, 261, 264, 275
- overflow, 22, 23, 185, 230, 241, 242
  - prevention of, 22
- parallel computers
  - simultaneous method on, 117
- parallel computing, 67, 68
- parallel network topologies, 119
- parallel processors, 9
  - minimum number of, 9
- parallelism, 103
- partial fractions, 82
- pejorative manifold, 258, 263, 267
- pejorative root, 264, 266–268
- permutation method
  - of estimating errors, 160
- perturbation, 44, 81, 186, 246, 258, 267, 270, 294, 295, 299, 306
  - coefficientwise, 295, 300
  - infinitesimal, 296
  - method, 160

- multiplicity-preserving, 267
  - normwise, 295
  - unrestricted, 266
- perturbed polynomial, 305
- point-slope method, 173
- polar and spherical coordinates, 140
- polar coordinates, 191
- polynomial, xiii, xviii, 40
  - low-degree, 23
  - roots of, xiii, xiv
- polynomial remainder sequence
  - (see prs), 48
- power form, 186
- power method, 210
  - for companion matrix, 209
  - inverse, 242
  - shifted inverse, 245, 246
- power-of-x form, 312
- powers of x
  - evaluation of, 10
  - parallel computation, 9
- precision
  - infinite, 43
- program for simultaneous methods, 121
- prs
  - normal, 49
  - Primitive, 48
  - Reduced, 48, 49
  - Subresultant, 49
- PRSS method
  - mixed interval-point version, 103
- pseudo-inverse, 274
- pseudo-remainder, 48, 49
- pseudospectra, 299
- pseudospectrum, 294, 296
- pseudozero set, 294, 295, 299
- QR factorization, 241, 256, 260, 268, 272, 275
- QR iteration, 235, 239, 254, 256, 257
  - modified, 235
- QR method, 215, 216, 218, 220, 223–226, 230, 231, 233, 289, 305, 314
  - description of, 225
  - shifted, 236, 253
- QUADPACK, 286
- quadratic, xiii, 13, 43, 92, 108, 120
  - factor, 111, 113, 169, 181
- quadrature rule, 144
- quadruple precision, 299
- quartic, xiii
- quotient
  - $p(x)$  divided by  $z-x$ , 2
  - partial, 63
- QZ-algorithm, 306
- radicals
  - solution by, xiii
- radius, 101, 172
  - of disk, 85
- random errors, 232
- random numbers, 252, 253
- random variables, 190
- range, 162
  - evaluation of, 162
  - of interval, 165
- rank deficiency, 265
- rationals, 47
- real factorization methods, 108
- real root case, 120, 141, 175, 215
- reciprocal polynomial, 209, 224, 294
- rectangle
  - enclosing, 171
  - in interval methods, 84
- rectangular region, 286
- rectangular rule, 144
- recurrence relation, 112, 178, 312
- recursive methods, 90
- references
  - for bounds, 31
  - for chapter 1, 33
  - for chapter 2, 51
  - for chapter 3, 65
  - for chapter 4, 123
  - for chapter 5, 196
  - for chapter 6, 316
- regular termination, 215, 220
- relative error, 160



- remainder, 47
  - partial, 48
  - sign of, 47
- repeated symmetric single-step method, 103
- residuals, 19, 45, 46
- residue theorem, 279
- resultant, 49
- reverse RQ step, 239
- root, xiii, xiv
  - approximate, 47
  - delta-approximated, 246
  - k-fold, 43
- root error
  - absolute, 307
- roots
  - common, 113
  - distinct, 215, 286
  - multiple, 40, 42
  - of unity, 282
  - real, 37
  - real-isolation of, 55
  - simple, 39, 43, 217, 220
  - upper bound on, 39
- rounding
  - after operation, 17
- rounding error, 4, 6, 22, 43, 55, 100, 102, 154, 156, 158, 160, 164, 186, 187, 222, 302
  - alternative expression, 6
  - bound on, 4, 7, 163
  - estimate of, 6, 7
  - in evaluation, 4, 17, 160
  - in simultaneous method, 100
  - increased, 67
- Routh's theorem, 51
- Routh-Hurwitz criterion, xiv
- RQ step, 242
- Runge-Kutta method, 144
- S-series, 308
- Sakurai's method, 84
- sample point, 16
- scalar product
  - precise, 17, 19
- scaling, 22, 184, 232, 298, 312, 314
  - factor, 22
  - Fortran implementation, 23
- SCARFS
  - program, 195
- Schroeder's method, 224
- Schur-Cohn algorithm, 313, 314
- secant method, xvii, 24, 25
  - optimal, 194
- secant-like method, 194
- secant-Newton method, 195
- second derivative
  - methods involving, 92
- semi-diagonal
  - of rectangle, 90
  - of rectangular interval, 102
- sensitivity, 263, 267, 294
  - structure-constrained, 268
- shared memory, 173
- Sherman-Morrison-Woodbury formula, 244
- shift, 226, 227
  - exceptional, 235
- sign variation, 38, 53, 56, 59, 60, 181
  - number of, 53
- signal processing, xiv, 14
- similarity transformation, 297
- Simpson
  - true inventor of "Newton's method", 135
- simultaneous methods, xiii, xvi, 11, 67, 73, 117, 161
  - order of, 114
- singular matrix, 111, 113
- singular value, 258–260, 262, 266, 267, 271, 272, 274
- singular vector, 260, 261, 271
- slope method, 173
- sparsity structure, 275
- spectral radius, 106
- speed-up
  - on parallel processor, 117, 119
- spherical coordinates, 141
- splitting point, 176

- square root, 98, 235
- square root method, 84
  - Petkovic's variation, 98
  - with Halley correction (Gauss-Seidel version), 107
- square roots
  - methods involving, 92
- stability, xiv, xv, 221, 241
  - of zerofinding algorithm, 295
- staggered format, 18
- standard deviation
  - of error, 160
- stationary iteration, 293
- stereographic projection, 140
- stopping criterion, 131, 132, 135, 158, 161, 164
  - Adams', 6, 117
  - Garwick's, 7, 159
  - Igarashi's, 6, 159
  - Vignes', 160
- Sturm sequence, xvi, 26, 37–40, 47, 50, 51, 90, 119, 182, 183
  - using integers, 47
- Sturm's theorem, 38, 51, 53, 186
- subinterval, 162, 165, 174
- submatrix
  - principal, 176, 245
- subresultant method, 51
- symbolic algebra, 43
- system of nonlinear equations, 211
- Szego polynomial, 311–313
  
- Taylor's series, 2, 191
- Taylor's theorem, 132, 142
- termination
  - of iterative methods, 100
- Toeplitz matrix, 110
  - triangular, 110
- trace
  - preservation of, 234
- transfer function, xiv
- trapezoidal rule, 144, 282
- triangular form
  - upper, 256, 257
- triangular linear system, 275
- underflow, 22, 23, 185, 242
  - minimization of, 22
  - reduction of, 22
- union
  - of intervals, 164
- vector machine, 230
- Vincent's theorem, 54, 55
  - Akritas' improvement, 56
  
- WDK correction, 89, 99, 114
- WDK method, 67, 68, 75, 76, 78, 81, 84, 91, 108, 111, 115, 117, 266
  - (nearly always converges), 72
  - asynchronous parallel version, 118
  - Atanassova's generalization of, 90
  - convergence of, 77, 78
  - disk version, 87, 88, 102, 116
  - for multiple roots, 80
  - Gauss-Seidel version, 115
  - modification of, 103
  - on parallel processor, 118
  - parallel or total-step variation, 103
  - rectangular interval version of, 89
  - serial or single-step variation, 103
  - variation of, 90
- Weierstrass' correction
  - see WDK correction, 96
- Weierstrass' method, 67
- Werner's method, 67, 145
- Wilf's method, 51
- Wilkinson's polynomial, 242, 253, 299
- work
  - in iteration, 25
  
- z-transform, xiv
- Zeng's program MULTROOT, 314