EXPLAINING AND AMELIORATING THE ILL CONDITION OF ZEROS OF POLYNOMIALS

by

David Granville Hough

Memorandum No. UCB/ERL M77/30

6 May 1977

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

EXPLAINING AND AMELIORATING THE ILL CONDITION OF ZEROS OF POLYNOMIALS

David Granville Hough

Ph.D. Computer Science

Sponsor: Office of Naval Research

W. Kahan
Chairman of Committee

Abstract

Physical systems can frequently be modeled by polynomial equations. Then interesting properties of the systems can be determined from the zeros of the polynomials. Standard codes compute those zeros from the coefficients in a stable fashion. But what should be done if the zeros are inherently hypersensitive to changes in the coefficients of their polynomials? Newly developed methods can be used to explain such an ill conditioned polynomial by exhibiting a nearby polynomial with one or more multiple zeros which are well conditioned. Furthermore these methods can be abused by uncritically replacing the ill conditioned polynomial with the well conditioned one nearby. When such a replacement is unwarranted, bounds can be obtained on the variation of the zeros corresponding to the uncertainty in the coefficients. One way to obtain such bounds is to exploit the nearby well conditioned polynomial to obtain a revision of the classical Puiseux fractional power series expansions of the zeros.

These notions have been investigated experimentally in a long series of computer calculations. In the course of these calculations the existing stock of numerical techniques has been augmented. A new way is now known for computing the condition numbers which measure the condition of zeros. The previously known equations to be solved for the nearest polynomial with a single multiple zero are now joined by

equations for the nearest polynomial with a complex conjugate pair of double zeros and equations for the nearest polynomial with several distinct double zeros. All these equations have simplified forms because certain Lagrange multipliers vanish in the complex case. But some examples demonstrate that when only real perturbations are considered, the Lagrange multipliers do not always vanish. Finally, there is some theory about the location of the nearest polynomial with a double zero.

The numerical experiments show that Newton's method may be used successfully to solve the equations in the cases of greatest interest when the expected result is sufficiently simple. The techniques may also be applied to polynomials such as Wilkinson's famous example whose zeros are the integers from 1 to 20. But then the numerical results suggest that that ill conditioned polynomial can not be explained successfully as a small perturbation of a well conditioned polynomial. Instead Wilkinson's polynomial lies in a region of polynomial space whose geometry seems to be exceptionally complicated.

Bounds on uncertainties in zeros corresponding to uncertainties in coefficients are customarily computed with Taylor series. For ill conditioned simple zeros these Taylor series have radii of convergence that are much too small. The well conditioned multiple zeros of a nearby polynomial are not amenable to Taylor series expansions but may be expanded in a Puiseux fractional power series. These fractional power series, however, also have unsatisfactory regions of convergence. But by choosing a different starting point the convergence problem of the Puiseux series can be overcome to produce, in principle, series that converge rapidly throughout the region of interest. In practice

those series are used to produce realistic bounds on the uncertainties in the zeros. Full exploitation of these techniques awaits adequate facilities for symbolic algebra.

ACKNOWLEDGMENTS

The research to be described owes much to my advisor, Professor W. Kahan, who suggested the topic and patiently guided my research to completion.

Professor B.N. Parlett obtained much of the financial support for this and other numerical analysis projects at Berkeley. Research assistantships and computer time were mostly provided by the Office of Naval Research Contract Number N00014-69-A-0200-1017 for which Parlett is principal investigator.

Professors Parlett and T. Kato served on the dissertation committee and provided numerous valuable suggestions. At various times the Computer Center of the University of California, Berkeley, and the Applied Mathematics Division of Argonne National Laboratory provided financial support, office space, and computer time for which I am grateful.

Ms. Ruth Suzuki did the typing in her usual expert fashion.

CONTENTS

CHAPTER

I INTRODUCTION AND MOTIVATION 1

1. What is the Problem? 1
2. What is Ill Condition? 3
3. Examples of Definitions 8
4. What is Ill Condition of Zeros of Polynomials? 11
5. Treating the Symptoms of Ill Condition 14
6. Explaining Ill Condition 16
7. What Do We Do with the Explanation? 23
8. Survey of Previous Results 25
9. Summary of Findings 28
10. Notation 30

II COMPUTING CONDITION NUMBERS FOR ZEROS OF POLYNOMIALS . 34

1. Definition of Condition Numbers for Simple Zeros 34
2. Definition of Condition Numbers for Multiple Zeros 37
3. Condition Numbers for n-tuple Zeros 42
4. Resolution of Condition Number into Components 43
5. Computing σ for Arbitrary Norms -- Dual Method . 46
6. Computing σ for ℓ_2 Norms -- Dual Method 50
7. Computing σ for ℓ_2 Norms -- Primal Method . . . 52
8. Computational Details 53
9. Condition Numbers for Complex Conjugate Zeros of Real Polynomials 54
10. Computing σ_c for ℓ_2 Norms 56
11. General Condition Numbers 60
12. Application of the Idea of General Condition Number 62
13. Condition Number vs. Distance to Submanifold . 64

III FINDING THE NEAREST POLYNOMIAL WITH AN m-TUPLE ZERO . . 66

1. Introduction 66
2. The Nearest Polynomial with an n-tuple Zero . . 67
3. The Nearest Polynomial with a Fixed Double Zero 70
4. The Nearest Polynomial with a Double Zero . . . 73
5. The Nearest Polynomial with a Fixed m-tuple Zero 76
6. The Nearest Polynomial with an m-tuple Zero, No Longer Fixed 78
7. Computational Details: The Equation to Solve for the Nearest m-tuple Zero 82
8. The Second Derivative of $\|q\|$ 87
9. The Last Lagrange Multiplier is Zero at a Minimum 91

CHAPTER

III
10. Another Kind of Second Derivative 94
11. Computational Details: A Constrained Hessian for ν 97

IV FINDING THE NEAREST REAL POLYNOMIAL WITH A COMPLEX CONJUGATE PAIR OF m-TUPLE ZEROS 100

1. Introduction 100
2. The Nearest Polynomial with a Complex Conjugate Pair of m-tuple Zeros 101
3. Divided Differences for the Equations for a Complex Conjugate Double Zero 107
4. Computational Details: The Equations to Solve for a Complex Conjugate Pair of Double Zeros . 110
5. The Rows of B are Linearly Independent 113
6. The Last Lagrange Multiplier is Zero 116

V FINDING THE NEAREST POLYNOMIAL WITH MORE THAN ONE MULTIPLE ZERO 121

1. Introduction 121
2. The Nearest Polynomial with Several Multiple Zeros 122
3. The Last Lagrange Multipliers are Zero 125
4. Equations for k Real Double Zeros 128
5. Deflation for Several Double Zeros 132
6. The Equations for Two Real Double Zeros 133

VI LOCATION THEORY FOR NEAREST POLYNOMIALS WITH A DOUBLE ZERO 138

1. Introduction 138
2. No Complex Solutions for Certain Real Polynomials 142
3. Counterexample 144
4. A Bound on the Solutions ζ 147
5. Propositions for Real Quadratic Polynomials . . 150
6. Swindle Results for Real Quadratic Polynomials 154
7. The Smallest Circle Containing Two Zeros Need Not Contain a ζ 158
8. Infinitesimal Location Theory 163

VII PERTURBATION THEORY FOR MULTIPLE ZEROS OF POLYNOMIALS . 166

1. Introduction 166
2. Classical Theory of Expansions of Algebraic Functions 169
3. Failure of Classical Taylor and Puiseux Series Expansions 176
4. Why Find the Nearest Polynomial with the Multiple Zero? 181

CHAPTER

VII
5. Resolving Expansions into Components 184
6. A Practical Technique for Bounding Changes in Zeros 190
7. An Example of Expansions 207

VIII EXPERIMENTAL METHODS 222

1. Introduction 222
2. How the Equations were Solved 223
3. How Do We Know the Answers are Correct? 225
4. Computed Checks on Results 228
5. Setting Up a Problem with a Known Solution . . 232

IX NONPATHOLOGICAL EXPERIMENTAL RESULTS 237

1. Introduction 237
2. n-tuple Zeros 239
3. Returning to a Double Zero 240
4. Returning to a Triple Zero 242
5. Returning to Two Double Zeros 243
6. Returning to a Complex Conjugate Pair of Double Zeros 244
7. A Polynomial with Several Pairs of Complex Conjugate Zeros 245
8. An Uninteresting Polynomial 246
9. Zeros in a Circle 247
10. Summary 248

X WHAT'S WRONG WITH WILKINSON'S POLYNOMIAL? 249

1. Wilkinson's Polynomial 249
2. Coefficients and Condition Numbers for Wilkinson's Polynomial 250
3. The Nearest Polynomial with a Double Zero . . . 258
4. Interesting Polynomials Near Wilkinson's . . . 261
5. Discussion of Results 268
6. Numerical Results for Translation 271
7. Zeros in Geometrical Progression 273

XI CONCLUDING REMARKS 274

APPENDICES 277

1. Using the Zeros of a Polynomial to Compute Its Coefficients 277
2. Simultaneous Evaluation of a Polynomial and Some of its Derivatives 280
3. Partial Derivatives of a Deflated Function of a Complex Variable 283

APPENDICES
4. Computing the Divided Differences Required in the Equations to be Solved for Complex Conjugate Double Zeros in Chapter IV 285
5. Computing the Divided Differences Required in the Equations to be Solved for Two Real Double Zeros in Chapter V 288
6. The Lagrange Multiplier Theorem 294

REFERENCES 295

CHAPTER I

INTRODUCTION AND MOTIVATION

1. What is the Problem?

The research to be reported in the following chapters deals with "ill condition" of the zeros of polynomials. "Ill condition" means unusually great sensitivity of the zeros to changes in the coefficients of the polynomial.

Consider the following example: a physicist has determined that a parameter of interest may be determined by finding the zeros of a polynomial. He computes the coefficients of the polynomial and solves for its zeros with any of a number of computer codes which find zeros of polynomials. Then the computer states that his polynomial of degree six has the following zeros:

$$-2.0$$
$$-1.0$$
$$+ .99999998 \pm .000104625\ i$$
$$+2.0$$
$$+3.0$$

Perhaps being distrustful, the physicist computes the coefficients of the polynomial which has exactly these zeros. He finds that those reconstituted coefficients agree with the original coefficients of the polynomial he gave the computer to well within the uncertainty in the coefficients, which were derived from experimental data. He will usually find that the differences between those sets of coefficients are comparable in size to a few rounding errors, so he seems to have no grounds for complaint with the computed result.

None the less there may be sound physical reasons why the answers he seeks can not have imaginary components. Then why do they appear in his answer? Is he justified in ignoring them? The methods proposed in the following chapters provide a way of dealing with these questions.

Those methods would "explain" the physicist's quandary as follows. First they would show that the two complex conjugate zeros are extremely ill conditioned. That is, small changes in the coefficients comparable with experimental error could easily cause them to undergo much larger real or complex changes. The ill condition arises from the fact that the physicist's polynomial is very close to a polynomial with a double zero. In fact, the methods we will discuss show that changing each coefficient of the polynomial by as little as one part in 10^9 suffices to cause the polynomial to have a double zero at 1.0. That double zero is well conditioned, in a sense to be explained later. Therefore the physicist might "ameliorate" the condition of the answers to his problem by accepting a double zero at 1.0 in place of the complex conjugate pair <u>if</u> the experimental uncertainties in the coefficients exceed one part in 10^9 <u>and</u> there is physical justification for assuming that his answer should be in the form of a double zero. Where that justification is lacking, the ill condition of the result is a warning signal that a misjudgment in the design of the experiment and computation may have invalidated the results.

2. What is Ill Condition?

We turn now to precise definitions of terms like posedness, condition, and stability. The terms have been defined by numerical analysts in many different and sometimes inconsistent ways; our definitions will be those used by W. Kahan in numerical analysis courses at the University of California, Berkeley [18]. These definitions are also close to those in the widely used text by Dahlquist and Björck [6].

The definitions to follow make sense if one thinks of a problem having a definite set of input data and a similar set of output data which we call the solution. For instance, in the problem of determining the n complex zeros of an n'th degree polynomial, the $n+1$ coefficients of the polynomial are the input data and the n zeros are the solution. In contrast, the "problem" of finding a polynomial approximating a given function is incomplete until we specify a criterion for choosing the best approximation. That criterion could be regarded as fixed, and hence part of the problem, or subject to change, and hence part of the data.

If furthermore the data are regarded as uncertain, then the information on the size of the uncertainty becomes part of the data. This information is often expressed in terms of a metric or norm on the space from which the input data are drawn. The norm itself may also be part of the input data if it is subject to change. The purpose of the norm on the input data, for example, is to provide a way for the problem poser to specify which inputs are so close together as to be indistinguishable from his point of view. In addition, there may be a norm on the output solution with a similar purpose. As we shall

see, the poser may be obliged to provide these norms even if the input data are regarded as exact.

Within this framework a problem is well posed if it (1) has a solution which (2) is unique and (3) varies continuously when the input data vary continuously. Consequently an ill posed problem may, for some input data, have several solutions or none or the solution may change discontinuously when the input data is changed continuously. The answer to the question of whether a problem is well posed is either yes or no.

Given a problem that is analytically well posed, we call it well conditioned if changes that we consider negligible in the input data can only cause changes in the solution that we also consider negligible. Conditioning can be measured by computing the partial derivatives of the solution with respect to changes in the input data. If the appropriate norm of these partial derivatives, called the condition number, is too large, the problem is ill conditioned. Unlike posedness, then, there is not a sharp break between well and ill conditioned problems, but rather a continuum.

From our point of view, stability is a property of algorithms, rather than problems, and relates to the question, "Does this algorithm always produce a solution as good as can be expected, considering the condition of the problem?" Interesting numerical algorithms almost always fail to produce the mathematically correct solution to a problem. This is because such algorithms usually commit rounding errors due to finite precision arithmetic and truncation errors due to terminating infinite analytical processes after a finite number of steps.

A stable algorithm has the property that the uncertainty it contributes to the solution of a problem is not much larger than the uncertainty that would be associated with small changes in the input data. Figures I.1 and I.2 illustrate a stable algorithm applied to an ill conditioned problem. A stable algorithm applied to a well conditioned problem yields nearly the correct answer. Many stable algorithms, moreover, can be shown to deliver the exact solution of a problem with input data very near the given input data, even if that data is ill conditioned.

To conclude the definitions, recall that the key to the problem of the physicist in section 1 was to find the polynomial with a double zero nearest his polynomial. In general, the polynomials with one or more multiple zeros form a subset of the space of all polynomials. These subsets have been called pejorative manifolds by W. Kahan [17], because polynomials near a pejorative manifold always have some ill conditioned zeros. Since they are the only manifolds that interest us, we will use the term manifold in subsequent chapters to mean one of these pejorative manifolds. Thus the manifold of n'th degree monic polynomials with one m-tuple zero is a surface with dimensionality $n-m+1$ in the space of all n'th degree monic polynomials.

The distinction between wrong answers caused by an ill conditioned problem and wrong answers caused by an unstable algorithm applied to a well conditioned problem is well known in the west mostly because of the work of Wilkinson [34]. But similar concepts are also present in the contemporaneous work of the Soviet author V. Zaguskin [37]. Zaguskin defines condition numbers with respect to small finite rather than infinitesimal perturbations. In well conditioned cases his

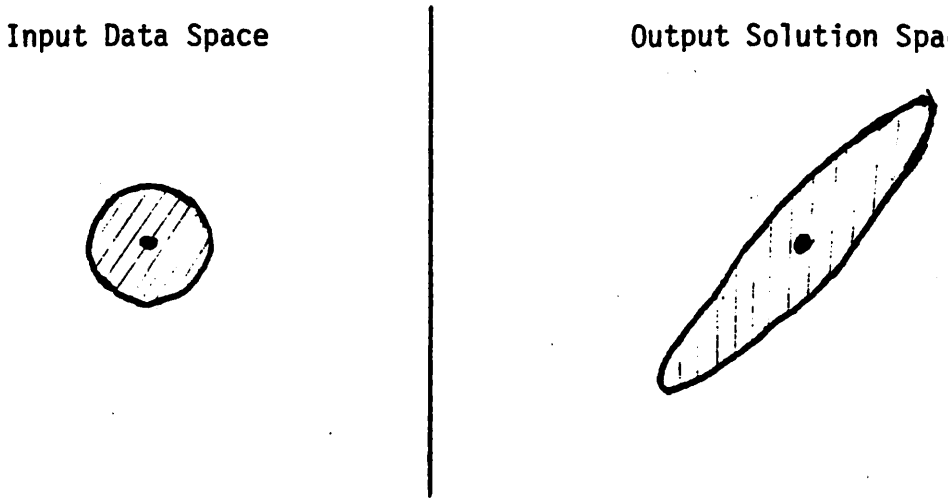

Figure I.1. Effect of ill conditioning: a ball in the input space maps into a cigar-shaped region in the solution space.

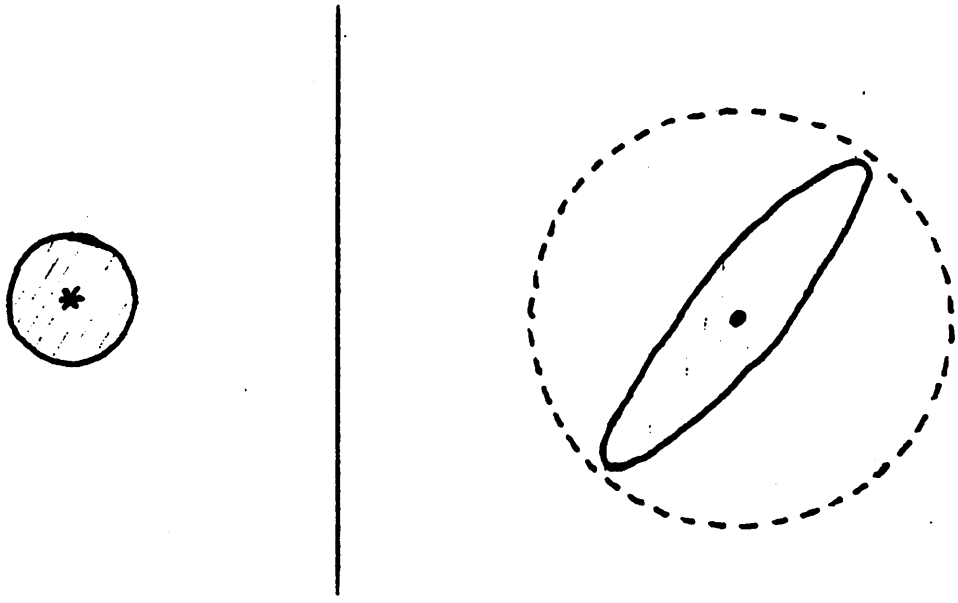

Figure I.2. A stable algorithm maps the input point * into the region bounded by the dotted ball which is not much larger than the image of the input ball.

methods give an idea of how much the zeros of a polynomial may vary as the polynomial varies within its finite uncertainty. In chapter VII we will show how such notions may be applied even for an ill conditioned polynomial. There we will show how to develop the whole series of which the infinitesimal condition number is simply a bound on the first term.

3. Examples of Definitions

An example might help to clarify the definitions of the previous section. Consider the problem of finding the smaller real zero of the quadratic polynomial

$$f(x) = x^2 + 2x + 1 - \varepsilon \quad \text{for} \quad |\varepsilon| \leq 0.1 \ .$$

We see that for $\varepsilon = 0$, there is a real double zero; for $\varepsilon < 0$ there are no real zeros; for $\varepsilon > 0$ there are two distinct real zeros. Since in some cases of the input there is no solution to this problem, it is ill posed.

Suppose we restrict the problem so $0 \leq \varepsilon \leq 0.1$. Now the problem has become well posed but ill conditioned. Consider the dependence of the zeros of f on ε:

$$x_{\pm} = -1 \pm \sqrt{\varepsilon} \ ,$$

$$\frac{\partial x_{\pm}}{\partial \varepsilon} = \pm 1/(2\sqrt{\varepsilon}) \ .$$

So as $\varepsilon \to 0$ this condition number becomes arbitrarily large in magnitude. Any small error in the original data or in the computation may be magnified by an arbitrarily large factor. Note how in this case, as in many others, approaching ill posedness corresponds to worsening condition. See Kahan [17].

What are the pejorative manifolds in the quadratic case? There is just one, the manifold of quadratics with double zeros. In the space of quadratics

$$x^2 + bx + c \ ,$$

the manifold of polynomials with double zeros is just the subset of

polynomials with

$$b^2 = 4c \ .$$

It is evident that the previous polynomial

$$x^2 + 2x + 1 - \varepsilon$$

lies rather near this manifold; that nearness causes the ill condition of its zeros.

Stability may be illustrated by considering the problem of finding the small real zero $\tilde{x}$ of the polynomial

$$x^2 - 2x + \delta \ ,$$

for $|\delta| \leq 10^{-20}$. The usual formula yields

$$\tilde{x} = 1 - \sqrt{1-\delta} \ .$$

On most computers there will be numbers δ large enough to be representable but small enough that the computed value of $1 - \delta$ is 1. In this case the computed $\tilde{x} = 0$. For many purposes this is unacceptably far from the correct answer which is $\tilde{x} \doteq \frac{1}{2}\delta$. A check of condition numbers shows that they are small. That the fault lies with the algorithm implementing the usual formula, rather than with the problem, can be seen by considering another less well known but equivalent formula for the zero:

$$\tilde{x} = \delta/(1 + \sqrt{1-\delta}) \ .$$

An algorithm implementing this formula will compute an approximately correct answer for small δ even in the face of rounding error.

This should come as no surprise since this polynomial is obviously far from the pejorative manifold.

4. What is Ill Condition of Zeros of Polynomials?

The chapters to come will discuss methods for dealing with ill conditioned zeros of polynomials. In order to see why such methods might be useful, we consider first the problem of finding the zeros of a polynomial from its coefficients. Several algorithms are now known which are not only stable in the sense outlined above, but also are more efficient than other (unstable) methods. Best known of these is that of Jenkins and Traub [14]; another good one is Brian Smith's version of Laguerre's method [30]. FORTRAN implementations of both these algorithms are available in the IMSL library [13]. The stability of these algorithms may be shown for a specific problem by computing the coefficients of a polynomial whose zeros are exactly the zeros computed by the algorithm. Then the coefficients of the original polynomial do not differ much from the coefficients of the polynomial recomputed from the numerical solution.

But if we happen to know the exact zeros of the original polynomial, we may find that they differ greatly from the zeros that were computed. If this is the case -- that a stable algorithm has produced results that are more than slightly wrong -- then the problem must be ill conditioned. In the previous section we saw that the condition of zeros of a quadratic polynomial was related to how nearly the polynomial came to having a double zero. It is a basic fact about the zeros of analytic functions that nearness to a function with a multiple zero corresponds to ill condition of the zeros.

As a simple example consider the analytic function

$$f(\tau) = (\tau-\alpha)^m g(\tau)$$

where $g(\tau)$ is analytic and $g(\alpha) \neq 0$. If $f(\tau)$ is perturbed by $\varepsilon h(\tau)$, $h(\alpha) \neq 0$, then the perturbed zeros β satisfy

$$f(\beta) - \varepsilon h(\beta) = 0 ,$$

so

$$\varepsilon = (g(\beta)/h(\beta))(\beta-\alpha)^m .$$

In chapter VII we will see that the last equation can be transformed to express $\beta - \alpha$ as a power series in $\varepsilon^{1/m}$. Thus there are m zeros β which converge to α as $\varepsilon \to 0$.

Implicit differentiation reveals the dependence of a solution β on the data ε:

$$\frac{d\beta}{d\varepsilon} = \frac{1}{\varepsilon} \cdot \frac{1}{m\left(\frac{g(\beta)}{\varepsilon h(\beta)}\right) + \left(\frac{g'(\beta)}{g(\beta)} - \frac{h'(\beta)}{h(\beta)}\right)} .$$

As $\varepsilon \to 0$, $\beta \to \alpha$, $g(\beta) \to g(\alpha)$, and $h(\beta) \to h(\alpha)$. Simultaneously the condition number $\left|\frac{d\beta}{d\varepsilon}\right|$ increases like $1/(|\varepsilon|^{1-1/m})$ without bound, so the condition of each β becomes infinitely bad.

One way to visualize the meaning of the condition number is to think of the process of finding a zero of a polynomial as a mapping from the space of polynomials into the complex plane. Then we can ask how an infinitesimal neighborhood in polynomial space is mapped into the complex plane. If that neighborhood is spherical then its image will usually look elliptical. In a well conditioned case the ellipse is small; in an ill conditioned case large. In the case of an infinitesimal neighborhood of a polynomial with a multiple zero, the image is a large star-shaped region.

The research to be described is motivated by the desire to know how large these image regions may become for polynomials within a finite ball. The condition number tells how large the ellipses may be in the infinitesimal case; it can be used to bound the first term of a power series. Just when that first term is large, however, the power series turns out to have a short radius of convergence. In fact, if a manifold of polynomials with multiple zeros runs through the ball, then the usual power series can not converge at every point in the ball.

But by exploiting that manifold as described in chapter VII we may be able to get, in principle, a different kind of series that converges throughout the ball. The notion underlying that series may be used, in practice, to obtain a bound on the size of the image of the ball.

If the polynomial from which we expand lies on a manifold, the nature of series expansions of its multiple zeros is different than when the polynomial lies off the manifold. The series includes fractional powers of the perturbations. This is not a severe handicap. However it may be that there are a priori reasons for knowing that the only significant perturbations are those which are along the manifold and maintain multiplicities. Then reasonable condition numbers can be defined which are finite with respect to those perturbations. Furthermore the expansions used to bound the changes in the zeros take much simpler forms.

5. Treating the Symptoms of Ill Condition

Large condition numbers are a warning that small changes in the input data cause large changes in the solution of a problem. In the next section we consider ways of identifying the underlying difficulty, but now we will merely treat the symptoms: substantial changes in our answers are being caused by seemingly insignificant changes in our data or by rounding and truncation errors in our algorithms.

If our data is derived experimentally, we could try to perform more careful experiments in order to get the variation in our answers within acceptable limits. If the data is not subject to empirical uncertainties, then the errors in our algorithms are the cause of our symptoms. We may use increased precision to reduce the effect of rounding errors, and we may carry out more steps of infinite processes to reduce truncation errors. For polynomials, this would mean carrying out more steps of iterative processes such as Newton's method.

If the coefficients of a polynomial are known exactly, then rational arithmetic may be used to determine the zeros to any required accuracy. Pinkert [41] discusses such a method. These methods are relatively slow on present computers, but they do eliminate ill condition as a factor affecting accuracy of computed zeros. Exact arithmetic methods are inappropriate, however, when the coefficients are not precisely known; then explicit account should be taken of ill condition.

Changing the algorithm does not change the condition of the problem, but an unstable algorithm can aggravate our symptoms of ill condition. Sometimes we can reformulate the problem to take advantage of a stable algorithm. In other cases we can reformulate the problem

to make it better conditioned.

Thus we will see later that the condition of a zero of a polynomial may sometimes be improved by translating the polynomial so that the zero to be found is near the origin. In certain cases this may be helpful, but care must be taken that the translation is computed with insignificant rounding error. The translation of the coefficients is computed effectively by evaluating the polynomial and n of its derivatives. Usually such translations must be performed in higher precision when ill conditioned zeros are involved. Stewart [31] shows that the effect of such translations, carried out in conventional fashion, is comparable to the effect of rounding errors in the coefficients of the original polynomial. Kahan [18] has shown that unconventional algorithms can sometimes do better than would be expected from [31], but his algorithm is a fluke.

If one is concerned with numerical treatment of a polynomial that arises experimentally, it may be that careful translation is the most reasonable method of "ameliorating" ill condition that has no obvious source. Such translation is justified if the zeros represent a physical quantity whose origin is arbitrary. The coordinates of a point on a line, for instance, are sometimes arbitrary, but not if something interesting, such as a body exerting a central force, occurs at the origin.

However performed, translation amounts to attacking the problem of ill condition piecemeal, one zero at a time, rather than trying to deal with the overall condition of the problem. And the results of translation in no way "explain" the ill condition.

6. Explaining Ill Condition

The methods to be presented later try to "explain" ill condition by finding the nearest polynomial with all its zeros well conditioned. That polynomial will be on one of the pejorative manifolds of polynomials with multiple zeros. At the end of chapter II we will see that if an m-tuple zero is sufficiently ill conditioned there must be a polynomial with an m+1-tuple zero fairly close by. So we may in succession try to find the nearest polynomial with a double zero, a triple zero, two double zeros, and so on. We may count ourselves successful if we find that one of these nearest polynomials has all of its zeros well conditioned and yet is close enough to our original polynomial. When we are successful, our starting polynomial may be explained as a small perturbation of a polynomial with some multiple zeros, all of which are well conditioned.

The reader with some experience may feel that the nearest such polynomial should be apparent from inspection of the distribution of zeros, for ill conditioned zeros often form obvious clusters. After all, an m-tuple zero subjected to a suitably small perturbation will usually split up into m distinct zeros, and such configurations should be easily recognized. However, the ill conditioned simple zeros scatter so quickly that they may soon lose their clustered aspect. As we shall see later when we discuss Wilkinson's polynomial, it is sometimes impossible to guess just by inspection of the zeros what the nearest polynomial with well conditioned zeros might be like.

We may find, moreover, that no small perturbation will get us to a polynomial with all zeros well conditioned. Rather, by moving increasing distances we may increasingly improve the condition of the

zeros, but in order to improve the condition of all zeros as much as we want it is necessary to move much further than we want. Wilkinson's polynomial seems to be of this sort; it is discussed in chapter X. There is no natural division between the polynomials which are explainable and those which are not; however we set a somewhat arbitrary boundary by our choice of norm and tolerance.

If we do find a nearby polynomial with all of its zeros well conditioned with respect to variations that maintain multiplicities, then we might say that moving to the new polynomial has ameliorated the problem of ill condition. Such a viewpoint makes sense only if the new polynomial is indistinguishable from the original and it is reasonable to hypothesize that the original problem could have a built in constraint in favor of multiple zeros. This constraint may have existed unrecognized heretofore, or perhaps there was no convenient algorithmic way to provide for it when finding the zeros of the polynomial from the coefficients. Such a constraint may reveal itself in the following way: an experimental system has the property that the observed parameters always seem to be well conditioned functions of the controllable parameters. The mathematical model for the system, however, might lack that well conditioned relation of output to input. Should we add something to the model? We could add a constraint in favor of some multiplicity structure, e.g. one double zero, that is inspired by a feature of the physical system. For instance a symmetry in the experimental system might correspond to a double zero in the polynomial.

Constraints upon the form of the solution should not be imposed merely to obtain a well conditioned solution. Not all experimental

systems are well conditioned, and not all problems should have well conditioned solutions. Suppressing annoying numerical properties may be equivalent to ignoring the most important and interesting features of the system. It may be that the observed ill condition corresponds to an important feature of the problem that is not properly reflected in our theory. In other cases ill condition may mean that the problem we seek to solve is so close to being ill posed that it is senseless to try to solve the problem in the presence of error.

Example. Figure I.3 is an example of a physical system. It is the well known damped harmonic oscillator discussed in elementary physics courses; see, e.g., Kibble [20]. A mass m may travel up and down. It is attached through a spring to the roof; the other end is attached to a shock absorber (dashpot). If the mass is moved from its rest position and released it will eventually return to its rest position, because of friction forces in the dashpot. The goal of an engineer might be to design the dashpot so that the mass will return to its rest position as quickly as possible after a perturbation. By adjusting the dashpot, the mass may be caused to return to its rest position as rapidly as possible without oscillation. The system is then said to be critically damped. The engineer may decide that the spring force on m is $-kx$ for a $k > 0$ which can be measured to perhaps three significant figures. An investigation of the friction forces of the fluid in the dashpot might confirm that the friction forces on m can be approximated by $-d\dot{x}$ for a constant $d > 0$, which can again be measured to a few figures. Finally the mass itself can be measured.

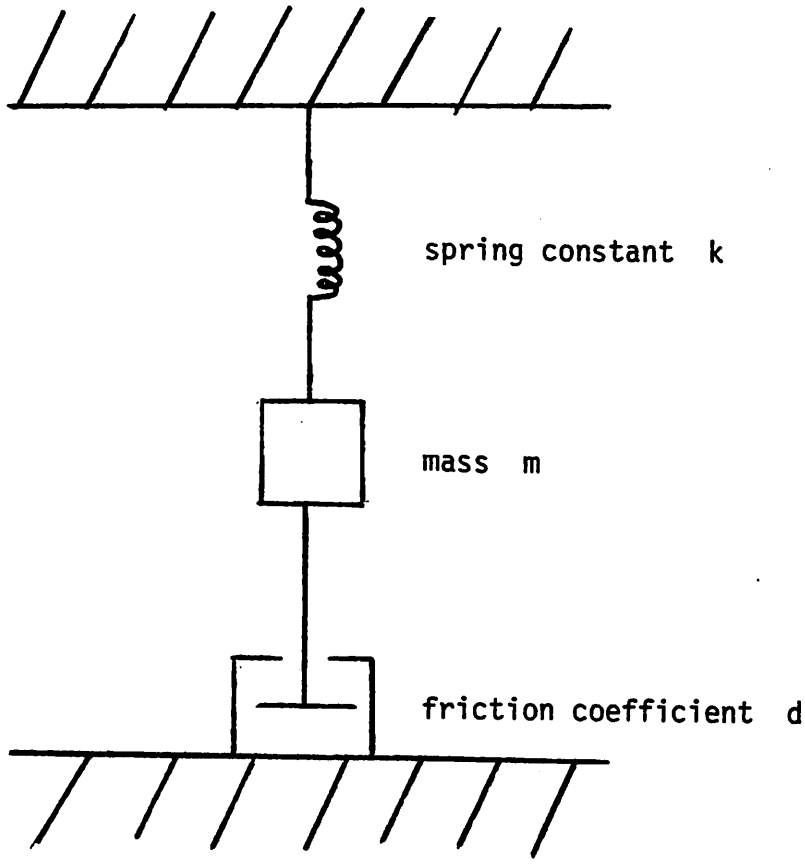

Figure I.3. A damped harmonic oscillator.

Then the mathematical model corresponding to the stated physical assumptions is that the restoring force on m is $-kx - d\dot{x}$ so

$$m\ddot{x} + d\dot{x} + kx = 0 ,$$

and $x(0) = x_0$ and $\dot{x}(0) = v_0$ are the initial conditions. The solutions to such linear ordinary differential equations with constant coefficients are usually linear combinations of exponentials $e^{c_+ t}$ and $e^{c_- t}$ where c_+ and c_- are the zeros c of the quadratic polynomial

$$mc^2 + dc + k \quad .$$

If $c_+ = c_-$ then the solutions are linear combinations of $e^{c_+ t}$ and $te^{c_+ t}$. The quantity to be minimized is the maximum time constant for the components of the solution. The time constant for e^{ct} is defined to be $-1/\mathrm{Re}\ c$ which corresponds to the non-oscillatory, decaying part of the motion of m. (The oscillatory part is governed by Im c.) Then

$$\max\left(\frac{-1}{\mathrm{Re}\ c_+}, \frac{-1}{\mathrm{Re}\ c_-}\right) = \begin{cases} \dfrac{2m}{d - \sqrt{d^2 - 4mk}} & \text{for } d \geq \sqrt{4mk} , \\ \dfrac{2m}{d} & \text{for } 0 \leq d \leq \sqrt{4mk} . \end{cases}$$

For $d \geq 0$ this is minimized by letting $d^2 = 4mk$. In that case $c_+ = c_-$.

Given m and k the engineer can compute an optimal d which he can obtain approximately by adjusting the dashpot.

The engineer may then mass produce these assemblies. Of course there will be variations within tolerances in m, k, and d. Some of

the assemblies will probably exhibit oscillatory motions when perturbed. Then the question will arise: are these variations from unit to unit due to the normal variation of components within tolerances, or is there an error in the design, or in the claimed tolerances?

We can resolve this question by asking: given the polynomial corresponding to one of the production units.

$$p(c) = c^2 + (\frac{d}{m})c + (\frac{k}{m}) ,$$

is the nearest polynomial with a double zero within the distance allowed by the tolerances on $(\frac{d}{m})$ and $(\frac{k}{m})$? If Δ_d is the tolerance on $(\frac{d}{m})$ and Δ_k the tolerance on $(\frac{k}{m})$ then we might measure perturbations

$$q(c) = \alpha c + \beta$$

by

$$\|q\|^2 = (\frac{\alpha}{\Delta_d})^2 + (\frac{\beta}{\Delta_k})^2 .$$

Then if the distance to the nearest polynomial with a double zero were less than $\sqrt{2}$ in this norm, the components would likely be within tolerance.

Suppose we have adjusted the assembly to be critically damped. Then we may carefully measure m, k, and d. If we wanted to compute the time constant from the data and the model, we would be wise to incorporate a constraint in favor of double zeros in our polynomial solver, for that constraint corresponds to a fact we know about the physical system.

In contrast, if we carefully measured m, k, and d on an (unadjusted) assembly from the production line, and we wished to

compute the time constant, it would be folly to incorporate a constraint for a double zero in the polynomial solver. If we did we would always think that the assembly was critically damped.

Even when the assembly is at or near critical damping, where small changes in m, d, or k produce large changes in c_+ or c_-, such small changes produce only small changes in the solution of the differential equation, measured in an appropriate norm. That is, an important feature of the physical system is well conditioned. We encounter ill conditioning numerically because we choose to think of the solution of the equation as a sum of exponentials. As a consequence of this point of view we then solve a polynomial equation to find the time constants of the exponentials. Solving the polynomial equation is the step that may be ill conditioned.

Similar mechanical problems are used as examples in the text of Carnahan, Luther, and Wilkes [4, exercises 4.23-4.26 and example 3.1]. There the natural circular vibrational frequencies of mechanical systems with several components are computed. These frequencies are obtained from eigenvalues of symmetric matrices. Multiple eigenvalues merely mean that two different modes of circular vibration happen to have the same frequency because of chance or some physical symmetry. Viewed as an eigenvalue problem, eigenvalues of symmetric matrices are always well conditioned [5]. An inappropriate reformulation of an eigenvalue problem as a polynomial problem is responsible for the ill conditioned zeros Carnahan et al obtain in some of the numerical results given in their example 3.1.

7. What Do We Do With the Explanation?

Once the nearest polynomial has been found which "explains" some ill conditioned problem, what should be done next?

If we just substitute the zeros of the ameliorated or regularized polynomial for the zeros of the original polynomial, we may be guilty of covering up important features of the problem.

One way to investigate those features is to answer the following question: How do the zeros of the polynomial vary when the coefficients of the polynomial vary within their respective uncertainties? When all zeros are well conditioned this question is easily answered by expressing changes in the zeros as a Taylor series in changes in the polynomial, of which only the first term or two are needed because the series converges quickly.

In the interesting case, however, we find that a conventional Taylor series approach will not work for ill conditioned zeros. The radius of convergence of the series never exceeds the distance to the nearest polynomial with a multiple zero. If we actually move to that nearest polynomial, we then find that conventional fractional power series expansion methods still tend to founder because of short radii of convergence.

In chapter VII these problems are discussed and a method is proposed for obtaining expansions for changes in zeros that converge in a much larger region than conventional techniques. The proposed method depends on using the nearest well conditioned polynomial as a starting point for an expansion in two phases. The first phase retains the multiplicity structure of the starting point while the second phase continues in a conventional manner. Thus the symbolic

determination of a series expansion depends on numerical means for determining the most suitable starting point. Most of the difficulty of the problem is in the numerical part. Analytical difficulties preclude getting the actual expansions, but the idea may be used in a very practical way to get bounds for the changes in the zeros as the coefficients vary throughout the entire region of interest. Smith [42] explains how Gerschgorin circles may also be exploited to obtain similar bounds.

8. Survey of Previous Results

Prior to the computer era relatively little attention was devoted to the problem of ill conditioned simple zeros beyond recognizing that small perturbations tended to break up multiple zeros into ill conditioned simple zeros. Thus the multiple zeros themselves were usually unfairly considered to be ill conditioned. The behavior of multiple zeros under perturbation has long been a matter of interest to analysts and algebraists; the fractional power series discussed in chapter VII have been known since the eighteenth century.

Another facet of multiple zeros is their effect on convergence of zero finding algorithms. It has long been known, for instance, that the convergence of Newton's method is only linear in the vicinity of a multiple zero. Consequently much effort has been expended in developing zero finding iterations that perform better near multiple zeros. Such methods have been discussed by Traub [33] and Ostrowski [25], among others; Stewart's is a recent example of such work [32].

James Daniel [7] has recently studied the problem of improving approximations to multiple zeros. He suggests that averages of clustered ill conditioned simple zeros may be taken to determine the multiple zero of which they are apparently approximations. The examples he cites show that his suggestion may sometimes be helpful for double zeros and perhaps for higher multiplicities if accuracy requirements are not very stringent. Daniel's work has not been incorporated in any widely available codes for polynomial zeros. The reason may be that a conventional zero-finding code with deflation would, in the vicinity of an m-tuple zero, find first an ill conditioned member of an m-member cluster. Then it would find an ill conditioned member of

an m-1-member cluster caused by perturbing an m-1-tuple zero which is not the same as the m-tuple zero of the original problem. Then the m ill conditioned zeros that are averaged together at the end are not all perturbations of the same multiple zero and consequently this average does not make a very good estimate of any multiple zero.

To J. Wilkinson [34] must go credit for publicizing the fact that ill condition and apparent clustering are not equivalent characteristics of zeros of polynomials. This fact does not seem to be explicitly recognized previous to Wilkinson's work. The polynomials he chose as examples are still being studied profitably as in chapter X of the present work.

Wilkinson also brought to the attention of many readers the facts that condition could not only be rigorously defined but could be measured as well.

In 1975 Dunaway [8] proposed a different method for dealing with polynomials with multiple zeros. Her work is based on the fact that the greatest common divisor (GCD) of such a polynomial and its derivative is a polynomial whose factors are the multiple zeros of the original polynomial, but of multiplicity one less. GCD algorithms have long been used for studying polynomials whose coefficients are exactly known. Recent work by Collins [5] and others has been in the context of symbolic algebra systems employing exact rational arithmetic.

Dunaway's idea was to implement a traditional GCD algorithm in standard finite precision floating point arithmetic. There the key problem is determining when a term in a polynomial remainder sequence may be considered to vanish, indicating that an approximate GCD has been found. As Dunaway remarks, that is a difficult problem in finite

precision arithmetic. She does not give details as to how she resolved it, and it is not clear that her procedure could be automated. If that were possible, it might be an attractive method for investigating the multiplicity structure of the zeros of polynomials without specifying that structure in advance. In contrast, the methods to be presented in subsequent chapters require that one specific structure be investigated at a time -- one double zero, a triple zero, two double zeros, etc.

The present investigation is based on the work of W. Kahan described in [17]. Kahan displayed the connection between ill condition and nearness to the manifold of polynomials with multiple zeros. In [17] and also in [19] he determined how to compute condition numbers and how to derive the equations to be solved for the nearest polynomial with a double or triple zero. He also perceived that the manifolds could be exploited to provide a better way to express perturbed zeros as an expansion in terms of the perturbation.

Kahan went as far as theory unaided by extensive computational experience could be expected to go; this dissertation supplies some of that computational experience and some of the theoretical extensions motivated by that experience.

9. Summary of Findings

The principal original results of this research are:

1) A new method for computing condition numbers for zeros of polynomials, valid for certain norms only, is presented in chapter II.

2) The equations to be solved for the nearest polynomial with two complex conjugate double zeros, two double zeros, and three or more double zeros are presented in chapters IV and V.

3) When k complex multiple zeros are sought, the equations that need to be solved are less complicated than might have been thought at first. It is shown that k Lagrange multipliers may be assumed to vanish for any interesting solutions. This result, previously known [19] for the case of a single multiple zero, has been extended to the case of several multiple zeros and the case of a complex conjugate pair of multiple zeros in chapters IV and V. But a counterexample has been discovered which indicates that, in the most common case of a real polynomial subject only to real perturbations, these results are not always applicable.

4) Some results on the location of the nearest polynomial with a double zero are given in chapter VI.

5) The details of a new technique for bounding changes in the zeros of a polynomial are presented in chapter VII. This technique, originally suggested by W. Kahan, exploits nearby manifolds of polynomials with multiple zeros whereas conventional techniques are usually hindered by the presence of those same manifolds.

6) Extensive computer codes of methods presented in earlier chapters were prepared to test the theory experimentally. In chapter IX examples are given of successful application of these codes.

7) Extensive computer results are given in chapter X to support the conclusion that one polynomial mentioned by Wilkinson [34] is intrinsically not amenable to treatment of the type proposed in the previous section, due to its position near a particularly complicated part of the manifold of polynomials with double zeros.

10. Notation

In the following chapters we will consider perturbations of monic algebraic polynomials p, of degree n, with real or complex coefficients:

$$p(\tau) = \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j} .$$

We will usually follow the conventions of using lower case Greek letters for scalars, lower case Romen letters other than i through n for vectors and polynomials, and capital Roman letters for matrices, non-linear operators on vectors, and sometimes for functions. But p_j and A_{ij} will usually represent scalar elements of p and A. R^n and C^n represent the real and complex vector spaces of dimension n.

The perturbations will be polynomials of degree at most $n-1$, not usually monic:

$$q(\tau) = \sum_{j=1}^{n} q_j \tau^{n-j} .$$

We identify the space of perturbations q of a polynomial p with a vector space of dimension n and, in the obvious basis

$$\{\tau^{n-1}, \tau^{n-2}, \ldots, \tau, 1\} ,$$

the elements of the vectors are the coefficients of the polynomials:

$$q = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix} \sim q(\tau) = \sum_{j=1}^{n} q_j \tau^{n-j} .$$

Any norm for R^n or C^n may now be imposed. We will be interested in a weighted ℓ_2 norm on C^n defined by

$$\|q\|_W = (q^*Wq)^{1/2} = \|W^{1/2}q\|_2 ,$$

where q^* denotes conjugate transpose and W is Hermitian positive definite and usually diagonal as well. In the diagonal case we write

$$\|q\|_W = \sqrt{\sum_{j=1}^{n} w_j |q_j|^2} .$$

There is a dual space of functionals u^* which has the usual norm

$$\|u^*\|_W = \sup_q \frac{|u^*q|}{\|q\|_W} = (u^*W^{-1}u)^{1/2} ,$$

or

$$\sqrt{\sum_{j=1}^{n} |u_j|^2/w_j}$$

in the diagonal case. Most often the functional we are interested in is e_ζ^*, the functional that evaluates a polynomial q at ζ: $e_\zeta^* q = q(\zeta)$. In our basis $e_\zeta^* = (\zeta^{n-1}\zeta^{n-2}\ldots\zeta\ 1)$.

One frequently used operator is the derivative operator D which maps C^n to C^n and has the matrix form

$$D = \underbrace{\begin{bmatrix} 0 & & & \\ n-1 & 0 & & O \\ & n-2 & 0 & \\ & & \ddots & \ddots \\ O & & & 0 \\ & & & 1 \quad 0 \end{bmatrix}}_{n} \Bigg\}n$$

We can for instance write $e_\zeta^* D^k$ for the functional which evaluates the k'th derivative of a polynomial at ζ. In fact we will often be interested in the operator which computes a polynomial and its first

$m-1$ derivatives at ζ. We will define it as

$$A_\zeta = \underbrace{\left.\begin{bmatrix} e_\zeta^* \\ e_\zeta^* D \\ \vdots \\ e_\zeta^* D^{m-1} \end{bmatrix}\right\} m}_{n}$$

so

$$A_\zeta q = \begin{bmatrix} q(\zeta) \\ \vdots \\ q^{(m-1)}(\zeta) \end{bmatrix} .$$

Corresponding operators $\tilde{D}$ and $\tilde{A}$ can be defined for polynomials of degree n; their matrices operate on vectors of dimension $n+1$. Then

$$\tilde{A}_\zeta p = \begin{bmatrix} p(\zeta) \\ \vdots \\ p^{(m-1)}(\zeta) \end{bmatrix} .$$

$\tilde{A}_\zeta$ is m by $(n+1)$.

It is handy to note here that the m rows of A_ζ are independent for $m \leq n$. For if we apply A_ζ to the vector q representing $(\tau-\zeta)^k$ we find

$$A_\zeta q = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ k! \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{position } k+1 .$$

By letting k run from 0 to $m-1$ we find that the rank of A_ζ is indeed m.

Frequently we will be using ζ as a symbol for a multiple zero of a nearby polynomial and α will be a symbol for a zero of the original polynomial. We will write e^* for e_ζ^* and A for A_ζ. In chapters II and VII, however, A will be an m-1 by n matrix

$$A \equiv \begin{pmatrix} e_\alpha^* \\ \vdots \\ e_\alpha^* D^{m-2} \end{pmatrix} .$$

Those chapters also use the n by n-m+1 matrix

$$P_{m-1} = \begin{pmatrix} 1 & & 0 \\ (m-1)(-\alpha) & \ddots & \\ \vdots & \ddots & 1 \\ \vdots & & (m-1)(-\alpha) \\ (-\alpha)^{m-1} & & \vdots \\ & \ddots & \\ 0 & & (-\alpha)^{m-1} \end{pmatrix}$$

Multiplying an n-m+1 vector q by P_{m-1} corresponds to multiplying a polynomial of degree $n-m$, $q(\tau)$, by $(\tau-\alpha)^{m-1}$. The columns of P_{m-1} are linearly independent since $(\tau-\alpha)^{m-1}q(\tau) \neq 0$ if $q \neq 0$.

When presenting numerical results we will often use FORTRAN E-format, e.g.

.123E-5 means $.123 \times 10^{-5}$.

CHAPTER II

COMPUTING CONDITION NUMBERS FOR ZEROS OF POLYNOMIALS

1. Definition of Condition Numbers for Simple Zeros

In this chapter we explain several ways to compute condition numbers for zeros of polynomials. In the last section we see why ill condition is always associated with nearness to a polynomial with one or more double zeros.

Condition numbers are intended to be a numerical measurement of condition. They tell us how large a change in the solution may result from a given change in the data. In general, for a problem which converts m input data items d_i into n components of a solution s_j, there could be nm condition numbers $\gamma_{ij} \equiv |\Gamma_{ij}|$, $\Gamma_{ij} \equiv \frac{\partial s_j}{\partial d_i}$, and the condition of the problem could be defined to be a norm of the matrix of Γ_{ij}. If there is a norm $\|\cdot\|_S$ defined on the solution and a norm $\|\cdot\|_D$ defined on the data, then the most suitable norm for Γ, the matrix of Γ_{ij}, is

$$\|\Gamma\| \equiv \sup_d \left(\frac{\|\Gamma d\|_S}{\|d\|_D}\right) .$$

One could just as well consider relative condition numbers,

$$\tilde{\gamma}_{ij} = \left|\frac{d_i}{s_j}\right| \cdot \left|\frac{\partial s_j}{\partial d_i}\right|$$

as long as $s_j \neq 0$.

For our purposes we will generally consider a separate condition number for each zero of a polynomial but we will lump together changes in the coefficients and measure the combined change by means of a norm.

Let p be a monic polynomial of degree n,

$$p(\tau) = \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j} ,$$

and let δp be a perturbing polynomial of degree $n-1$, not necessarily monic, representing a change in the coefficients:

$$\delta p(\tau) = \sum_{j=1}^{n} \delta p_j \tau^{n-j} .$$

Let α be a zero of $p(\tau)$ and $\alpha + \delta\alpha$ a zero of $p(\tau) + \delta p(\tau)$.

Definition. The (absolute) condition number, γ, of α with respect to changes δp is

$$\gamma \equiv \lim_{\Delta \to 0} \left(\sup_{\substack{\delta p \text{ with} \\ \|\delta p\| = \Delta}} \frac{|\delta\alpha|}{\|\delta p\|} \right) . \tag{1.1}$$

As we have seen, this limit is infinite for multiple zeros α, a defect which we shall remedy shortly.

There is one aspect of ill condition of zeros of polynomials that may surprise those accustomed to thinking of ill condition primarily in terms of systems of linear equations. In that context norms are usually chosen in such a way that the condition number of a matrix with respect to inversion is never less than 1. There is no such natural choice of norms for zeros of polynomials and their condition numbers may take on any positive value. We shall see in chapters IX and X that well conditioned zeros can be very well conditioned indeed:

in a certain reasonable norm, the condition number of one of the zeros of Wilkinson's polynomial is about 1.E-16.

Our definition of condition and condition number is similar to that of Wilkinson [34], and is also a special case of a more general formulation proposed by Rice [27]. Both Rice and Wilkinson also propose relative condition numbers which we would define as

$$\gamma_{rel} \equiv \frac{\gamma}{|\alpha|}$$

for $\alpha \neq 0$. In this case we would choose a norm for δp which would measure relative changes in the coefficients. An example is

$$\|\delta p\| = \left(\sum_{j=1}^{n} \left| \frac{\delta p_j}{p_j} \right|^2 \right)^{1/2}$$

if all $p_j \neq 0$. Other norms can be devised suitable for the case when some p_j is zero. It is the responsibility of the definer of a problem to decide the appropriate norm. For instance, if none of the zeros of p are 0, then the polynomial $\tilde{p}(\tau)$, whose positive zeros are the moduli of the zeros of p, may be used to define a norm:

$$\|\delta p\| = \left(\sum_{j=1}^{n} \left| \frac{\delta p_j}{\tilde{p}_j} \right|^2 \right)^{1/2} .$$

None of the $\tilde{p}_j$ are 0 as long as $p_n \neq 0$.

2. Definition of Condition Numbers for Multiple Zeros

The previous discussion shows that our definition of condition number does not make sense for a multiple zero, which would apparently have an infinite condition number. That infinite condition is caused by the fact that most arbitrary infinitesimal perturbations applied to a polynomial with a multiple zero tend to break up that multiple zero into ill conditioned simple zeros.

In order to have a sensible definition of condition number for a multiple zero we must only allow perturbations which do not destroy the multiple zero. Here is an example: consider a real monic cubic polynomial,

$$p(\tau) = (\tau-\alpha)^2(\tau-\beta) = \tau^3 - (2\alpha+\beta)\tau^2 + (2\alpha\beta+\alpha^2)\tau - \alpha^2\beta \quad ,$$

and small quadratic perturbations,

$$q(\tau) = q_1\tau^2 + q_2\tau + q_3 \quad ,$$

which preserve the multiplicity of α so that

$$p(\tau) + q(\tau) = (\tau - (\alpha+\varepsilon))^2(\tau - (\beta+\theta)) \quad .$$

We discover that

$$\begin{aligned} q_1 &= 2\varepsilon + \theta \quad , \\ q_2 &= 2\alpha\varepsilon + 2\beta\varepsilon + 2\alpha\theta + (2\varepsilon\theta + \varepsilon^2) \quad , \\ q_3 &= 2\alpha\beta\varepsilon + \alpha^2\theta + (2\alpha\varepsilon\theta + \beta\varepsilon^2 + \varepsilon^2\theta) \quad , \end{aligned}$$

where the parentheses segregate higher order terms which we shall ignore. Thus the three parameters q_i are defined in terms of the two variables ε and θ. We can choose any two of the q_i as the

independent parameters of the perturbation and solve for ε in terms of them. Thus if we choose q_1 and q_2, we find

$$\varepsilon = (q_2 - \alpha q_1)/(2(\beta-\alpha))$$

and

$$\theta = (\beta q_1 - q_2)/(\beta-\alpha)$$

to first order in ε and θ.

Then we can see that the ratio of change in solution (ε) to change in data (q_1) is

$$\frac{\varepsilon}{q_1} = \frac{q_2/q_1 - \alpha}{2(\beta-\alpha)}$$

which will be well defined unless $\beta = \alpha$, which would mean that the multiplicity of α was not two, as we thought, but actually three.

In general let

$$p(\tau) = (\tau-\alpha)^m q(\tau) , \qquad q(\alpha) \neq 0 .$$

Definition. The condition number of α is

(2.1)
$$\gamma \equiv \lim_{\Delta\to 0} \left[\sup_{\substack{\text{over } \delta p \text{ maintaining} \\ \text{multiplicity of } \alpha \\ \text{with } \|\delta p\| = \Delta}} \frac{|\delta\alpha|}{\|\delta p\|} \right] .$$

In order to appreciate graphically what is meant by constraining perturbations to maintain multiplicity, consider the drawings in Figures II.1-II.3 of the space of monic real cubic polynomials. That space is three dimensional and the set of small perturbations about a point in that space is a closed ball. The drawings are based on a norm in which closed balls look like spheres; see Figure II.1.

The set of monic real cubic polynomials with double zeros is a two dimensional algebraic surface (manifold). The set of small perturbations maintaining multiplicity of a double zero is the intersection of the ball and that manifold. If the manifold were a plane that set might be an oval. In general that set resembles a bent coin or an ellipse warped into three dimensions; see Figure II.2.

The double zero is well behaved in the face of perturbations that keep the polynomial on the manifold but away from the one dimensional submanifold of real cubic polynomials with a triple zero. That submanifold is an algebraic curve and a subset of the surface mentioned previously. The set of small perturbations maintaining a triple zero is the intersection of the ball and that curve -- amounting to a segment of the curve, as in Figure II.3.

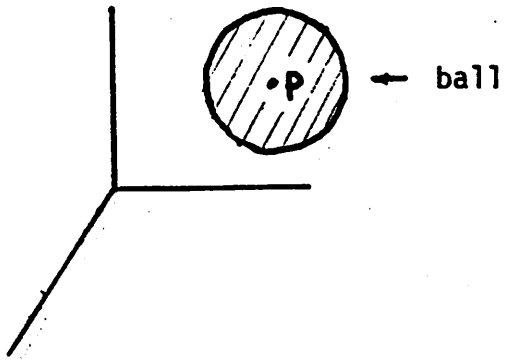

Figure II.1. A small ball about p in R^3 containing perturbations δp such that $\|\delta p\| \leq \Delta$.

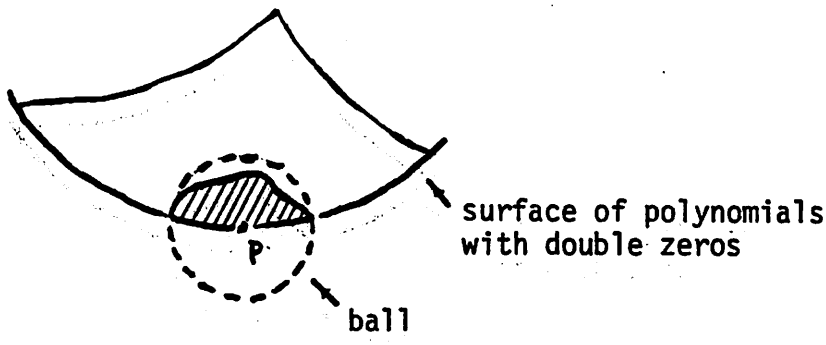

Figure II.2. The set of small perturbations about p maintaining a double zero resembles a bent coin.

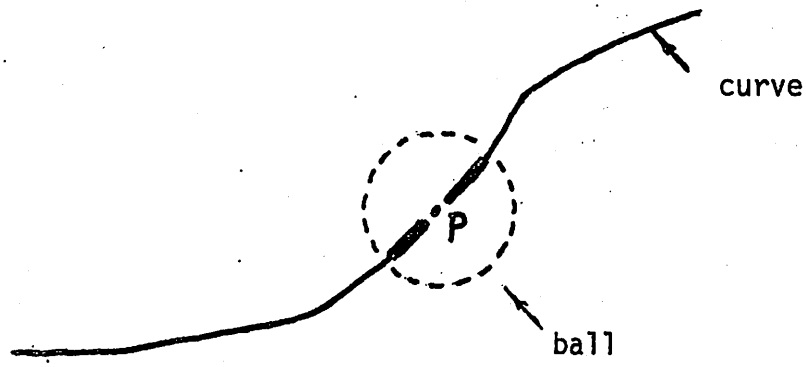

Figure II.3. The set of small perturbations about p maintaining a triple zero is a segment of a curve.

3. Condition Numbers for n-tuple Zeros

As a start we derive a condition number for the simplest case, that of a single n-tuple zero. When the polynomial has the form

$$p(\tau) = (\tau-\alpha)^n = \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j} ,$$

where

$$p_j = \binom{n}{j}(-\alpha)^j , \qquad \binom{n}{j} \equiv \frac{n!}{(n-j)!j!} ,$$

then δp has the form

$$\begin{aligned} \delta p(\tau) &= (\tau - (\alpha+\delta\alpha))^n - (\tau-\alpha)^n \\ &= \sum_{j=1}^{n} \binom{n}{j} \{(-\alpha-\delta\alpha)^j - (-\alpha)^j\} \tau^{n-j} \\ &= (-\delta\alpha) \sum_{j=1}^{n} \binom{n}{j} \cdot j \cdot (-\alpha)^{j-1} \tau^{n-j} \quad \text{to first order} . \end{aligned}$$

Then, recognizing an expression for $(\tau-\alpha)^{n-1}$,

$$\gamma = \sup_{\delta p} \frac{|\delta\alpha|}{\|\delta p\|} = \frac{1}{n\|(\tau-\alpha)^{n-1}\|} = \frac{|\alpha|}{\left\| \sum_{j=1}^{n} j p_j \tau^{n-j} \right\|} .$$

In particular for the diagonal W norms

$$\gamma = \frac{|\alpha|}{\sqrt{\sum_{j=1}^{n} w_j \cdot j^2 \cdot |p_j|^2}} ,$$

except if $\alpha = 0$,

$$\gamma = \frac{1}{n\sqrt{w_1}} .$$

4. Resolution of Condition Number into Components

We show now that the condition number we have defined is a product of two independent factors. Thus for the polynomial

$$p(\tau) = (\tau-\alpha)^m \prod_{j=m+1}^{n} (\tau-\zeta_j)$$

the condition number for α will be shown to be

$$\gamma = \frac{\frac{1}{m}\sigma}{\prod_{j=m+1}^{n} |\alpha-\zeta_j|}$$

where the numerator σ/m will depend on the zero α but not on the other zeros ζ_j. The denominator depends on the other zeros ζ_j but not on m nor on the norm. We require that $\alpha \neq \zeta_j$ so that m is indeed the true multiplicity of α.

W. Kahan demonstrated this fact in [17] after showing that, for a monic polynomial of degree n, an m-tuple zero may be regarded as an analytic function of the first $n+1-m$ coefficients of that polynomial. This may be compared to the well known result that a simple zero is an analytic function of the n coefficients of a monic polynomial. In both cases analyticity is confined to regions in which the zero does not increase or decrease in multiplicity.

We shall infer the resolution of the condition number directly, however. Let

$$p(\tau) = (\tau-\alpha)^m q(\tau)$$

and let δp represent infinitesimal variations in p such that $p+\delta p$ has a multiple zero $\alpha+\delta\alpha$ of multiplicity m. Then

$$(p+\delta p)(\tau) = (\tau - (\alpha+\delta\alpha))^m \cdot (q+\delta q)(\tau) \ ,$$

and in consequence, keeping only first order terms, we find

$$\delta p(\tau) = (\tau-\alpha)^{m-1}\{(\tau-\alpha)\delta q(\tau) - mq(\tau)\delta\alpha\} \ .$$

Thus δp is displayed as a function of δq and $\delta\alpha$.

We claim

$$\gamma = \sup_{\left(\begin{smallmatrix}\text{constrained}\\ \delta p\end{smallmatrix}\right)} \frac{|\delta\alpha|}{\|\delta p\|} = \frac{1}{m}\frac{1}{|q(\alpha)|} \sup_{\left(\begin{smallmatrix}r \text{ of degree}\\ \leq n-m\end{smallmatrix}\right)} \frac{|r(\alpha)|}{\|(\tau-\alpha)^{m-1}\cdot r(\tau)\|} \ ,$$

and we prove it by showing the one-to-one correspondence between such r and such δp. Namely let

$$r(\tau) = (\tau-\alpha)\delta q(\tau) - mq(\tau)\delta\alpha$$

so

$$\delta p(\tau) = (\tau-\alpha)^{m-1} r(\tau) \ .$$

Since δp has degree $\leq n-1$, r has degree $\leq n-m$. The <u>dimension</u> of the <u>vector</u> r is $n-m+1$, however, since the polynomial $r(\tau)$ is not monic.

Any such r defines δp and hence δq uniquely:

$$\delta\alpha = \frac{-r(\alpha)}{mq(\alpha)} \ , \qquad \delta q(\tau) = \frac{r(\tau) + mq(\tau)\delta\alpha}{\tau - \alpha} \ .$$

The numerator of the expression for $\delta q(\tau)$ does vanish when $\tau = \alpha$ so that expression is indeed a polynomial rather than a rational function. Therefore we may write

$$\frac{|\delta\alpha|}{\|\delta p\|} = \frac{1}{m|q(\alpha)|} \cdot \frac{|r(\alpha)|}{\|(\tau-\alpha)^{m-1} r(\tau)\|}$$

and, since $q(\tau) = \prod_{j=m+1}^{n} (\tau-\zeta_j)$, then $|q(\alpha)| = \prod_{j=m+1}^{n} |\alpha-\zeta_j|$.

As claimed, then, we may write the condition number for α as

$$\gamma = \frac{\frac{1}{m}\sigma}{\prod_{j=m+1}^{n} |\alpha-\zeta_j|}, \tag{4.1}$$

and

$$\frac{1}{m}\sigma = \frac{1}{m} \sup_{\left(\begin{array}{c}\text{degree } r \\ \leq n-m\end{array}\right)} \frac{|r(\alpha)|}{\|(\tau-\alpha)^{m-1} r(\tau)\|} \tag{4.2}$$

is the part of the condition number that is independent of the other zeros ζ_j. The next few sections will be devoted to explaining how to compute σ.

5. Computing σ for Arbitrary Norms -- Dual Method

W. Kahan [19] has provided the following method for computing σ in arbitrary norms. We shall see that it leads to solving a standard kind of linear approximation problem, namely

$$\sigma = \min_{\ell^*} \|s^* + \ell^* A\|/(m-1)! \ ,$$

for vectors s^* and ℓ^* and a matrix A to be defined.

To prove the statement above, write the formula for σ as

$$\sigma = \sup_{\left(\begin{smallmatrix} r \text{ of degree} \\ \leq n-m \end{smallmatrix}\right)} \frac{|e_\alpha^* r|}{\|P_{m-1} r\|} = \sup_{\left(\begin{smallmatrix} y \text{ of degree} \\ \leq n-1 \end{smallmatrix}\right)} \frac{|e_\alpha^* Z S y|}{\|P_{m-1} S y\|}$$

where $y \in C^n$ and S is a map from C^n onto C^{n-m+1}. Z is the operator which fills out $n-m+1$-vectors with zeros to form n-vectors:

$$Z = \underbrace{\begin{pmatrix} & 0 & \\ \hline 1 & & \\ & \ddots & 0 \\ & & \\ 0 & & 1 \end{pmatrix}}_{n-m+1} \Bigg\} n$$

ZS is required to be a projector. Finally P_{m-1} is the linear operator from C^{n-m+1} to C^n mentioned in chapter I which represents multiplication by $(\tau-\alpha)^{m-1}$.

Our goal is to transform the sup problem into a dual min problem. We therefore state a duality theorem of Buck [3]. The setting for the theorem is a normed vector space E with its dual space of functionals E^*. If M is a subspace in E and $M^\perp$ its annihilator in E^*,

the theorem states

$$\sup_{x \in M} \frac{|v_0^* x|}{\|x\|} = \min_{v^* \in M^\perp} \|v_0^* - v^*\| .$$

For the application at hand, E is C^n. $M = \{P_{m-1}Sy \mid y \in C^n\}$. Then $M^\perp = \{v^* \mid v^* P_{m-1} S = 0\}$. We discover

$$\sup_{y \in C^n} \frac{|v_0^* P_{m-1} Sy|}{\|P_{m-1}Sy\|} = \min_{(u^* P_{m-1} S = v_0^* P_{m-1} S)} \|u^*\| .$$

Then if there is a v_0^* such that $v_0^* P_{m-1} S = e_\alpha^* ZS$ we will have the sup we seek, expressed as a min.

Since the columns of P_{m-1} are linearly independent, the range space of P_{m-1}^* must have full dimension so the equation $Z^* e_\alpha = P_{m-1}^* v_0$ may be solved for v_0. Therefore

$$\sigma = \min_{(u^* P_{m-1} S = e_\alpha^* ZS)} \|u^*\| . \tag{5.1}$$

Let us see what the solutions of $u^* P_{m-1} S = e_\alpha^* ZS$ are; among them we will find that of minimal norm. As in chapter I let D^k denote the operator which maps polynomials to their k'th derivatives. Then we find that

$$u^* = e_\alpha^* D^{m-1}/(m-1)!$$

is one solution of the equation. For consider any $y(\tau)$ and let $r(\tau)$ be its image; $r = Sy$. Then

$$\begin{aligned} e_\alpha^* D^{m-1} P_{m-1} r &= \{(\tau-\alpha)^{m-1} r(\tau)\}^{(m-1)}(\alpha) \\ &= (m-1)! r(\alpha) = (m-1)! e_\alpha^* Zr . \end{aligned}$$

The next step is to determine the solutions of the homogeneous equation $u^*P_{m-1}S = 0$. The rank of S is $n-m+1$, as is the rank of P_{m-1}, and therefore their product. Since u^* has dimension n, the null space of $(P_{m-1}S)^*$ must have dimension $m-1$. Therefore we seek a subspace of solutions u^* of dimension $m-1$.

We may easily verify that $\{e_\alpha^*, e_\alpha^*D, \ldots, e_\alpha^*D^{m-2}\}$ is a set of solutions to $u^*P_{m-1}S = 0$, because $e_\alpha^*D^kP_{m-1}r = \{(\tau-\alpha)^{m-1}r(\tau)\}^{(k)}(\alpha) = 0$ for $0 \le k \le m-2$. These $m-1$ linearly independent solutions therefore form a basis for the solution space and we may insert the general solution of the inhomogeneous equation in the formula (5.1) to get

$$\sigma = \frac{1}{(m-1)!} \min_{\lambda_k} \left\| e_\alpha^*D^{m-1} + \sum_{k=0}^{m-2} \lambda_k e_\alpha^*D^k \right\| . \tag{5.2}$$

If we write the $m-1$ vector $\ell^* = (\lambda_0, \lambda_1, \ldots, \lambda_{m-2})$, the $m-1$ by n matrix

$$A = \begin{pmatrix} e_\alpha^* \\ e_\alpha^*D \\ \vdots \\ e_\alpha^*D^{m-2} \end{pmatrix} ,$$

and the vector $s^* = e_\alpha^*D^{m-1}$, we have

$$\sigma = \min_{\ell^*} \|s^* + \ell^*A\|/(m-1)! . \tag{5.3}$$

Consequently σ may be found by solving the indicated linear approximation problem, as claimed.

In the special case $\alpha = 0$ we find

$$\sigma = \|(0 \cdots 0\ \underset{\substack{\uparrow \\ n-m+1 \text{ position}}}{1}\ 0 \cdots 0)\| \ .$$

6. Computing σ for ℓ_2 Norms -- Dual Method

We now evaluate σ for ℓ_2 norms. First we note that $\|u^*\|_W = \|u^*W^{-1/2}\|_2$ and, using the theory of least squares, the minimal residual may be expressed as

$$\sigma = \min_{\ell} \|W^{-1/2}s + W^{-1/2}A^*\ell\|_2 / (m-1)!$$
$$= \{s^*(W^{-1} - W^{-1}A^*(AW^{-1}A^*)^{-1}AW^{-1})s\}^{1/2} / (m-1)! \ .$$

In particular, if $m = 1$ then $A = 0$ and $s = e_\alpha$ so

$$\sigma = (e_\alpha^* W^{-1} e_\alpha)^{1/2} = \left(\sum_{j=1}^{n} |\alpha^2|^{n-j}/w_j\right)^{1/2} . \tag{6.1}$$

If $m = 2$, then $A = e_\alpha^*$, $s = D^*e_\alpha$, and after some computation we find

$$\sigma^2 = \frac{1}{|\zeta|^2}\left\{\sum (n-j)^2 |\alpha^2|^{n-j}/w_j - \frac{|\sum (n-j)|\alpha^2|^{n-j}/w_j|^2}{\sum |\alpha^2|^{n-j}/w_j}\right\} ,$$

or in a computationally more economical form,

$$\sigma^2 = \frac{\sum_{j=1}^{n-1} \frac{1}{w_j}|\alpha^2|^{n-j-1}\left\{\sum_{k=j+1}^{n} \frac{1}{w_k}|\alpha^2|^{n-k}(k-j)^2\right\}}{\sum_{j=1}^{n} |\alpha^2|^{n-j}/w_j} .$$

For $m > 2$,

$$\sigma = \frac{1}{(m-1)!} \min_{\lambda_k} \left\{ \sum_{j=1}^{n} \frac{1}{w_j} \left| \frac{(n-j)!}{(n-j-m+1)!}\alpha^{n-j-m+1} + \sum_{k=0}^{m-2} \lambda_k \frac{(n-j)!}{(n-j-k)!}\alpha^{n-j-k} \right|^2 \right\}^{1/2} .$$

This may be written in conventional least squares format as

$$\sigma = \min_{\ell} \|\hat{s}-\hat{A}^*\ell\|_2/(m-1)!$$

where

$$\hat{s}_j = \frac{(n-j)!}{(n-j-m+1)!}\,\alpha^{n-j-m+1}/(w_j)^{1/2}\,,$$

$$\hat{A}^*_{j,k} = \frac{(n-j)!}{(n-j-k)!}\,\alpha^{n-j-k}/(w_j)^{1/2}\,.$$

Finally, if $\alpha = 0$, then $\sigma = 1/(w_{n+1-m})^{1/2}$.

7. Computing σ for ℓ_2 Norms -- Primal Method

In the previous section we computed σ by solving the dual problem. Our goal now is to find σ directly. First convert the expression

$$\sigma = \sup_{\left(r \text{ of degree} \atop \leq n-m\right)} \frac{|r(\alpha)|}{\|(\tau-\alpha)^{m-1} r(\tau)\|}$$

into the vector notation:

$$\sigma = \sup \frac{|e_\alpha^* r|}{\|P_{m-1} r\|} .$$

But if we define a new norm $\|r\|_P \equiv \|P_{m-1} r\|$ then by definition

$$\sigma = \|e_\alpha^*\|_P$$

in the dual norm. Now $\|r\|_P = \|(P_{m-1}^* W P_{m-1})^{1/2} r\|_2$ in our ℓ_2 norm so

$$\sigma = \|e_\alpha^* (P_{m-1}^* W P_{m-1})^{-1/2}\|_2 \tag{7.1}$$

$$= (e_\alpha^* (P_{m-1}^* W P_{m-1})^{-1} e_\alpha)^{1/2} .$$

We can check this result by comparison with the simplest case, $m = 1$. Then $P_0 = I$ and

$$\sigma^2 = e_\alpha^* W^{-1} e_\alpha = \sum_{j=1}^{n} |\alpha^2|^{n-j}/w_j$$

which is just the result obtained in the previous section.

8. Computational Details

We shall see how to compute the non-zero elements of $P_{m-1}{}^{*}WP_{m-1}$. Let P denote a generalized matrix of the P_{m-1} type corresponding to multiplication by a monic polynomial $t(\tau)$ of degree d. For instance, if $m = 3$, P_2 corresponds to $(\tau-\alpha)^2 = \tau^2 - 2\alpha\tau + \alpha^2$. Then $t_0 = 1$, $t_1 = -2\alpha$, and $t_2 = \alpha^2$ are the elements of t. P has the form of an n by n-d matrix

$$\begin{pmatrix} 1 & & 0 \\ t_1 & \ddots & \\ \vdots & \ddots & 1 \\ t_d & & t_1 \\ & \ddots & \vdots \\ 0 & & t_d \end{pmatrix}$$

so

$$P_{ij} = \begin{cases} t_{i-j} & \text{if } j \le i \le j+d \text{ ,} \\ 0 & \text{otherwise .} \end{cases}$$

Then

$$(P^{*}WP)_{ij} = \begin{cases} \sum_{k=\max(i,j)}^{k=\min(i,j)} w_k t_{k-i}{}^{*} t_{k-j} & \text{if } |i-j| \le d \text{ ,} \\ 0 & \text{otherwise ,} \end{cases}$$

so this matrix has bandwidth $2d+1$ in addition to being positive definite Hermitian.

9. Condition Numbers for Complex Conjugate Zeros of Real Polynomials

The formulas derived in the previous sections were valid for complex zeros of a complex polynomial subject to complex perturbations. It is easy to verify that the same formulas apply for real zeros of a real polynomial subject to real perturbations. The case of complex zeros of a real polynomial subject to real perturbations, however, is more complicated. The requirement that the perturbed polynomial remain real amounts to an extra constraint. We now define condition numbers that reflect this constraint. Let

$$p(\tau) = (\tau-\alpha)^m(\tau-\bar{\alpha})^m q(\tau) , \quad q(\alpha) \neq 0 ,$$

represent a real polynomial with a complex m-tuple zero at α and consequently at $\bar{\alpha}$ as well, with $\operatorname{Im} \alpha \neq 0$. Considering infinitesimal perturbations we define

$$(p+\delta p)(\tau) = (\tau - (\alpha+\delta\alpha))^m(\tau - (\bar{\alpha}+\overline{\delta\alpha}))^m(q+\delta q)(\tau)$$

and to first order we find

$$\delta p(\tau) = (\tau-\alpha)^{m-1}(\tau-\bar{\alpha})^{m-1}[(\tau-\alpha)(\tau-\bar{\alpha})\delta q(\tau) - 2mq(\tau)\{(\operatorname{Re}\ \delta\alpha)\tau - \operatorname{Re}(\bar{\alpha}\delta\alpha)\}]$$

Definition. The condition number of α with respect to real perturbations of p is

$$\gamma \equiv \lim_{\Delta\to 0} \sup_{\substack{\text{constrained } \delta p \\ \text{with } \|\delta p\| = \Delta}} \frac{|\delta\alpha|}{\|\delta p\|} . \tag{9.1}$$

Let

$$r(\tau) = (\tau-\alpha)(\tau-\bar{\alpha})\delta q(\tau) - 2mq(\tau)\{(\operatorname{Re}\ \delta\alpha)\tau - \operatorname{Re}(\bar{\alpha}\delta\alpha)\} .$$

Then real δq and complex $\delta\alpha$ define r uniquely. Conversely,

$$\delta\alpha = (\sqrt{-1}\, r(\alpha))/(2m(\mathrm{Im}\ \alpha)q(\alpha))$$

and

$$\delta q(\tau) = \frac{r(\tau) + 2mq(\tau)\{(\mathrm{Re}\ \delta\alpha)\tau - \mathrm{Re}(\bar{\alpha}\delta\alpha)\}}{(\tau-\alpha)(\tau-\bar{\alpha})} .$$

As before we can verify that the expression for δq defines a polynomial rather than a rational function.

Thus there is a one-to-one correspondence between r and $(\delta\alpha,\delta q)$. Substituting in (9.1) we find

$$\gamma = \frac{1}{2m|\mathrm{Im}\ \alpha||q(\alpha)|} \sup_{\substack{r \text{ of degree} \\ \leq n-2m+1}} \frac{|r(\alpha)|}{\|(\tau-\alpha)^{m-1}(\tau-\bar{\alpha})^{m-1}r(\tau)\|}$$

or

$$\gamma = \frac{1}{2m|\mathrm{Im}\ \alpha||q(\alpha)|}\sigma_c . \tag{9.2}$$

Thus in this case as well, the condition number consists of (1) a numerator $\sigma_c/(2m|\mathrm{Im}\ \alpha|)$ independent of the other zeros ζ_j, and (2) a denominator $|q(\alpha)| = \prod_{j=2m+1}^{n} |\alpha-\zeta_j|$.

The limit $\mathrm{Im}\ \alpha \to 0$ corresponds to α and $\bar{\alpha}$ coalescing to form a zero of greater multiplicity $2m$. Therefore the condition number becomes infinite as $\mathrm{Im}\ \alpha \to 0$.

10. Computing σ_c for ℓ_2 Norms

We turn now to the problem of computing σ_c by a method similar to the primal method for computing σ. Define C_{m-1} mapping R^{n-2m+2} into R^n as the operator corresponding to multiplication by $(\tau-\alpha)^{m-1}(\tau-\bar{\alpha})^{m-1}$ for complex α. Then in matrix form, C_1 for instance is n by n-2:

$$C_1 = \begin{pmatrix} 1 & & 0 \\ -2\,\mathrm{Re}\,\alpha & \ddots & \\ |\alpha|^2 & \ddots & 1 \\ & \ddots & -2\,\mathrm{Re}\,\alpha \\ 0 & & |\alpha|^2 \end{pmatrix} .$$

Consequently

$$\sigma_c^2 = \sup_r \frac{|r(\alpha)|^2}{\|C_{m-1}r\|^2} = \sup_r \frac{r^* e_\alpha e_\alpha^* r}{r^* C_{m-1}^* W C_{m-1} r} .$$

As before $C_{m-1}^* W C_{m-1}$ is real symmetric positive definite so $(C_{m-1}^* W C_{m-1})^{-1/2}$ exists. We find that

$$\sigma_c^2 = \sup_{\hat{r}} \frac{\hat{r}^* (C_{m-1}^* W C_{m-1})^{-1/2} e_\alpha e_\alpha^* (C_{m-1}^* W C_{m-1})^{-1/2} \hat{r}}{\hat{r}^* \hat{r}} .$$

The supremum is over <u>real</u> $\hat{r}$ but the matrix $e_\alpha e_\alpha^*$ is <u>complex</u> so a Rayleigh quotient argument does not apply directly. Instead write $e_\alpha^* = u^* + iv^*$ where

$$u^* = \mathrm{Re}(e_\alpha^*) = (\mathrm{Re}(\alpha^{n-1}) \cdots \mathrm{Re}\,\alpha \quad 1)$$

and

$$v^* = \mathrm{Im}(e_\alpha^*) = (\mathrm{Im}(\alpha^{n-1}) \cdots \mathrm{Im}\,\alpha \quad 0) .$$

Then observe that for any <u>real</u> s,

$$s^* e_\alpha e_\alpha^* s = s^*(uu^* + vv^*)s \ .$$

Applying the Rayleigh quotient theorem now we find

$$\sigma_c^{\,2} = \max \text{ eigenvalue}[(C_{m-1}^* W C_{m-1})^{-1/2}(uu^*+vv^*)(C_{m-1}^* W C_{m-1})^{-1/2}]$$
$$= \max \text{ eigenvalue}[xx^* + yy^*]$$

where

$$x = (C_{m-1}^* W C_{m-1})^{-1/2} u \ ,$$
$$y = (C_{m-1}^* W C_{m-1})^{-1/2} v \ .$$

A rank two matrix has two positive eigenvalues which can be found by reduction to a matrix of dimension two. For an eigenvalue λ and an eigenvector $(\theta x+\phi y)$,

$$(\theta x + \phi y) = (xx^* + yy^*)(\theta x + \phi y) \ .$$

Therefore

$$\begin{pmatrix} x^*x & x^*y \\ y^*x & y^*y \end{pmatrix}\begin{pmatrix} \theta \\ \phi \end{pmatrix} = \lambda\begin{pmatrix} \theta \\ \phi \end{pmatrix}$$

and λ is an eigenvalue of the indicated two by two matrix. The largest eigenvalue of that matrix is

$$\lambda_{max} = \frac{1}{2}(x^*x + y^*y + ((x^*x-y^*y)^2 + 4|x^*y|^2)^{1/2}) \qquad (10.1)$$

where

$$x^*x = u^*(C_{m-1}^* W C_{m-1})u \ ,$$
$$x^*y = u^*(C_{m-1}^* W C_{m-1})v \ ,$$

etc. Then

(10.2) $$\sigma_c^2 = \lambda_{max}$$

and

(10.3) $$\gamma_c = \frac{\sqrt{\lambda_{max}}}{2m|\operatorname{Im}\alpha||q(\alpha)|} .$$

What does this result mean in the case $m = 1$? For comparison, suppose we computed the condition number γ of the same complex α using the general formula for complex polynomials (4.1, 6.1). The result is

$$\gamma = \sqrt{e_\alpha^* W^{-1} e_\alpha} / (2m|\operatorname{Im}\alpha||q(\alpha)|) .$$

To compute σ_c note that $x^*x = u^*W^{-1}u$, etc., and

$$\lambda_{max} = \tfrac{1}{2}(e_\alpha^* W^{-1} e_\alpha + \Delta^{1/2})$$

where $\Delta = (x^*x - y^*y)^2 + 4|x^*y|^2$. From the Cauchy-Schwartz inequality we can deduce that

$$(e_\alpha^* W^{-1} e_\alpha)^2 = (x^*x + y^*y)^2 \geq \Delta \geq 0 ,$$

and consequently

$$\tfrac{1}{2}(e_\alpha^* W^{-1} e_\alpha) \leq \lambda_{max} \leq e_\alpha^* W^{-1} e_\alpha .$$

Then we find that

(10.4) $$1 \leq \gamma/\gamma_c \leq \sqrt{2}$$

for $m = 1$.

When $m > 1$, however, the discrepancy between these condition numbers can be much greater. In fact, as $\operatorname{Im}\alpha \to 0$ for fixed $\operatorname{Re}\alpha$ and $m \geq 2$, γ/γ_c increases without bound. The condition numbers differ because γ maintains the multiplicity of only one zero intact

but γ_c maintains intact the multiplicities of two zeros.

Computational Details for σ_c

The computation of σ_c is similar to that of σ, except the matrix C_{m-1} corresponds to multiplication by $t(\tau) = (\tau-\alpha)^{m-1}(\tau-\bar{\alpha})^{m-1}$, a polynomial of degree $d = 2m-2$. Then C_{m-1} is n by n-2m+2, and

$$(C_{m-1}{}^{*}WC_{m-1})_{ij} = \begin{cases} \sum_{k=\max(i,j)}^{k=\min(i,j)+d} w_k t_{k-i}{}^{*} t_{k-j} , & |i-j| \leq d , \\ 0 & \text{otherwise} . \end{cases}$$

11. General Condition Numbers

The first condition numbers we considered reflected the condition of a zero subject to infinitesimal perturbations that maintain the multiplicity of (only) that zero. The second condition numbers reflected condition with respect to perturbations that maintain the multiplicity of that zero and its complex conjugate. We can go further, restricting the class of allowable perturbations to those that maintain whatever multiplicity structure we consider important in the other zeros.

For instance, let

$$p(\tau) = \left(\prod_{k=1}^{K} (\tau-\alpha_k)^{m_k}\right) q(\tau) ,$$

where

$$q(\alpha_k) \neq 0 , \quad 1 \leq k \leq K ,$$

and we consider only perturbations of the form

$$(p+\delta p)(\tau) = \prod_{k=1}^{K} (\tau - (\alpha_k+\delta\alpha_k))^{m_k} (q+\delta q)(\tau)$$

so that

$$\delta p(\tau) = \left(\prod_k (\tau-\alpha_k)^{m_k-1}\right)\left\{\left(\prod_k (\tau-\alpha_k)\right)\delta q(\tau) - q(\tau)\sum_k \left(m_k \delta\alpha_k \prod_{j\neq k} (\tau-\alpha_j)\right)\right\} .$$

In the usual way define the condition number γ of α with respect to such constrained perturbations to find that

$$(11.1) \quad \gamma = \frac{1}{m|q(\alpha_1)| \prod_{k=2}^{K} |\alpha_1-\alpha_k|} \cdot \sup_{\left(\deg r \leq n-1-\sum(m_k-1)\right)} \frac{|r(\alpha)|}{\left\| \left(\prod_{k=1}^{K} (\tau-\alpha_k)^{m_k-1}\right) r(\tau) \right\|} .$$

In the ℓ_2 case we can write the sup as

$$\sigma_G^2 = \sup \left(\frac{r^*ee^*r}{r^*G^*WGr}\right)$$

where G is the operator corresponding to multiplication by

$$(\tau-\alpha_1)^{m_1-1}(\tau-\alpha_2)^{m_2-1}\cdots(\tau-\alpha_K)^{m_K-1} .$$

Then as before, in the case of complex perturbations of a complex polynomial,

$$\sigma_G^2 = e^*(G^*WG)^{-1}e$$

where

$$e^* = (\alpha_1^{n-1}\ \alpha_1^{n-2}\ \cdots\ \alpha_1\ 1) .$$

The case of real perturbations of a real polynomial with real α_1 is similar. If α_1 is a complex zero of a real polynomial, however, then one of the other $\alpha_k = \bar{\alpha}_1$, and

$$\sigma_G^2 = \frac{1}{2}(x^*x + y^*y + \{(x^*x-y^*y)^2 + 4|x^*y|^2\}^{1/2}) ,$$

where $x^*x = u^*(G^*WG)^{-1}u$, $y^*y = v^*(G^*WG)^{-1}v$, etc., as in the previous section.

12. Application of the Idea of General Condition Number

Let

$$p(\tau) = (\tau-\alpha)^m(\tau-\bar{\alpha})^m q(\tau)$$

be a real polynomial with complex α. We have defined γ_c, the condition of α with respect to real changes which maintain conjugate m-tuple zeros $\alpha+\delta\alpha$ and $\bar{\alpha}+\overline{\delta\alpha}$. We want to compare γ_c to γ_2, the condition of α with respect to complex changes that maintain m-tuple zeros $\alpha+\delta\alpha$ and $\bar{\alpha}+\delta\beta$. $\delta\alpha$ and $\delta\beta$ are no longer necessarily complex conjugate.

We have seen that

$$\gamma_c = \frac{1}{2m|\operatorname{Im}\alpha||q(\alpha)|}\cdot\frac{1}{\sqrt{2}}\cdot\sqrt{x^*x + y^*y + \sqrt{(x^*x-y^*y)^2 + 4|x^*y|^2}}$$

where $x^*x = u^*(C_{m-1}{}^*WC_{m-1})^{-1}u$, $u^* = \operatorname{Re}(e_\alpha{}^*)$, etc. C_{m-1} corresponds to $(\tau-\alpha)^{m-1}(\tau-\bar{\alpha})^{m-1}$.

To compute γ_2, let

$$p(\tau) = (\tau-\alpha)^m(\tau-\bar{\alpha})^m q(\tau)\ .$$

Then

$$\gamma_2 = \frac{1}{m|q(\alpha)|}\cdot\frac{1}{|\alpha-\bar{\alpha}|}\sqrt{e_\alpha{}^*(GWG)^{-1}e_\alpha}$$

where G also corresponds to $(\tau-\alpha)^{m-1}(\tau-\bar{\alpha})^{m-1}$. Since $G = C_{m-1}$,

$$\gamma_2 = \frac{1}{2m|\operatorname{Im}\alpha||q(\alpha)|}\sqrt{x^*x+y^*y}$$

and

$$1 \le \frac{\gamma_2}{\gamma_c} \le \sqrt{2}\ . \tag{12.1}$$

In contrast to (10.4), our present result is independent of m. It means that the restriction to only real perturbations does not affect the condition number by a very large factor compared to a condition number that allows complex perturbations that maintain the multiplicities of the same number of complex zeros.

13. Condition Number vs. Distance to Submanifold

Now that we have a definition for condition number, we shall show why ill condition prompts us to look for the nearest polynomial with a more multiple zero. Consider the polynomial

$$p(\tau) = (\tau-\alpha)^m q(\tau) \ .$$

Then the condition of α is

$$\gamma = \frac{\frac{1}{m}\sigma}{|q(\alpha)|} = \frac{(m-1)!\sigma}{|p^{(m)}(\alpha)|} \ .$$

Consider the second polynomial

$$\hat{p}(\tau) = (\tau-\alpha)^m (q(\tau)-q(\alpha)) \ .$$

This polynomial has an m+1-tuple zero α. Further if

$$\Delta \equiv \|p-\hat{p}\| = |q(\alpha)| \|(\tau-\alpha)^m\| \ ,$$

then

$$\Delta = \frac{\frac{1}{m}\sigma \|(\tau-\alpha)^m\|}{\gamma} \ . \tag{13.1}$$

That is, if n, m, α, and the norm are regarded as fixed, then ill condition (large γ) always implies that there is a nearby polynomial with an m+1-tuple zero. Furthermore, the <u>closest</u> such polynomial may be much closer than the estimate above.

W. Kahan has suggested [17] that ill condition may be explained by exhibiting the nearest polynomial with a higher order zero. In the vector space of polynomials with m-tuple zeros, that corresponds to finding the closest point on the manifold of polynomials with m+1-tuple

zeros. If that m+1-tuple zero is still ill conditioned, then there must be a nearby polynomial on the submanifold of polynomials with m+2-tuple zeros.

In the chapters that follow we shall describe ways of finding the nearest polynomial with an m-tuple zero.

CHAPTER III

FINDING THE NEAREST POLYNOMIAL WITH AN m-TUPLE ZERO

1. Introduction

In the first chapter we discussed why we might wish to find the nearest polynomial with an m-tuple zero. Now we will demonstrate how to set up the equations to be solved. The problem amounts to a constrained optimization, and in general we find we must solve a non-analytic equation in a complex variable.

We first consider the simplest cases of the problem: finding the nearest real polynomial with an n-tuple zero or with a double zero.

Then we discuss the equations to be solved for the stationary points which include the nearest complex polynomial with an m-tuple zero. Finally we explain two kinds of second derivatives which may be used for deciding which stationary points are actually minima.

2. The Nearest Polynomial with an n-tuple Zero

We will start by considering the simplest case -- that of finding the nearest polynomial with an n-tuple zero. We suppose that we have a monic polynomial

$$p(\tau) = \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j}$$

and we wish to find another polynomial

$$q(\tau) = (\tau-\zeta)^n = \tau^n + \sum_{j=1}^{n} \binom{n}{j}(-\zeta)^j \tau^{n-j} \quad , \quad \binom{n}{j} = \frac{n!}{j!(n-j)!}$$

such that $\|p-q\|$ is a minimum.

Since

$$p - q = \begin{bmatrix} \vdots \\ p_j - \binom{n}{j}(-\zeta)^j \\ \vdots \end{bmatrix}$$

and depends only on ζ we can easily find the equation to be solved for stationary points with respect to a given norm. We will demonstrate the equation for the weighted ℓ_2 norms as follows:

If we let the raised dot $\dot{}$ represent $\frac{\partial}{\partial \operatorname{Re}\zeta}$ or $\frac{\partial}{\partial \operatorname{Im}\zeta}$ we find

$$(\|r\|_w^2)^{\cdot} = \dot{r}^* W r + r^* W \dot{r} = 2 \operatorname{Re}(r^* W \dot{r}) \ .$$

For stationarity we require then $\operatorname{Re}(r^* W \dot{r}) = 0$. Thus

$$0 = \operatorname{Re} \sum_{j=1}^{n} \left(p_j - \binom{n}{j}(-\zeta)^j\right)^* w_j \left(-\binom{n}{j} \cdot j \cdot (-\zeta)^{j-1} \cdot (-1) \cdot \dot{\zeta}\right)$$

$$= \operatorname{Re} \sum_{j=1}^{n} w_j \cdot j \binom{n}{j} (-\zeta)^{j-1} \left(p_j - \binom{n}{j}(-\zeta)^j\right)^* \dot{\zeta}$$

or

$$(2.1) \qquad f(\zeta) \equiv \sum_{j=1}^{n} w_j \cdot j \binom{n}{j} (-\zeta^*)^{j-1} \left(p_j - \binom{n}{j}(-\zeta)^j\right) = 0 .$$

$f(\zeta)$ is thus our first example of a non-analytic function of a complex variable ζ. To find a zero would in general require solving a system of two equations in two real variables.

In the most interesting case, however, we would be interested in real perturbations $q - p$ of a real polynomial p. If ζ were complex then $q - p$ could not be real, so we need only consider cases for which ζ is real. Then the real function $f(\zeta)$ is

$$(2.2) \quad f(\zeta) = w_1 n(p_1 + n\zeta) + \zeta \sum_{j=2}^{n} w_j \cdot j \binom{n}{j} (-\zeta)^{j-2} \left\{\binom{n}{j}\zeta^2(-\zeta)^{j-2} - p_j\right\} .$$

We write $f(\zeta)$ in this way for comparison with the expression for $f'(\zeta)$:

$$(2.3) \quad f'(\zeta) = w_1 n^2 + \sum_{j=2}^{n} w_j \cdot j \cdot \binom{n}{j} \cdot (-\zeta)^{j-2} \left\{(2j-1)\binom{n}{j}\zeta^2(-\zeta)^{j-2} - (j-1)p_j\right\} .$$

Then we may use Newton's method from a suitable starting point to find a stationary point ζ. $f(\zeta)$ is evidently a real polynomial of odd degree $2n - 1$ so it does have at least one real zero. We shall see later that even when $n = 2$ there may be more than one real zero. We could in principle find all the zeros of f with a conventional polynomial zero finding technique, but we would have to reject most of those zeros as irrelevant since they would be complex.

In practice it appears that when Newton's method is started from $\zeta = -p_1/n$, convergence occurs quickly to a stationary value which

appears to be a reasonable candidate for a global minimum. This choice of starting point makes sense because, when we consider $p(\tau) = (\tau-\zeta_0)^n + \varepsilon q(\tau)$ for infinitesimal perturbations εq, the solution turns out to be $\zeta = \zeta_0 - \frac{1}{n}\varepsilon q_1 = -\frac{p_1}{n}$.

Even in the apparently simple case of finding the nearest n-tuple zero we encounter most of the characteristic difficulties of the more complicated cases of m-tuple zeros for $m < n$. In the next sections we will explore these cases in detail.

3. The Nearest Polynomial with a Fixed Double Zero

In the present section we will solve the following problem: given a real polynomial

$$p(\tau) = \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j} ,$$

what is the least real perturbation

$$q(\tau) = \sum_{j=1}^{n} q_j \tau^{n-j}$$

such that $p+q$ has a specified real double zero ζ? We will measure perturbations q by the familiar ℓ_2 norms $\|q\|_w^2 = q^T W q = \sum_{j=1}^{n} w_j q_j^2$.

Our problem is to minimize $\|q\|_w^2$ subject to the constraints that $p(\tau) + q(\tau) = (\tau-\zeta)^2 r(\tau)$ for some r of degree $n-2$. Using the notations of the chapter on condition numbers, then, our problem is to find r to minimize

$$\|P_2 r - p\|_w = \|W^{1/2} P_2 r - W^{1/2} p\|_2 .$$

Recall that P_2 is the operator which multiplies polynomials of degree $n-2$ by $(\tau-\zeta)^2$.

The solution of this linear least squares problem is

$$r = (W^{1/2} P_2)^T W^{1/2} p .$$

Then

$$q = (P_2 (P_2^* W P_2)^{-1} P_2 W - 1) p .$$

Thus we can solve this problem by the usual least squares method. But when we do not specify ζ in advance that method is inapplicable

since P_2 now depends on ζ. Therefore we will look at a dual formulation of the problem that can easily be expanded when we allow ζ to vary.

So now when we minimize $\|q\|_w^2$ subject to $(p+q)(\zeta) = 0$ and $(p+q)'(\zeta) = 0$ we will apply Lagrange multipliers according to the conventional formulation. Namely we will seek the stationary points of

$$\phi = \sum_{j=1}^{n} w_j (q_j)^2 + \lambda_0 (p(\zeta) + q(\zeta)) + \lambda_1 (p'(\zeta) + q'(\zeta))$$

with respect to changes in q_j. We note that $q(\zeta) = \sum_{j=1}^{n} q_j \zeta^{n-j}$ so $\frac{\partial (q(\zeta))}{\partial q_j} = \zeta^{n-j}$ and $\frac{\partial (q'(\zeta))}{\partial q_j} = (n-j)\zeta^{n-j-1}$. Thus

$$0 = \frac{\partial \phi}{\partial q_j} = 2w_j q_j + \lambda_0 \zeta^{n-j} + \lambda_1 (n-j)\zeta^{n-j-1}$$

whence

$$q_j = \frac{-1}{2w_j}\{\lambda_0 \zeta^{n-j} + \lambda_1 (n-j)\zeta^{n-j-1}\}, \quad j < n \quad \text{and} \quad q_n = \frac{-\lambda_0}{2w_n}.$$

To determine λ_0 and λ_1 we will use the constraints:

$$0 = (p+q)(\zeta) = p(\zeta) + (-\tfrac{1}{2}) \sum_{j=1}^{n-1} \{\frac{1}{w_j}(\lambda_0 (\zeta^2)^{n-j} + \lambda_1 (n-j)\zeta(\zeta^2)^{n-j-1})\} - \frac{\lambda_0}{2w_n}$$

$$0 = (p+q)'(\zeta) = p'(\zeta) + (-\tfrac{1}{2}) \sum_{j=1}^{n-1} \frac{1}{w_j}(\lambda_0 (n-j)\zeta(\zeta^2)^{n-j-1} + \lambda_1 (n-j)^2 (\zeta^2)^{n-j-1}).$$

The above may be written as a linear system of equations:

$$\begin{bmatrix} \sum_{j=1}^{n} \frac{1}{w_j}(\zeta^2)^{n-j} & \zeta\sum_{j=1}^{n-1} \frac{1}{w_j}(n-j)(\zeta^2)^{n-j-1} \\ \zeta\sum_{j=1}^{n-1} \frac{1}{w_j}(n-j)(\zeta^2)^{n-j-1} & \sum_{j=1}^{n-1} \frac{1}{w_j}(n-j)^2(\zeta^2)^{n-j-1} \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \end{bmatrix} = 2 \begin{bmatrix} p(\zeta) \\ p'(\zeta) \end{bmatrix}$$

If we write $\sigma_k = \sum_{j=1}^{n} \frac{1}{w_j}(n-j)^k(\zeta^2)^{n-j}$ then

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \end{bmatrix} = \frac{2}{\sigma_0\sigma_2-\sigma_1^2} \begin{bmatrix} \sigma_2 & -\zeta\sigma_1 \\ -\zeta\sigma_1 & \zeta^2\sigma_0 \end{bmatrix} \begin{bmatrix} p(\zeta) \\ p'(\zeta) \end{bmatrix}$$

and

$$q_j = \frac{-1}{w_j} \frac{\zeta^{n-j}}{\sigma_0\sigma_2-\sigma_1^2}\{(\sigma_2-(n-j)\sigma_1)p(\zeta) + \zeta(-\sigma_1+(n-j)\sigma_0)p'(\zeta)\} .$$

Then

$$(3.1) \quad q(\tau) = \frac{1}{\sigma_1^2-\sigma_0\sigma_2} \sum_{j=1}^{n} \frac{1}{w_j}\{(\sigma_2-(n-j)\sigma_1)p(\zeta) + \zeta(-\sigma_1+(n-j)\sigma_0)p'(\zeta)\} \cdot \zeta^{n-j}\tau^{n-j}$$

is the smallest perturbation moving $p(\tau)$ to the manifold of polynomials having double zeros at ζ. The distance may be calculated to be

$$\|q\|_w = \left\{\frac{\sigma_2(p(\zeta))^2 - 2\sigma_1 p(\zeta)(\zeta p'(\zeta)) + \sigma_0(\zeta p'(\zeta))^2}{\sigma_0\sigma_2 - \sigma_1^2}\right\}^{1/2} .$$

The foregoing calculation is invalid when $\zeta = 0$. In that case $q_n = -p_n$, $q_{n-1} = -p_{n-1}$, and $q_j = 0$, $1 \le j \le n-2$.
$\|q\|_w^2 = w_{n-1}(p_{n-1})^2 + w_n(p_n)^2$.

4. The Nearest Polynomial with a Double Zero

After the complicated expressions of the previous section, one would expect worse from the following problem: given real p, find real q such that $p+q$ has a real double zero ζ not fixed in advance, so that ζ may vary. The final expressions to be derived are surprisingly simple, however.

We could solve this problem by differentiating with respect to ζ the final expression for $\|q\|_w^2$ of the previous section. It will be more enlightening, however, to make a fresh start. The direct linear least squares solution method won't work now, and we must solve the problem with Lagrange multipliers. Thus we seek the stationary points of

$$\nu = \sum_{j=1}^{n} w_j (q_j)^2 + \lambda_0 (p+q)(\zeta) + \lambda_1 (p+q)'(\zeta)$$

with respect to variations in q_j and ζ. Then as before

$$0 = \frac{\partial \nu}{\partial q_j} = 2w_j q_j + \lambda_0 \zeta^{n-j} + \lambda_1 (n-j)\zeta^{n-j-1} ,$$

but now, in addition,

$$0 = \frac{\partial \nu}{\partial \zeta} = \lambda_0 (p+q)'(\zeta) + \lambda_1 (p+q)''(\zeta) .$$

We exploit the constraint $(p+q)'(\zeta) = 0$ to see that

$$0 = \lambda_1 (p+q)''(\zeta) .$$

Remarkably enough, either one of the Lagrange multipliers is identically zero or else the unknown ζ is not only a double but a triple zero of $p+q$. It turns out that stationary points with $(p+q)''(\zeta) = 0$

and $\lambda_1 \neq 0$ are almost never minima; see section 9. Accepting that assertion for the time being, assume $\lambda_1 = 0$. Then

$$q_j = \frac{-1}{2w_j} \lambda_0 \zeta^{n-j} .$$

From the constraint $(p+q)(\zeta) = 0$, we find

$$p(\zeta) = \frac{1}{2}\lambda_0 \sum_{j=1}^{n} \frac{1}{w_j} \zeta^{n-j} \cdot \zeta^{n-j}$$

so

$$\lambda_0 = \frac{2p(\zeta)}{\sigma_0}$$

and

$$q_j = \frac{-p(\zeta)}{\sigma_0} \frac{1}{w_j} \zeta^{n-j} ,$$

$$q(\tau) = \frac{-p(\zeta)}{\sigma_0} \sum_{j=1}^{n} \frac{1}{w_j} \zeta^{n-j} \tau^{n-j} .$$

We still don't know ζ, but we can exploit the constraint $(p+q)'(\zeta) = 0$ to find

$$-p'(\zeta) = \sum_{j=1}^{n} q_j (n-j) \zeta^{n-j-1} = \frac{-p(\zeta)}{\sigma_0} \sum \left(\frac{n-j}{w_j}\right) \zeta^{n-j+n-j-1}$$

and

$$\frac{\zeta p'(\zeta)}{p(\zeta)} = \frac{\sigma_1}{\sigma_0} = \frac{\sum_{j=1}^{n-1} \left(\frac{n-j}{w_j}\right) (\zeta^2)^{n-j}}{\sum_{j=1}^{n} \left(\frac{1}{w_j}\right) (\zeta^2)^{n-j}} \tag{4.1}$$

is the equation to be solved for ζ. Apparently it could be written as a polynomial equation of degree $3n - 2$. We will devote several sections to discussions of ways to solve this equation. Let it suffice

to say that when p is real, the equation always has a solution $\zeta = 0$, and when $n \geq 2$ is even and $p_{n-1} \neq 0$ it always has at least one other real solution as well.

Once a solution ζ has been found, the corresponding distance is

$$\|q\|_w = \frac{|p(\zeta)|}{\sqrt{\sigma_0}} = \frac{|\zeta p'(\zeta)|}{\sigma_1} \sqrt{\sigma_0} .$$

There are usually several real solutions ζ and, surprisingly, most of them are local minima, rather than maxima or saddle points. It turns out that the maxima are usually the stationary points with $(p+q)''(\zeta) = 0$. A difficult, unsolved problem is to find the ζ corresponding to a global minimum of $\|q\|$ without having to find <u>all</u> the solutions ζ.

5. The Nearest Polynomial with a Fixed m-tuple Zero

Using the notation of Chapter I we will now show how to find the nearest polynomial with an m-tuple zero ζ. We wish to minimize $\|q\|_W^2 = q^*Wq$ subject to $\tilde{A}p + Aq = 0$.

We may find the linear least squares solution directly. The vector $W^{1/2}q$ of least Euclidean norm solving $(AW^{-1/2})(W^{1/2}q) = -\tilde{A}p$ is just $(W^{1/2}q) = (AW^{-1/2})^{\dagger}(-\tilde{A}p)$, where † denotes pseudo inverse. Since A has more columns than rows, and the rows are linearly independent,

$$(AW^{-1/2})^{\dagger} = W^{-1/2}A^*(AW^{-1}A^*)^{-1} ,$$

whence

$$q = -W^{-1}A^*(AW^{-1}A^*)^{-1}\tilde{A}p . \tag{5.1}$$

Consequently

$$\|q\|_W = \left((\tilde{A}p)^*(AW^{-1}A^*)^{-1}\tilde{A}p\right)^{1/2} .$$

To compare this with our earlier results for real double zeros, we let $m = 2$ and recall that when $m = 2$,

$$A \equiv \begin{pmatrix} e^* \\ e^*D \end{pmatrix} ,$$

so

$$AW^{-1}A^* = \begin{pmatrix} e^*W^{-1}e & e^*W^{-1}D^*e \\ e^*DW^{-1}e & e^*DW^{-1}D^*e \end{pmatrix} .$$

We can derive expressions for the matrix elements in terms of the

$$\sigma_k \equiv \sum_{j=1}^{n} \frac{(n-j)^k}{w_j} |\zeta^2|^{n-j} .$$

Notice that this is a redefinition of the σ_k replacing the previous definition $\sum_{j=1}^{n} \frac{(n-j)^k}{w_j} (\zeta^2)^{n-j}$ which is not suitable for complex ζ. Then

$$e^*W^{-1}e = \sigma_0 ,$$
$$e^*DW^{-1}e = \frac{1}{\zeta}\sigma_1 = (e^*W^{-1}D^*e)^* ,$$
$$e^*DW^{-1}D^*e = \frac{1}{|\zeta|^2}\sigma_2 .$$

Therefore

$$(AW^{-1}A^*)^{-1} = \frac{|\zeta|^2}{\sigma_0\sigma_2-\sigma_1^2}\begin{pmatrix} \frac{1}{|\zeta|^2}\sigma_2 & \frac{-1}{\zeta^*}\sigma_1 \\ \frac{-1}{\zeta}\sigma_1 & \sigma_0 \end{pmatrix}$$

and

$$\|q\|_w = \left[\frac{\sigma_2|p(\zeta)|^2-2\sigma_1 \mathrm{Re}(p^*(\zeta)\zeta p'(\zeta))+\sigma_0|\zeta p'(\zeta)|^2}{\sigma_0\sigma_2-\sigma_1^2}\right]^{1/2} .$$

Apparently the major difference between the previous real case and the present complex case is that expressions like $(\theta)^2$ have been replaced by expressions like $|\theta|^2$. The effect of this change will be that the equations to be solved for ζ, when it is <u>not</u> fixed in advance, will no longer be analytic.

6. The Nearest Polynomial with an m-tuple Zero, No Longer Fixed

Our problem appears similar to that in a previous section: minimize $\|q\|_W$ subject to $\tilde{A}p + Aq = 0$. The difference is that the ζ on which $\tilde{A}$ and A depend is no longer fixed, and a linear least squares theory is no longer applicable. As we have just seen, if we do hold ζ fixed, we can write q as a non-analytic function of ζ. Therefore we can find a directional derivative of q if we think of ζ as a function of a real parameter θ: $\zeta = \zeta_0 + \theta\dot{\zeta}$. Then $\frac{d\zeta}{d\theta} = \dot{\zeta}$ and if

$$\nu \equiv q^*Wq$$

then

$$\frac{d\nu}{d\theta} = \dot{\nu} = \dot{q}^*Wq + q^*W\dot{q} = 2\,\mathrm{Re}\,(q^*W\dot{q})$$

since W is constant. At a stationary point of ν we would require $\dot{\nu} = 0$ for all $\dot{q}$, including that particular one which makes $q^*W\dot{q}$ real. From that case we conclude that

$$0 = q^*W\dot{q}$$

is the condition for stationarity.

But q is constrained in the values it may take. When we differentiate that constraint we find $\dot{\tilde{A}}p + \dot{A}q + A\dot{q} = 0$. Since $(e^{\dot{*}}) = (\cdots(\zeta^{\dot{n-j}})\cdots) = (\cdots(n-j)\zeta^{n-j-1}\dot{\zeta}\cdots) = e^*D\dot{\zeta}$, we conclude that $\dot{A} = AD\dot{\zeta}$. Therefore the constraint on q and $\dot{\zeta}$ is $(\tilde{A}Dp + ADq)\dot{\zeta} + A\dot{q} = 0$.

The idea of constrained optimization is that every pair $(\dot{q},\dot{\zeta})$ which satisfies the constraint should also satisfy the stationarity property, i.e., in the notation of the Lagrange multiplier theorem (Appendix 6),

$$Bx = 0 \;\Rightarrow\; y^*x = 0\ ,$$

where

$$x = \begin{pmatrix} \dot{q} \\ \hline \dot{\zeta} \end{pmatrix}$$

$$B = (A \mid \tilde{\tilde{A}}Dp + ADq)$$

and $$y^* = (q^*W \mid 0) \ .$$

The Lagrange multiplier theorem just cited assures us that y may be written $y = B^*\ell$ for some vector ℓ of Lagrange multipliers. For convenience we will write

$$\ell = \begin{pmatrix} \ell_0 \\ \vdots \\ \ell_{m-1} \end{pmatrix} .$$

Then

$$\begin{pmatrix} Wq \\ 0 \end{pmatrix} = \begin{pmatrix} A^* \\ \hline (\tilde{A}Dp+ADq)^* \end{pmatrix}\ell \ . \tag{6.1}$$

But since $\tilde{A}p + Aq = 0$ is the constraint, $(\tilde{A}Dp + ADq)^*\ell = 0 \Rightarrow ((p+q)^{(m)}(\zeta))^*\ell_{m-1} = 0$ and we are therefore faced with the two possibilities we saw in the $m = 2$ case: either the last Lagrange multiplier is zero, or the zero ζ has one higher multiplicity than we had planned. By examining the second derivative $\ddot{v}$ in a subsequent section we will find that stationary points with extra multiplicity corresponding to minima of v always have $\ell_{m-1} = 0$. Therefore we may always assume that $\ell_{m-1} = 0$ at interesting stationary points.

Continuing we find $Wq = A^*\ell$ so $q = W^{-1}A^*\ell$. Then the constraint implies $(AW^{-1}A^*)\ell = -\tilde{A}p$. Although $AW^{-1}A^*$ is Hermitian positive definite and therefore invertible, we would find that ℓ_{m-1} would not come out to be zero except for certain special ζ's. These special

values of ζ must correspond to the stationary points of ν. To find out what they are, we write $\ell = \begin{pmatrix} \hat{\ell} \\ 0 \end{pmatrix}$ and

$$\tilde{A}p + (AW^{-1}A^*)\begin{pmatrix} \hat{\ell} \\ 0 \end{pmatrix} = 0$$

or

$$(\tilde{A}p \quad AW^{-1}A^*Z)\begin{pmatrix} 1 \\ \hat{\ell} \end{pmatrix} = 0 .$$

Here

$$Z = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ & & 1 \\ 0 & \cdots & 0 \end{pmatrix} \begin{matrix} \} m \\ \\ \\ \\ \underbrace{\qquad}_{m-1} \end{matrix}$$

and it has the effect of removing the last column of $AW^{-1}A^*$. The resulting homogeneous equation above obviously has a nontrivial solution so the matrix is singular. Therefore

$$0 = \det(\tilde{A}p \mid AW^{-1}A^*Z) \tag{6.2}$$

is the equation to be solved to find the ζ's corresponding to interesting stationary points of ν.

To see what kind of equation it is, consider the case $m = 2$:

$$AW^{-1}A^* = \begin{pmatrix} e^*W^{-1}e & e^*W^{-1}D^*e \\ e^*DW^{-1}e & e^*DW^{-1}D^*e \end{pmatrix}$$

so

$$0 = \det \begin{pmatrix} p(\zeta) & e^*W^{-1}e \\ p'(\zeta) & e^*DW^{-1}e \end{pmatrix} = \frac{1}{\zeta}\sigma_1 p(\zeta) - \sigma_0 p'(\zeta) ,$$

which we may write

$$\frac{\zeta p'(\zeta)}{p(\zeta)} = \frac{\sum_{j=1}^{n} \frac{(n-j)}{w_j} |\zeta^2|^{n-j}}{\sum_{j=1}^{n} \frac{1}{w_j} |\zeta^2|^{n-j}} \,. \tag{6.3}$$

This equation is evidently <u>not</u> that of an analytic function. We shall return to it later. Supposing for now that we have found an acceptable solution ζ for the equation above; we can then evaluate $\hat{\ell}$ from

$$AW^{-1}A^*Z\hat{\ell} = -\tilde{A}p$$

in any of a variety of ways; the obvious way is to solve

$$(Z^*AW^{-1}A^*Z)\hat{\ell} = -Z^*\tilde{A}p \,.$$

This equation is the same as

$$\hat{A}W^{-1}\hat{A}^*\hat{\ell} = -\hat{\tilde{A}}p$$

where $\hat{A}$ is one dimension smaller than A, i.e., $A = \begin{pmatrix} \hat{A} \\ e^*D^{m-1} \end{pmatrix}$. Then $q = W^{-1}\hat{A}^*\hat{\ell}$ and finally

$$\|q\|_W = (\hat{\ell}^*\hat{A}W^{-1}\hat{A}^*\hat{\ell})^{1/2} = \left((\hat{\tilde{A}}p)^*(\hat{A}W^{-1}\hat{A}^*)^{-1}(\hat{\tilde{A}}p)\right)^{1/2} \,.$$

For the case $m = 2$ that we considered previously,

$$\begin{aligned} \hat{A}W^{-1}\hat{A}^* &= \sigma_0 \,, \\ \hat{\ell} &= -p(\zeta)/\sigma_0 \,, \\ q &= (-p(\zeta)/\sigma_0)W^{-1}e \,, \end{aligned} \tag{6.4}$$

and

$$\|q\|_W = |p(\zeta)|/\sqrt{\sigma_0} \,.$$

7. Computational Details: The Equation to Solve for the Nearest m-tuple Zero

As we have seen, in order to find the nearest polynomial with a double zero, we must solve the equation

$$h(\tau) \equiv \sigma_1 p(\tau) - \sigma_0 \tau p'(\tau) = 0$$

where

$$\sigma_k \equiv \sum_{j=1}^{n} \frac{(n-j)^k}{w_j} |\tau^2|^{n-j} .$$

We will see that there are various ways of solving this equation for its zeros ζ when τ and p are real, but for the more general complex case there do not seem to be many methods that work. We will usually solve this equation by means of Newton's method applied to two real equations in two real unknowns. In this section we will provide the expressions necessary for Newton's method in the case of an m-tuple zero.

The equation we have to solve is in this form:

$$0 = \det(\tilde{A}p \mid AW^{-1}A^*Z)$$

or, written out,

$$0 = \begin{vmatrix} p(\zeta) & e^*W^{-1}e & \cdots & e^*W^{-1}(D^{m-2})^*e \\ p'(\zeta) & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ p^{(m-1)}(\zeta) & e^*D^{m-1}W^{-1}e & \cdots & e^*D^{m-1}W^{-1}(D^{m-2})^*e \end{vmatrix} .$$

Now $e^*D^iW^{-1}(D^j)^*e = \frac{1}{\zeta^i(\zeta^j)^*}\left[\sum_{k=1}^{n} \frac{1}{w_k} \frac{(n-k)!}{(n-k-i)!} \frac{(n-k)!}{(n-k-j)!} |\zeta^2|^{n-k}\right] \equiv \frac{\sigma_{ij}}{\zeta^i(\zeta^j)^*}$.

By multiplying rows and columns by powers of ζ and ζ^* we can

rewrite the determinant without changing its value as

$$0 = \begin{vmatrix} p(\zeta) & \sigma_{00} & \cdots & \sigma_{0,m-2} \\ \zeta p'(\zeta) & \sigma_{10} & & \vdots \\ \vdots & \vdots & & \vdots \\ \zeta^{m-1} p^{[m-1]}(\zeta) & \sigma_{m-1,0} & \cdots & \sigma_{m-1,m-2} \end{vmatrix} \equiv f(\zeta) .$$

In this form it is obvious that the expansion in terms of minors from the first column will yield

$$\begin{aligned} f(\zeta) &= \Delta_0 p(\zeta) - \Delta_1(\zeta p'(\zeta)) + \cdots + (-1)^{m-1}\Delta_{m-1}(\zeta^{m-1} p^{[m-1]}(\zeta)) \\ &= (\Delta_0 \ -\Delta_1 \ \cdots \ (-1)^{m-1}\Delta_{m-1}) \begin{pmatrix} p(\zeta) \\ \zeta p'(\zeta) \\ \vdots \\ \zeta^{m-1} p^{[m-1]}(\zeta) \end{pmatrix} \\ &\equiv v^* u . \end{aligned}$$

Thus f may be expressed as a scalar product of (1) a vector u of analytic functions of ζ and (2) a vector v of functions depending only on σ_{ij} and hence only on $|\zeta^2|$. In fact the Δ_j are <u>real</u> analytic functions of the <u>real</u> variable $|\zeta^2|$.

The two real equations which we shall solve by Newton's method are $\operatorname{Re} f = 0$ and $\operatorname{Im} f = 0$, that is,

$$\begin{aligned} v^* \operatorname{Re} u &= 0 , \\ v^* \operatorname{Im} u &= 0 . \end{aligned} \tag{7.1}$$

Now

$$\frac{\partial \operatorname{Re} f}{\partial \operatorname{Re} \zeta} = \left(\frac{\partial v}{\partial \operatorname{Re} \zeta}\right)^* \operatorname{Re} u + v^* \left(\frac{\partial \operatorname{Re} u}{\partial \operatorname{Re} \zeta}\right) = (v')^* \frac{\partial(|\zeta^2|)}{\partial \operatorname{Re} \zeta} \operatorname{Re} u + v^* \operatorname{Re} u'$$

$$= 2 \operatorname{Re} \zeta \operatorname{Re}((v')^* u) + \operatorname{Re}(v^* u') . \tag{7.2a}$$

Similarly

$$\text{(7.2b)}\qquad \begin{aligned} \frac{\partial \operatorname{Re} f}{\partial \operatorname{Im} \zeta} &= 2 \operatorname{Im}\zeta \operatorname{Re}((v')^*u) - \operatorname{Im}(v^*u') , \\ \frac{\partial \operatorname{Im} f}{\partial \operatorname{Re} \zeta} &= 2 \operatorname{Re}\zeta \operatorname{Im}((v')^*u) + \operatorname{Im}(v^*u') , \\ \frac{\partial \operatorname{Im} f}{\partial \operatorname{Im} \zeta} &= 2 \operatorname{Im}\zeta \operatorname{Im}((v')^*u) + \operatorname{Re}(v^*u') . \end{aligned}$$

In general v^* is a vector whose components are functions of the σ_{ij} which can in turn be written as functions of the σ_k defined earlier. Then $\sigma_k' = \frac{1}{|\zeta^2|}\sigma_{k+1}$.

For the case $m = 2$ we have $v^* = (\sigma_1 - \sigma_0)$ and $u = \begin{pmatrix} p(\zeta) \\ \zeta p'(\zeta) \end{pmatrix}$. Then

$$(v')^*u = \frac{1}{|\zeta|^2}\{\sigma_2 p(\zeta) - \sigma_1 \zeta p'(\zeta)\}$$

and

$$v^*u' = \{\sigma_1 p'(\zeta) - \sigma_0(\zeta p'(\zeta) - p''(\zeta))\}$$

are the quantities required in the expressions for the partial derivatives. Those partial derivatives enable us to compute the Jacobian matrix required for Newton's method in two dimensions.

The case for $m = 3$ is more complicated. In accordance with the previous formulation,

$$\Delta_0 = \begin{vmatrix} \sigma_{10} & \sigma_{11} \\ \sigma_{20} & \sigma_{21} \end{vmatrix} = \sigma_1(\sigma_3 - \sigma_2) - (\sigma_2 - \sigma_1)\sigma_2 = \sigma_1\sigma_3 - \sigma_2^2 ,$$

$$\Delta_1 = \sigma_0\sigma_3 - \sigma_0\sigma_2 - \sigma_1\sigma_2 + \sigma_1^2 ,$$

$$\Delta_2 = \sigma_0\sigma_2 - \sigma_1^2 .$$

For simplicity we will make a slight change:

$$(7.3)\qquad v^*u = v^*\begin{pmatrix}1&0&0\\0&1&0\\0&-1&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&1&1\end{pmatrix}u$$

$$= (\sigma_1\sigma_3-\sigma_2^2,-(\sigma_0\sigma_3-\sigma_1\sigma_2),\sigma_0\sigma_2-\sigma_1^2)\begin{pmatrix}p(\zeta)\\ \zeta p'(\zeta)\\ \zeta p'(\zeta)+\zeta^2 p''(\zeta)\end{pmatrix}$$

$$\equiv \hat{v}^*\hat{u}\ .$$

With v^* and u thus redefined,

$$(\hat{v}')^* = \frac{1}{|\zeta|^2}(\sigma_1\sigma_4-\sigma_2\sigma_3,-(\sigma_0\sigma_4-\sigma_2^2),\sigma_0\sigma_3-\sigma_1\sigma_2) \equiv \frac{1}{|\zeta|^2}(\hat{\Delta}_0,\hat{\Delta}_1,\hat{\Delta}_2)$$

$$(7.4)\qquad \hat{u}' = \begin{pmatrix}p(\zeta)\\ \zeta p'(\zeta)\\ \zeta^2 p'''+3\zeta p''+p'\end{pmatrix}\ .$$

It may be observed that expressions like $\sigma_1\sigma_3-\sigma_2^2$ involving subtraction of positive quantities will result in cancellation. Therefore we will rewrite those expressions. Let a typical term be

$$\hat{\Delta} = \sigma_a\sigma_b - \sigma_c\sigma_d\ .$$

Then

$$\hat{\Delta} = \Big(\sum\frac{(n-j)^a}{w_j}|\zeta^2|^{n-j}\Big)\Big(\sum\frac{(n-k)^b}{w_k}|\zeta^2|^{n-k}\Big)$$

$$-\Big(\sum\frac{(n-j)^c}{w_j}|\zeta^2|^{n-j}\Big)\Big(\sum\frac{(n-k)^d}{w_k}|\zeta^2|^{n-k}\Big)$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n}\frac{1}{w_j}\frac{1}{w_k}|\zeta^2|^{n-j}|\zeta^2|^{n-k}\{(n-j)^a(n-k)^b-(n-j)^c(n-k)^d\}\ .$$

This double sum has an entry for each position in an n by n square array, except for the diagonal entries which vanish. Therefore, we may add the i,j and j,i terms together and count only the terms

with $k > j$:

$$\hat{\Delta} = |\zeta^2| \sum_{j=1}^{n-1} \frac{1}{w_j} |\zeta^2|^{n-j-1} \Big[\sum_{k=j+1}^{n} \frac{1}{w_k} |\zeta^2|^{n-k} \{\cdot\} \Big] ,$$

$$\{\cdot\} \equiv (n-j)^a (n-k)^b + (n-j)^b (n-k)^a - (n-j)^c (n-k)^d - (n-j)^d (n-k)^c .$$

If we consider $\hat{\Delta}$ to be a function of a real variable $|\zeta^2|$ then we may define $\hat{\Delta}'$ as $\frac{\partial \hat{\Delta}}{\partial |\zeta^2|}$. Then

$$\hat{\Delta}' = \sum_{j=1}^{n-1} \frac{1}{w_j} |\zeta^2|^{n-j-1} \Big[\sum_{k=j+1}^{n} \frac{1}{w_k} |\zeta^2|^{n-k} (n-j+n-k) \{\cdot\} \Big] .$$

The expression $\{\cdot\}$ in the equations above has the following values:

$$\text{for } \hat{\Delta}_0, \quad (n-j)(n-k)(k-j)^2 ;$$

$$\text{for } \hat{\Delta}_1, \quad (n-k+n-j)(k-j)^2 ;$$

$$\text{for } \hat{\Delta}_2, \quad (k-j)^2 .$$

We may use these expressions for $\hat{\Delta}$ and $\hat{\Delta}'$ to compute $\hat{v}$ and $\hat{v}'$. Using the expressions for $\hat{u}$ (7.3) and $\hat{u}'$ (7.4) we may solve the equations for the nearest polynomial with a triple zero (7.1). The partial derivatives (7.2) are used by Newton's method.

8. The Second Derivative of $\|q\|$

We have just seen which equation must be solved to find the stationary points of $\|q\|$. Some of these points are local minima; others are maxima or saddle points. To investigate the nature of the stationary points we now develop expressions for directional second derivatives of $\|q\|_W^2$.

Suppose that $\zeta = \zeta_0 + \theta\dot{\zeta}$ for θ real. Let the function to be minimized be $\nu = q^*Wq$. As we have seen,

$$\frac{d\nu}{d\theta} = \dot{\nu} = 2\,\mathrm{Re}\,(q^*W\dot{q}) = 2\,\mathrm{Re}\,(\ell^*A\dot{q}) \ .$$

But the constraint $\tilde{A}p + Aq = 0$ implies $A\dot{q} = -(\tilde{A}\tilde{D}p + ADq)\dot{\zeta}$ so $\dot{\nu} = -2\,\mathrm{Re}\,(\ell^*(\tilde{A}\tilde{D}p + ADq)\dot{\zeta})$. Therefore

$$\ddot{\nu} = -2\,\mathrm{Re}\,\{\ell^*(\tilde{A}\tilde{D}p + ADq)\dot{\zeta} + \ell^*(\dot{\tilde{A}}\tilde{D}p + \dot{A}Dq)\dot{\zeta} + \ell^*AD\dot{q}\dot{\zeta}\} \ .$$

Differentiating $Wq = A^*\ell$ we find

$$W\dot{q} = \dot{A}^*\ell + A^*\dot{\ell} = D^*A^*\ell\dot{\zeta}^* + A^*\dot{\ell} \ .$$

Differentiating $(AW^{-1}A^*)\ell = -\tilde{A}p$ reveals that

$$\dot{A}W^{-1}A^*\ell + AW^{-1}\dot{A}^*\ell + AW^{-1}A^*\dot{\ell} = -\dot{\tilde{A}}p$$

or

$$ADW^{-1}A^*\ell\dot{\zeta} + AW^{-1}D^*A^*\ell\dot{\zeta}^* + AW^{-1}A^*\dot{\ell} = -\tilde{A}\tilde{D}p\dot{\zeta}$$

so

$$\dot{\ell} = -\,(AW^{-1}A^*)^{-1}\{\tilde{A}\tilde{D}p\dot{\zeta} + ADW^{-1}A^*\ell\dot{\zeta} + AW^{-1}D^*A^*\ell\dot{\zeta}^*\}$$

and

$$\dot{q} = W^{-1}D^*A^*\ell\dot{\zeta}^* + W^{-1}A^*\dot{\ell} \ .$$

Then

$$\ddot{\nu} = \mathrm{Re}(\phi\dot{\zeta}^2) + \psi|\dot{\zeta}|^2 \ ,$$

where

$$\phi = 4q^*WDW^{-1}A^*(AW^{-1}A^*)^{-1}(\tilde{A}\tilde{D}p + ADq) - 2\ell^*(\tilde{A}\tilde{D}^2p + AD^2q) ,$$

$$\psi = -2q^*WDW^{-1}D^*Wq + 2(\tilde{A}\tilde{D}p + ADq)^*(AW^{-1}A^*)^{-1}(\tilde{A}\tilde{D}p + ADq)$$
$$+ 2q^*WDW^{-1}A^*(AW^{-1}A^*)^{-1}AW^{-1}D^*Wq .$$

Thus

$$\ddot{v} = (\mathrm{Re}\,\dot{\zeta} \quad \mathrm{Im}\,\dot{\zeta}) \begin{pmatrix} \psi + \mathrm{Re}\,\phi & -\mathrm{Im}\,\phi \\ -\mathrm{Im}\,\phi & \psi - \mathrm{Re}\,\phi \end{pmatrix} \begin{pmatrix} \mathrm{Re}\,\dot{\zeta} \\ \mathrm{Im}\,\dot{\zeta} \end{pmatrix} . \tag{8.1}$$

The eigenvalues of the matrix are $\psi \pm |\phi|$. If $\psi > |\phi|$ then v is concave upward at ζ. If $|\phi| < -\psi$ then v is concave downward. Other possibilities correspond to more complicated geometries. For instance if $\psi \geq |\phi|$ at a stationary point, the point may be a minimum or a saddle point, depending on the third derivative.

To compute the components comprising $\ddot{v}$ note that

$$(AW^{-1}D^*Wq)_i = \sum_{k=1}^{n+1-i} (n-k)\frac{(n-k)!w_{k+1}}{(n-k-i+1)!w_k} q_{k+1}\zeta^{n-k-i+1}$$

and

$$q^*WDW^{-1}D^*Wq = \sum_{j=1}^{n-1} (n-j)^2 \frac{(w_{j+1})^2}{w_j} |q_{j+1}|^2 .$$

Special Cases for $\ddot{v}$

There are two cases in which the previous expression for $\ddot{v}$ may be simplified. The simplifications will become evident after we prove the

Lemma. $q^*WDW^{-1}D^*Wq = q^*WDW^{-1}A^*(AW^{-1}A^*)^{-1}AW^{-1}D^*Wq$ if and only if $m = n$ or $\ell_{m-1} = 0$.

Proof. (1) If $m = n$ then A is square and invertible so $(AW^{-1}A^*)^{-1} = (A^*)^{-1}WA^{-1}$.

(2) If $\ell_{m-1} = 0$ then $A^*v = D^*A^*\ell$ has a unique solution $v_0 = 0$, $v_1 = \ell_0, \ldots, v_{m-1} = \ell_{m-2}$. Also $\chi(v) = (A^*v - D^*A^*\ell)^*W^{-1}(A^*v - D^*A^*\ell) = 0$. That means that the linear least squares problem

$$W^{-1/2}A^*u = W^{-1/2}D^*A^*\ell = W^{-1/2}D^*Wq$$

has a solution u for which the residual $\chi(u)$ must vanish; otherwise v would be a better solution. In fact, since the rows of A are linearly independent, $u = v$. But there is another expression for u:

$$u = (AW^{-1}A^*)^{-1}AW^{-1}D^*Wq \ .$$

Then $\chi(u) = 0$ implies the desired result.

(3) Assume the hypothesis and that $m < n$; our goal is to show that $\ell_{m-1} = 0$. If we write $B = W^{-1/2}A^*$ then the hypothesis is

$$\ell^*ADW^{-1/2}(1 - BB^{\dagger})W^{-1/2}D^*A^*\ell = 0 \ . \tag{8.2}$$

The theory of the pseudo-inverse implies that $1 - BB^{\dagger}$ is positive semi-definite for any B. Therefore

$$(1 - BB^{\dagger})W^{-1/2}D^*A^*\ell = 0$$

and $D^*A^*\ell = A^*v$ for $v = B^{\dagger}W^{-1/2}D^*A^*\ell$. Since $m < n$ the rows of AD are linearly independent so the equation $\ell^*AD = v^*A$ has a unique solution v. By considering components we find that $v_0 = 0$ and therefore that $\ell_k = v_{k+1}$, $k = 0,1,\ldots,m-2$, and finally that $\ell_{m-1} = 0$, as claimed. Q.E.D.

The next simplification lemma is an easy consequence of the foregoing.

Lemma. If $m = n$ or $\ell_{m-1} = 0$, then

$$\ddot{v} = \mathrm{Re}(\phi\dot{\zeta}^2) + \psi|\dot{\zeta}|^2$$

with

$$\phi = 2\ell^*(\tilde{A}\tilde{D}^2p + AD^2q) ,$$

$$\psi = 2(\tilde{A}\tilde{D}p + ADq)^*(AW^{-1}A^*)^{-1}(\tilde{A}\tilde{D}p + ADq) .$$

Proof. The assertion about ψ is a direct corollary of the previous lemma. To prove the assertion about ϕ requires showing that $\ell^*(\tilde{A}\tilde{D}^2p + AD^2q) = q^*WDW^{-1}A^*(AW^{-1}A^*)^{-1}(\tilde{A}\tilde{D}p + ADq)$.

(1) If $m = n$ then we must show that

$$\ell^*_{m-2}(p+q)^{(n)}(\zeta) = \ell^*ADA^{-1}(\tilde{A}\tilde{D}p + ADq)$$

or

$$\ell^*_{m-2} = \ell^*ADA^{-1}y , \qquad y = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} .$$

But $A^{-1}y = x$ where x represents $(\tau-\zeta)^{n-1}/(n-1)!$. Then

$ADx = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ n! \\ 0 \end{pmatrix}$ so $\ell^*ADx = \ell^*_{m-2}$ as we wished to show.

(2) If $\ell_{m-1} = 0$ we must show that

$$\ell^*_{m-2}(p+q)^{(m)}(\zeta) = q^*WDW^{-1}A^*(AW^{-1}A^*)^{-1}y(p+q)^{(m)}(\zeta)$$

or $\ell^*_{m-2} = u^*y$ for the u^* of the previous lemma. The right hand side further reduces to $u^*_{m-1} = \ell^*_{m-2}$ as we sought to prove.

9. The Last Lagrange Multiplier is Zero at a Minimum

In a previous section we saw that there are two kinds of stationary points for the norm of the distance to the nearest polynomial with an m-tuple zero. Our object is to prove what we asserted then:

Proposition. Let ζ represent a stationary point for $\|q\|$ that is locally minimal with respect to complex perturbations. Then the last Lagrange multiplier $\ell_{m-1} = 0$.

Proof. We know that all stationary points for $\|q\|$ have either $\ell_{m-1} = 0$ or $(p+q)^{(m)}(\zeta) = 0$. Therefore we must show that if $(p+q)^{(m)}(\zeta) = 0$ and $\|q\|$ is locally minimal then $\ell_{m-1} = 0$. To do this we will examine the expression for the second derivative obtained in the previous sections.

The hypothesis, that $\tilde{A}\tilde{D}p + ADq = 0$, implies that

$$\phi = -2\ell^*_{m-1}(p+q)^{(m+1)}(\zeta)$$

and

$$\psi = -2q^*WD\{W^{-1} - W^{-1}A^*(AW^{-1}A^*)^{-1}AW^{-1}\}D^*Wq \ .$$

A minimum requires that $\psi \geq |\phi|$ or

$$-q^*WDW^{-1}\{W - A^*(AW^{-1}A^*)^{-1}A\}W^{-1}D^*Wq \geq |\ell_{m-1}||(p+q)^{(m+1)}(\zeta)| \ .$$

The quantity in $\{\cdot\}$ on the left is $1 - BB^\dagger$ where $B = W^{-1/2}A^*$. $1 - BB^\dagger$ is positive semidefinite for any B, so the left hand side must be ≤ 0. Since the right hand side is ≥ 0, both sides are exactly 0, so

$$q^*WDW^{-1}D^*Wq = q^*WDW^{-1}A^*(AW^{-1}A^*)^{-1}AW^{-1}D^*Wq$$

and

$$\ell^*_{m-1}(p+q)^{(m+1)}(\zeta) = 0 .$$

The first lemma of the last section tells us consequently that either $\ell_{m-1} = 0$, as claimed, or $m = n$. But if $m = n$, then

$$(p+q)^{(m)}(\zeta) = n! \neq 0 ,$$

contrary to the hypothesis that $\tilde{A}\tilde{D}p + ADq = 0$. This concludes the proof as originally worked out by W. Kahan [19].

Thus to find the nearest polynomial with a double zero it is only necessary to solve the simpler equations resulting from the assumption that the last Lagrange multiplier vanishes. In the case of a real polynomial, of course, it may happen that the nearest polynomial with a double zero is a complex polynomial with a complex double zero.

The situation is much more complicated if given a <u>real</u> polynomial, we see the nearest <u>real</u> polynomial with a double zero. Then three possibilities may arise: the nearest such polynomial may have a real double zero, a real triple zero, or a conjugate pair of complex double zeros. The last case is treated in the next chapter. That the second case may arise is illustrated by the following.

<u>Example</u>. Consider the real cubic polynomial whose roots are 1.0 and $.224 \pm .174i$. Let the weights in the usual norm be 1, 1000, and 10000. Then the nearest real polynomial with a double zero is the same as the nearest real polynomial with a triple zero, which is at $\zeta = .4235...$. The second Lagrange multiplier does not vanish at this ζ.

This example does not invalidate the proposition proved earlier in this section. If complex perturbations are allowed, then when double zeros are sought, $\zeta = .4235$ is a saddle point rather than a minimum. The nearest polynomials with double zeros turn out to have $\zeta = .4245 \pm .0993i$, and this ζ may be found by allowing the second Lagrange multiplier to vanish.

The example above was found by accident while searching for something else; see Chapter VI. As a practical matter it seems likely that such examples are quite rare, especially when normal weights are used. In all the other examples we have encountered, it was sufficient to find all the closest polynomials with double zeros and the closest with a complex conjugate pair of double zeros.

10. Another Kind of Second Derivative

In the previous sections we have discussed a directional second derivative for $\nu = q^*Wq$ which we compute by expressing ν as a function of ζ, the m-tuple zero. Another approach, which we could use numerically as a qualitative check on the previous method, is to compute a constrained Hessian matrix of partial second derivatives. In the next two sections we will define this idea and explain how such a matrix may be computed. Then the character of a stationary point may be construed from the signs of the eigenvalues of the constrained Hessian.

Let $f(x) = x^*Hx$ be a scalar function of the vector x. Then how does f vary when x is constrained to the nullspace of a given linear operator L^*? L^* is m by n with $m < n$.

We could choose a transformation P into a subspace of dimension $n - m$ so that the space P^*x satisfies the constraint. Then P^*HP would be the constrained Hessian and its signature would determine the nature of the stationary point.

As far as computational details go, we could let P be composed of columns from the QR factorization of L; see Figure III.1. P of course is not unique. We require L to be of full rank m; that is, none of the constraints are redundant. Then $\tilde{R}$ is invertible and $L^*x = R^*Q^*x = \tilde{R}^*H^*x$, so $L^*x = 0 \Leftrightarrow H^*x = 0$. Thus the columns of P span the space of x satisfying the constraint.

The QR factorization of a real rectangular matrix may be computed using the algorithm decompose in the Wilkinson-Reinsch compendium [35, pp. 113-114]. Q will be computed as a product of m orthogonal reflector matrices $(I - \beta uu^*)$. As each is computed, the corresponding

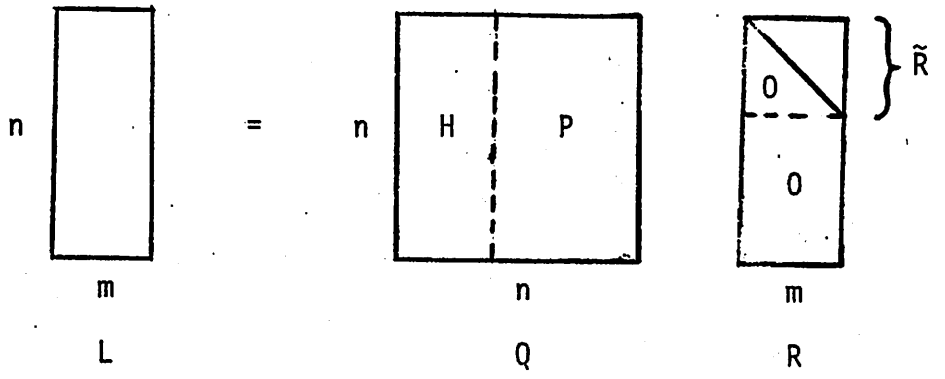

Figure III.1. The QR factorization of L.

similarity may be performed stepwise on H. If a represents a column of H and b* a row, then

$$(I - \beta uu^*)a = a - \beta(u^*a)u \ ,$$

$$b^*(I - \beta uu^*) = b^* - \beta(b^*u)u^* \ .$$

11. Computational Details: A Constrained Hessian for ν

We may apply the technique of the previous section to compute a Hessian matrix for $\nu = q^*Wq$ subject to the constraint $\tilde{A}p + Aq = 0$.

The constrained function to be minimized may be written

$$\Gamma = q^*Wq - \ell^*(\tilde{A}p + Aq)$$

with the Lagrange multipliers ℓ^* treated as independent of the variables q and ζ. Unfortunately the complex variables q appear in the equation non-analytically while the complex variable ζ appears analytically in $\tilde{A}$ and A. Therefore we will divide q, ℓ^*, and ζ into real and imaginary parts to have two sets of constraints:

$$\mathrm{Re}(\tilde{A}p + Aq) = 0$$

and

$$\mathrm{Im}(\tilde{A}p + Aq) = 0 .$$

Writing out the resulting expression for Γ in scalar form,

$$\Gamma = \sum_{j=1}^{n} w_j\{(\mathrm{Re}\ q_j)^2 + (\mathrm{Im}\ q_j)^2\} + \sum_{k=0}^{m-1} \mathrm{Re}\{\lambda_k (p+q)^{(k)}(\zeta)\}$$

where $\lambda_k = \rho_k - i\mu_k$. Then

$$\frac{\partial \Gamma}{\partial\, \mathrm{Re}\, q_j} = 2w_j\, \mathrm{Re}\, q_j + \sum_{k=0}^{m-1} \mathrm{Re}\left(\lambda_k(n,j,k)\zeta^{n-j-k}\right) ,$$

$$\frac{\partial \Gamma}{\partial\, \mathrm{Im}\, q_j} = 2w_j\, \mathrm{Im}\, q_j - \sum_{k=0}^{m-1} \mathrm{Im}\left(\lambda_k(n,j,k)\zeta^{n-j-k}\right) ,$$

$$\frac{\partial \Gamma}{\partial\, \mathrm{Re}\, \zeta} = \sum_{k=0}^{m-1} \mathrm{Re}\left(\lambda_k (p+q)^{(k+1)}(\zeta)\right) ,$$

$$\frac{\partial \Gamma}{\partial\, \mathrm{Im}\, \zeta} = - \sum_{k=0}^{m-1} \mathrm{Im}\left(\lambda_k (p+q)^{(k+1)}(\zeta)\right) ,$$

where $(n,j,k) = (n-j)!/(n-j-k)!$. The second derivatives are

$$\frac{\partial^2\Gamma}{(\partial \operatorname{Re} q_j)^2} = 2w_j = \frac{\partial^2\Gamma}{(\partial \operatorname{Im} q_j)^2} ,$$

$$\frac{\partial^2\Gamma}{(\partial \operatorname{Re} \zeta)^2} = \sum \operatorname{Re}(\lambda_k (p+q)^{(k+2)}(\zeta)) = - \frac{\partial^2\Gamma}{(\partial \operatorname{Im} \zeta)^2} ,$$

$$\frac{\partial^2\Gamma}{\partial \operatorname{Re} \zeta \, \partial \operatorname{Im} \zeta} = - \sum \operatorname{Im}(\lambda_k (p+q)^{(k+2)}(\zeta)) ,$$

$$\frac{\partial^2\Gamma}{\partial \operatorname{Re} q_i \, \partial \operatorname{Im} q_j} = \frac{\partial^2\Gamma}{\partial \operatorname{Re} q_i \, \partial \operatorname{Re} q_j} (i \neq j) = \frac{\partial^2\Gamma}{\partial \operatorname{Im} q_i \, \partial \operatorname{Im} q_j} (i \neq j) = 0 ,$$

$$\frac{\partial^2\Gamma}{\partial \operatorname{Re} q_j \, \partial \operatorname{Re} \zeta} = \sum (n,j,k+1) \operatorname{Re}(\lambda_k \zeta^{n-j-k-1}) = - \frac{\partial^2\Gamma}{\partial \operatorname{Im} q_j \, \partial \operatorname{Im} \zeta} ,$$

$$\frac{\partial^2\Gamma}{\partial \operatorname{Re} q_j \, \partial \operatorname{Im} \zeta} = - \sum (n,j,k+1) \operatorname{Im}(\lambda_k \zeta^{n-j-k-1}) = \frac{\partial^2\Gamma}{\partial \operatorname{Im} q_j \, \partial \operatorname{Re} \zeta} .$$

With these expressions for partial second derivatives we may construct the Hessian matrix H of the previous section. Then the second order change in Γ, for a small change

$$\delta x = \begin{pmatrix} \operatorname{Re} \delta q \\ \operatorname{Im} \delta q \\ \operatorname{Re} \zeta \\ \operatorname{Im} \zeta \end{pmatrix} ,$$

will be $\delta x^T H \delta x$.

The constraints on δx should appear in the matrix L. Those constraints may be found by differentiating $\operatorname{Re}(\tilde{A}p + Aq)$ and $\operatorname{Im}(\tilde{A}p + Aq)$. Then

$$\frac{\partial\ \mathrm{Re}(\tilde{A}p+Aq)}{\partial\ \mathrm{Re}\ q_j} = \mathrm{Re}(Au_j) = \frac{\partial\ \mathrm{Im}(\tilde{A}p+Aq)}{\partial\ \mathrm{Im}\ q_j}\ ,$$

$$\frac{\partial\ \mathrm{Im}(\tilde{A}p+Aq)}{\partial\ \mathrm{Re}\ q_j} = \mathrm{Im}(Au_j) = -\frac{\partial\ \mathrm{Re}(\tilde{A}p+Aq)}{\partial\ \mathrm{Im}\ q_j}\ ,$$

where u_j is the j'th column of the identity matrix. Also

$$\frac{\partial\ \mathrm{Re}(\tilde{A}p+Aq)}{\partial\ \mathrm{Re}\ \zeta} = \mathrm{Re}(\tilde{A}\tilde{D}p + ADq) = \frac{\partial\ \mathrm{Im}(\tilde{A}p+Aq)}{\partial\ \mathrm{Im}\ \zeta}\ ,$$

$$\frac{\partial\ \mathrm{Im}(\tilde{A}p+Aq)}{\partial\ \mathrm{Re}\ \zeta} = \mathrm{Im}(\tilde{A}\tilde{D}p + ADq) = -\frac{\partial\ \mathrm{Re}(\tilde{A}p+Aq)}{\partial\ \mathrm{Im}\ \zeta}\ .$$

Then the matrix L will be $2n+2$ by $2m$ and the matrix H will be $2n+2$ by $2n+2$.

It was necessary to resort to real arithmetic to deal with the non-analytic nature of the function Γ. If, however, we happen to be interested only in real changes in real q and ζ, then the dimensions corresponding to imaginary parts may be omitted, with considerable saving in computational effort to determine the signature of the constrained H.

CHAPTER IV

FINDING THE NEAREST REAL POLYNOMIAL WITH A COMPLEX CONJUGATE PAIR OF m-TUPLE ZEROS

1. Introduction

If we attempt to find the nearest polynomial with an m-tuple zero using the methods of the previous chapter, we sometimes find that one of the stationary points of $\|q\|$ corresponds to a complex m-tuple zero ζ, even if the starting polynomial p is real. Then q turns out to be complex. It might be more reasonable to restrict q to be real if p is real. Then we would find that the nearest real polynomial might have a real m-tuple zero, a real m+1-tuple zero, or a conjugate pair of complex m-tuple zeros.

In the present chapter we will develop the equations to be solved to find the nearest polynomial with a complex conjugate pair of m-tuple zeros. In that development we will take care to divide symbolically by $\operatorname{Im} \zeta$ to eliminate real solutions ζ that we usually do not want. Then we will develop an expression for the second derivative and show that we may assume that the last Lagrange multiplier vanishes, just as in the previous chapter.

2. The Nearest Polynomial with a Complex Conjugate Pair of m-tuple Zeros

Our goal is to minimize $\nu \equiv q^*Wq$ subject to $\tilde{A}p + Aq = 0$ and $\bar{\tilde{A}}p + \bar{A}q = 0$. We assume that the polynomial p is real, but the m-tuple zeros ζ and $\bar{\zeta}$ are complex with $\mathrm{Im}\,\zeta \neq 0$. At first we will not require q or W to be real.

The second constraint may be written $\tilde{A}p + A\bar{q} = 0$ and the constraints together imply $A\,\mathrm{Im}(q) = 0$, since p is real.

As in the previous chapter let ζ vary in a specified direction $\dot{\zeta}$ so $\zeta = \zeta_0 + \theta\dot{\zeta}$, θ real, and thus the directional derivative $\frac{d\zeta}{d\theta}$ is $\dot{\zeta}$. Then $\dot{\nu} = 2\,\mathrm{Re}(q^*W\dot{q})$.

The result of differentiating the constraints is

$$(\tilde{A}\tilde{D}p + ADq)\dot{\zeta} + A\dot{q} = 0$$

and

$$(\tilde{A}\tilde{D}p + AD\bar{q})\dot{\zeta} + A\dot{\bar{q}} = 0\ .$$

Thus if the vector of infinitesimal changes is

$$x = \begin{pmatrix} \mathrm{Re}\,\dot{q} \\ \mathrm{Im}\,\dot{q} \\ \mathrm{Re}\,\dot{\zeta} \\ \mathrm{Im}\,\dot{\zeta} \end{pmatrix},$$

then its constraint is $Cx = 0$, where

$$C = \begin{pmatrix} \mathrm{Re}\,A & -\mathrm{Im}\,A & \mathrm{Re}(\tilde{A}\tilde{D}p+ADq) & -\mathrm{Im}(\tilde{A}\tilde{D}p+ADq) \\ \mathrm{Im}\,A & \mathrm{Re}\,A & \mathrm{Im}(\tilde{A}\tilde{D}p+ADq) & \mathrm{Re}(\tilde{A}\tilde{D}p+ADq) \\ \mathrm{Re}\,A & \mathrm{Im}\,A & \mathrm{Re}(\tilde{A}\tilde{D}p+AD\bar{q}) & -\mathrm{Im}(\tilde{A}\tilde{D}p+AD\bar{q}) \\ \mathrm{Im}\,A & -\mathrm{Re}\,A & \mathrm{Im}(\tilde{A}\tilde{D}p+AD\bar{q}) & \mathrm{Re}(\tilde{A}\tilde{D}p+AD\bar{q}) \end{pmatrix}.$$

Then at a point where ν is stationary with respect to changes in q and ζ satisfying the constraint, $Cx = 0$ implies $y^*x = 0$ where

$$y^* \equiv \begin{pmatrix} \mathrm{Re}(q^*W) & -\mathrm{Im}(q^*W) & 0 & 0 \end{pmatrix} .$$

The notation x, y, and C has been chosen to conform to that of the Lagrange multiplier theorem of Appendix 6. That theorem states that

$$y^* = \begin{pmatrix} r^* & s^* & u^* & v^* \end{pmatrix} C$$

for a vector of Lagrange multipliers $(r^* \; s^* \; u^* \; v^*)$ of length $4m$. Therefore the components of y^* are

$$\mathrm{Re}(q^*W) = (r+u)^* \mathrm{Re}\, A + (s+v)^* \mathrm{Im}\, A , \tag{2.1}$$

$$-\mathrm{Im}(q^*W) = (s-v)^* \mathrm{Re}\, A + (u-r)^* \mathrm{Im}\, A , \tag{2.2}$$

$$\begin{aligned} 0 &= r^* \mathrm{Re}\, a_1 + s^* \mathrm{Im}\, a_1 + u^* \mathrm{Re}\, a_2 + v^* \mathrm{Im}\, a_2 , \\ 0 &= - r^* \mathrm{Im}\, a_1 + s^* \mathrm{Re}\, a_1 - u^* \mathrm{Im}\, a_2 + v^* \mathrm{Re}\, a_2 , \end{aligned} \tag{2.3}$$

where $a_1 = \tilde{A}\tilde{D}p + ADq$ and $a_2 = \tilde{A}\tilde{D}p + AD\bar{q}$.

Recall the formula $q^*W = \ell^*A$ from the previous chapter. The analogous formula now is

$$q^*W = \ell_1^* \mathrm{Re}\, A + \ell_2^* \mathrm{Im}\, A , \tag{2.4}$$

where

$$\begin{aligned} \ell_1^* &= (r+u)^* + i(s-v)^* , \\ \ell_2^* &= (s+v)^* + i(u-r)^* . \end{aligned}$$

Then substituting into the constraints yields

$$\begin{aligned} &\tilde{A}p + AW^{-1}(\mathrm{Re}\ A)^T\ell_1 + AW^{-1}(\mathrm{Im}\ A)^T\ell_2 = 0\ , \\ &\tilde{A}p + \overline{AW^{-1}(\mathrm{Re}\ A)^T\ell_1} + \overline{AW^{-1}(\mathrm{Im}\ A)^T\ell_2} = 0\ . \end{aligned} \tag{2.5}$$

This amounts to $4m$ real equations in $4m+2$ real unknowns, counting $\mathrm{Re}\ \zeta$ and $\mathrm{Im}\ \zeta$. As in the previous chapter, there must be a way of using (2.3) to eliminate some of the unknowns in (2.4).

Instead of pursuing this most general case, let us digress briefly to see what simplifying assumptions might be helpful.

Recall that for a Hermitian W,

$$\begin{aligned} q^*Wq = &(\mathrm{Re}\ q)^T(\mathrm{Re}\ W)(\mathrm{Re}\ q) + (\mathrm{Im}\ q)^T(\mathrm{Re}\ W)(\mathrm{Im}\ q) \\ &- 2(\mathrm{Re}\ q)^T(\mathrm{Im}\ W)(\mathrm{Im}\ q)\ . \end{aligned}$$

If q is real, then q^*Wq is independent of $\mathrm{Im}\ W$ so W might as well be taken to be real. From (2.2) and $A(\mathrm{Im}\ q) = 0$, moreover, we deduce that

$$\begin{aligned} -\mathrm{Im}(q^*W)(\mathrm{Im}\ q) &= 0 \\ &= (\mathrm{Im}\ q)^T(\mathrm{Re}\ W)(\mathrm{Im}\ q) - (\mathrm{Re}\ q)^T(\mathrm{Im}\ W)(\mathrm{Im}\ q)\ . \end{aligned}$$

Consequently if W is real, then $\mathrm{Im}\ q = 0$.

Therefore the simplifying assumption we will make is that W and q are real. Of course, real solutions q are the ones most likely to be of interest when p is real.

Returning to (2.2), with these assumptions we find

$$\begin{aligned} 0 &= (s-v)^*\mathrm{Re}\ A + (u-r)^*\mathrm{Im}\ A \\ &= \begin{pmatrix} s-v \\ u-r \end{pmatrix}^* B \end{aligned}$$

where

$$B \equiv \begin{pmatrix} \text{Re } A \\ \text{Im } A \end{pmatrix} .$$

We shall see in a subsequent section that the rows of B are linearly independent. Therefore $s = v$ and $u = r$, and (2.3) becomes

$$\ell^*(\tilde{A}\tilde{D}p + ADq) = 0 \tag{2.6}$$

for $\ell^* = 2(r^* - is^*)$. (2.4) becomes

$$q^*W = \text{Re}(\ell^*A) . \tag{2.7}$$

(2.6) and (2.7) are the equations for stationarity of real q and complex ζ with respect to complex variations in q and ζ. (2.5) becomes

$$\tilde{A}p + AW^{-1}\text{Re}(A^*\ell) = 0 , \tag{2.8}$$

which is only $2m$ real equations in $2m+2$ real unknowns.

As in Chapter III we might hope to apply (2.6), which implies that either the last Lagrange multiplier vanishes or else the multiplicity of ζ is $m+1$. In a subsequent section we shall see that we may reduce the dimension of (2.8) by one because the last Lagrange multiplier always vanishes at stationary points which are local minima.

Consequently we may assume the last Lagrange multiplier vanishes when solving (2.8), so the problem becomes one of solving $2m$ real equations in $2m$ real unknowns. The equations are linear in the $2m-2$ remaining Lagrange multipliers and very non-linear in $\text{Re } \zeta$ and $\text{Im } \zeta$. So as before we should eliminate the linear variables algebraically and solve for ζ numerically. If ζ were held fixed temporarily and

symbolic Gaussian elimination were attempted on the remaining system of $2m$ linear equations in $2m-2$ unknowns, one would obtain two expressions involving $\mathrm{Re}\ \zeta$ and $\mathrm{Im}\ \zeta$ which would be required to vanish. These last two expressions would be set to zero and solved numerically for $\mathrm{Re}\ \zeta$ and $\mathrm{Im}\ \zeta$.

We will leave the discussion of arbitrary m now and concentrate on the most interesting case, when $m = 2$. In this case (2.8) becomes much simpler. Then

$$\ell = \begin{pmatrix} -\lambda \\ 0 \end{pmatrix}$$

and

$$A = \begin{pmatrix} e^* \\ e^*D \end{pmatrix}$$

so

$$\begin{pmatrix} (\mathrm{Re}\ e^*)W^{-1}(\mathrm{Re}\ e) & -(\mathrm{Re}\ e^*)W^{-1}(\mathrm{Im}\ e) \\ (\mathrm{Im}\ e^*)W^{-1}(\mathrm{Re}\ e) & -(\mathrm{Im}\ e^*)W^{-1}(\mathrm{Im}\ e) \end{pmatrix}\begin{pmatrix} \mathrm{Re}\ \lambda \\ \mathrm{Im}\ \lambda \end{pmatrix} = \begin{pmatrix} \mathrm{Re}\ p(\zeta) \\ \mathrm{Im}\ p(\zeta) \end{pmatrix},$$

$$\begin{pmatrix} (\mathrm{Re}\ e^*D)W^{-1}(\mathrm{Re}\ e) & -(\mathrm{Re}\ e^*D)W^{-1}(\mathrm{Im}\ e) \\ (\mathrm{Im}\ e^*D)W^{-1}(\mathrm{Re}\ e) & -(\mathrm{Im}\ e^*D)W^{-1}(\mathrm{Im}\ e) \end{pmatrix}\begin{pmatrix} \mathrm{Re}\ \lambda \\ \mathrm{Im}\ \lambda \end{pmatrix} = \begin{pmatrix} \mathrm{Re}\ p'(\zeta) \\ \mathrm{Im}\ p'(\zeta) \end{pmatrix}.$$

Written out in detail for the usual W:

$$\begin{pmatrix} \sum(\mathrm{Re}\ \zeta^{n-j})^2/w_j & \sum(\mathrm{Re}\ \zeta^{n-j})(\mathrm{Im}\ \zeta^{n-j})/w_j \\ \sum(\mathrm{Re}\ \zeta^{n-j})(\mathrm{Im}\ \zeta^{n-j})/w_j & \sum(\mathrm{Im}\ \zeta^{n-j})^2/w_j \end{pmatrix}\begin{pmatrix} \mathrm{Re}\ \lambda \\ \mathrm{Im}\ \lambda \end{pmatrix} = \begin{pmatrix} \mathrm{Re}\ p(\zeta) \\ \mathrm{Im}\ p(\zeta) \end{pmatrix},$$

$$\begin{pmatrix} \sum(n-j)(\mathrm{Re}\ \zeta^{n-j})^2/w_j & \sum(n-j)(\mathrm{Re}\ \zeta^{n-j})(\mathrm{Im}\ \zeta^{n-j})/w_j \\ \sum(n-j)(\mathrm{Re}\ \zeta^{n-j})(\mathrm{Im}\ \zeta^{n-j})/w_j & \sum(n-j)(\mathrm{Im}\ \zeta^{n-j})^2/w_j \end{pmatrix}\begin{pmatrix} \mathrm{Re}\ \lambda \\ \mathrm{Im}\ \lambda \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{Re}\ \zeta p'(\zeta) \\ \mathrm{Im}\ \zeta p'(\zeta) \end{pmatrix}.$$

Write these last equations as

$$A_0\Lambda = x_0 \quad \text{and} \quad A_1\Lambda = x_1$$

for matrices A_0, A_1, and vectors Λ, x_0, and x_1. We could solve the equation

$$\tilde{F}(\zeta) \equiv A_0^{-1}x_0 - A_1^{-1}x_1 = 0 ,$$

or

$$\hat{F}(\zeta) \equiv \hat{D}_1 A_0^{\ddagger}x_0 - \hat{D}_0 A_1^{\ddagger}x_1 = 0 , \tag{2.9}$$

where $\ddagger$ denotes the adjoint and $\hat{D}_i$ denotes the determinant $\det(A_i)$; e.g.

$$A_0^{-1} = (\hat{D}_0)^{-1}A_0^{\ddagger} .$$

In the equation $\hat{F}(\zeta) = 0$, we have avoided explicit inverses at the cost of introducing extraneous solutions, by multiplying $\tilde{F}$ by $\hat{D}_0\hat{D}_1$. The equation $\hat{F}(\zeta) = 0$ may be solved trivially by any real ζ, since then the $\hat{D}$ vanish. Since only the complex solutions matter, the real solutions will just be a nuisance that will distract numerical procedures. Therefore we will discuss divided differences in the next sedtion to see whether we can avoid the numerical difficulties.

3. Divided Differences for the Equations for a Complex Conjugate Double Zero

The equation of the previous section

$$\hat{F}(\zeta) \equiv \hat{D}_1 A_0^{\dagger} x_0 - \hat{D}_0 A_1^{\dagger} x_1 = 0$$

has every real ζ among its solutions. The reason for this state of affairs is that $\tilde{F}$, the equation we really wished to solve, was multiplied by $\hat{D}_0\hat{D}_1$. Now

$$\hat{D}_0 = \{\sum(\mathrm{Re}\ \zeta^{n-j})^2/w_j\}\{\sum(\mathrm{Im}\ \zeta^{n-j})^2/w_j\} - \{\sum(\mathrm{Re}\ \zeta^{n-j})(\mathrm{Im}\ \zeta^{n-j})/w_j\}^2 .$$

But $\mathrm{Im}\ \zeta$ divides $\mathrm{Im}\ \zeta^k$ for any $k \geq 0$, as may be simply verified by induction. Therefore we could write

$$\hat{D}_0 = (\mathrm{Im}\ \zeta)^2[\{\sum(\mathrm{Re}\ \zeta^{n-j})^2/w_j\}\{\sum(\Delta_{n-j})^2/w_j\} - \{\sum(\mathrm{Re}\ \zeta^{n-j})\Delta_{n-j}/w_j\}]$$

where the standard divided difference symbol Δ means

$$\Delta_k \equiv \frac{\mathrm{Im}\ \zeta^k}{\mathrm{Im}\ \zeta} = \text{a polynomial in Im } \zeta \text{ and Re } \zeta \ .$$

We could similarly factor out $(\mathrm{Im}\ \zeta)^2$ from $\hat{D}_1$. It turns out, moreover, that for real polynomials p, $\mathrm{Im}\ \zeta$ divides $\mathrm{Im}(p(\zeta))$ and $\mathrm{Im}(\zeta p'(\zeta))$. We may denote these divided differences by Δ_p and $\Delta_{\zeta p'}$. Then $A_0^{\dagger} x_0$ is

$$\begin{pmatrix} (\mathrm{Im}\ \zeta)^2 & 0 \\ 0 & \mathrm{Im}\ \zeta \end{pmatrix} \begin{pmatrix} \sum(\Delta_{n-j})^2/w_j & -\sum(\Delta_{n-j}\,\mathrm{Re}\,\zeta^{n-j})/w_j \\ -\sum(\Delta_{n-j}\,\mathrm{Re}\,\zeta^{n-j})/w_j & \sum(\mathrm{Re}\ \zeta^{n-j})^2/w_j \end{pmatrix} \begin{pmatrix} \mathrm{Re}\ p(\zeta) \\ \Delta_p \end{pmatrix}$$

In all, then, $(\mathrm{Im}\ \zeta)^4$ divides the upper element of the vector $\hat{D}_1 A_0^{\dagger} x_0$ and $(\mathrm{Im}\ \zeta)^3$ divides the lower element. Have we found all possible

Im ζ factors? If we have, the equation will no longer be solved by every real ζ.

To answer the question, let ζ approach a real value. Then as Im $\zeta \to 0$,

$$\Delta_k \to \frac{d}{d\zeta}(\zeta^k) = k\zeta^{k-1} , \qquad \Delta_p \to \frac{d}{d\zeta}p(\zeta) = p'(\zeta) ,$$

$$\Delta_{\zeta p'} \to \frac{d}{d\zeta}(\zeta p'(\zeta)) = \zeta p''(\zeta) + p'(\zeta) .$$

Then when we substitute this information in the equation

$$F(\zeta) = \begin{bmatrix} F_1(\zeta) \\ F_2(\zeta) \end{bmatrix} \equiv \begin{bmatrix} (\text{Im } \zeta)^4 & 0 \\ 0 & (\text{Im } \zeta)^3 \end{bmatrix}^{-1} \hat{F}(\zeta) , \tag{3.1}$$

we find that, for instance,

$$\frac{\zeta^4 F_1(\zeta)}{\sigma_2} = (\sigma_1\sigma_3-\sigma_2^2)p(\zeta) - (\sigma_0\sigma_3-\sigma_1\sigma_2)\zeta p'(\zeta) + (\sigma_0\sigma_2-\sigma_1^2)(\zeta p''(\zeta)+p'(\zeta)) .$$

The right hand side is just the equation (III.7.1) to be solved to find the nearest polynomial with a real triple zero ζ.

Naively we might expect that the limiting case of equation (3.1), an equation for two complex conjugate double zeros, would look like the equation for one real quadruple zero, rather than a triple zero. That such is not the case shows how unreliable intuition can be when applied to these problems!

We may safely conclude, however, that all factors of Im ζ have been removed from (3.1). Ideally, the equation for a real triple zero should also be removed by algebraic means. That removal is such a formidable prospect that it seems more attractive just to numerically

prevent convergence to those real ζ's. Therefore we will solve $F(\zeta) = 0$ with F defined as in (3.1), with the $\operatorname{Im} \zeta$ factors removed symbolically but with convergence to the real triple zeros prevented numerically. The reader interested in the details of computing F may find them in the next few sections.

In the previous chapter we saw that the nearest real polynomial with a triple zero may sometimes also be the nearest real polynomial with a double zero. By numerically deflating the solutions for triple zeros we might be missing some interesting information, but experience has shown that, if the solutions for double zeros are unsatisfactory, then the triple zeros are much more efficiently found by solving the equations for triple zeros rather than allowing the solutions of the equations for complex conjugate pairs to coalesce.

4. Computational Details: The Equations to Solve for a Complex Conjugate Pair of Double Zeros

We aim to find zeros of the function

$$F(\zeta) = \begin{pmatrix} F_1(\zeta) \\ F_2(\zeta) \end{pmatrix} = \begin{pmatrix} (\mathrm{Im}\ \zeta)^4 & 0 \\ 0 & (\mathrm{Im}\ \zeta)^3 \end{pmatrix}^{-1} (\hat{D}_1 A_0^{\ddagger} x_0 - \hat{D}_0 A_1^{\ddagger} x_1) \ .$$

Therefore define

$$D_i = \hat{D}_i/(\mathrm{Im}\ \zeta)^2$$

and

$$\begin{pmatrix} t_i \\ b_i \end{pmatrix} = \begin{pmatrix} 1/(\mathrm{Im}\ \zeta)^2 & 0 \\ 0 & 1/\mathrm{Im}\ \zeta \end{pmatrix} A_i^{\ddagger} x_i$$

for $i = 0, 1$.

Now

$$D_0 = \sum(\mathrm{Re}\ \zeta^{n-j})^2/w_j \sum \Delta_{n-j}^2/w_j - (\sum(\mathrm{Re}\ \zeta^{n-j})\Delta_{n-j}/w_j)^2$$

and D_1 is the same, except $(n-j)/w_j$ replaces $1/w_j$. The formula may be rewritten

$$D_0 = \sum_{j=1}^{n-1} \frac{1}{w_j} \sum_{k=j+1}^{n} \frac{1}{w_k} |\zeta^4|^{n-k} \Delta_{k-j}^2 \ . \tag{4.1}$$

The formulas for the derivatives are

$$\frac{\partial D_0}{\partial\ \mathrm{Re}\ \zeta} = 2|\zeta|^2 \sum_{j=1}^{n-2} \frac{1}{w_j} \sum_{k=j+1}^{n-1} \frac{1}{w_k} \Delta_{k-j} |\zeta^4|^{n-k-1} [2(n-k)\mathrm{Re}\ \zeta\, \Delta_{k-j} + |\zeta|^2 \Delta_{k-j}^r] + \frac{2}{w_n} \sum_{j=1}^{n-1} \frac{1}{w_j} \Delta_{n-j} \Delta_{n-j}^r \ , \tag{4.2a}$$

$$(4.2b)\quad \frac{\partial D_0}{\partial\, \mathrm{Im}\,\zeta} = 2|\zeta|^2 \sum_{j=1}^{n-2} \frac{1}{w_j} \sum_{k=j+1}^{n-1} \frac{1}{w_k} \Delta_{k-j} |\zeta^4|^{n-k-1} [2(n-k)\,\mathrm{Im}\,\zeta\, \Delta_{k-j} + |\zeta|^2 \Delta^m_{k-j}]$$
$$+ \frac{2}{w_n} \sum_{j=1}^{n-1} \frac{1}{w_j} \Delta_{n-j} \Delta^m_{n-j} \,.$$

In the notation of Appendix 4,

$$\Delta^r_k \equiv \frac{\partial \Delta_k}{\partial\, \mathrm{Re}\,\zeta}\,, \qquad \Delta^m_k \equiv \frac{\partial \Delta_k}{\partial\, \mathrm{Im}\,\zeta}\,.$$

Now

$$t_0 = (\textstyle\sum \Delta^2_{n-j}/w_j)\,\mathrm{Re}\; p - (\sum (\mathrm{Re}\;\zeta^{n-j})\Delta_{n-j}/w_j)\Delta_p$$

and

$$\frac{\partial t_0}{\partial\, \mathrm{Re}\,\zeta} = (\textstyle\sum \Delta^2_{n-j}/w_j)\,\mathrm{Re}\; p' + (\mathrm{Re}\; p)\sum (2\Delta_{n-j}\Delta^r_{n-j})/w_j$$
$$- (\textstyle\sum (\mathrm{Re}\;\zeta^{n-j})\Delta_{n-j})(\frac{\partial}{\partial\, \mathrm{Re}\,\zeta}\Delta_p)$$
$$- \Delta_p \textstyle\sum (\mathrm{Re}\;\zeta^{n-j}\Delta^r_{n-j} + (n-j)\Delta_{n-j}\,\mathrm{Re}\,\zeta^{n-j-1})/w_j \;;$$

$$(4.3)\qquad \frac{\partial t_0}{\partial\, \mathrm{Im}\,\zeta} = -(\textstyle\sum \Delta^2_{n-j}/w_j)\,\mathrm{Im}\; p' + (\mathrm{Re}\; p)\sum (2\Delta_{n-j}\Delta^m_{n-j})/w_j$$
$$- (\textstyle\sum (\mathrm{Re}\;\zeta^{n-j})\Delta_{n-j})(\frac{\partial}{\partial\, \mathrm{Im}\,\zeta}\Delta_p)$$
$$- \Delta_p \textstyle\sum (\mathrm{Re}\;\zeta^{n-j}\Delta^m_{n-j} - (n-j)\Delta_{n-j}\,\mathrm{Im}\,\zeta^{n-j-1})/w_j \,.$$

Likewise

$$t_1 = (\textstyle\sum (n-j)\Delta^2_{n-j}/w_j)\,\mathrm{Re}\;\zeta p' - (\sum (n-j)\mathrm{Re}\;\zeta^{n-j}\Delta_{n-j}/w_j)\Delta_{\zeta p'}\,,$$
$$\frac{\partial t_1}{\partial\, \mathrm{Re}\,\zeta} = (\textstyle\sum (n-j)\Delta^2_{n-j}/w_j)\mathrm{Re}(\zeta p'' + p')$$
$$(4.4)\qquad + \mathrm{Re}(\zeta p')\textstyle\sum (n-j)(2\Delta_{n-j}\Delta^r_{n-j})/w_j - (\sum (n-j)\mathrm{Re}\;\zeta^{n-j}\Delta_{n-j})(\frac{\partial}{\partial\, \mathrm{Re}\,\zeta}\Delta_{\zeta p'})$$
$$- \Delta_{\zeta p'} \textstyle\sum (n-j)(\mathrm{Re}\;\zeta^{n-j}\Delta^r_{n-j} + (n-j)\Delta_{n-j}\,\mathrm{Re}\,\zeta^{n-j-1})/w_j \,.$$

The expression for $\frac{\partial t_1}{\partial \operatorname{Im}\zeta}$ may be obtained similarly by substituting $(n-j)/w_j$ for $1/w_j$ and $\zeta p'$ for p in the expression for $\frac{\partial t_0}{\partial \operatorname{Im}\zeta}$.

Continuing in the same fashion,

$$
\begin{aligned}
b_0 &= -\left(\sum \operatorname{Re}\zeta^{n-j}\Delta_{n-j}/w_j\right)\operatorname{Re} p + \left(\sum(\operatorname{Re}\zeta^{n-j})^2/w_j\right)\Delta_p\ , \\
b_1 &= -\left(\sum(n-j)\operatorname{Re}\zeta^{n-j}\Delta_{n-j}/w_j\right)\operatorname{Re}\zeta p' + \left(\sum(n-j)(\operatorname{Re}\zeta^{n-j})^2/w_j\right)\Delta_{\zeta p'}\ , \\
\frac{\partial b_0}{\partial \operatorname{Re}\zeta} &= -\left(\sum \operatorname{Re}\zeta^{n-j}\Delta_{n-j}/w_j\right)\operatorname{Re} p' \\
&\quad - (\operatorname{Re} p)\sum(\operatorname{Re}\zeta^{n-j}\Delta^r_{n-j} + (n-j)\Delta_{n-j}\operatorname{Re}\zeta^{n-j-1})/w_j \\
&\quad + \left(\sum(\operatorname{Re}\zeta^{n-j})^2/w_j\right)\frac{\partial \Delta_p}{\partial \operatorname{Re}\zeta} + \Delta_p \sum 2(n-j)\operatorname{Re}\zeta^{n-j}\operatorname{Re}\zeta^{n-j-1}/w_j \\
\frac{\partial b_0}{\partial \operatorname{Im}\zeta} &= \left(\sum \operatorname{Re}\zeta^{n-j}\Delta_{n-j}/w_j\right)\operatorname{Im} p' \\
&\quad - (\operatorname{Re} p)\sum(\operatorname{Re}\zeta^{n-j}\Delta^m_{n-j} - (n-j)\Delta_{n-j}\operatorname{Im}\zeta^{n-j-1})/w_j \\
&\quad + \left(\sum(\operatorname{Re}\zeta^{n-j})^2/w_j\right)\frac{\partial \Delta_p}{\partial \operatorname{Im}\zeta} - \Delta_p \sum 2(n-j)\operatorname{Re}\zeta^{n-j}\operatorname{Im}\zeta^{n-j-1}/w_j\ .
\end{aligned}
\tag{4.5}
$$

The formulas for the derivatives of b_1 can be obtained by the usual substitutions.

The formulas in this section may be used to implement Newton's method to solve the two real equations $F_1(\zeta) = 0$ and $F_2(\zeta) = 0$ for their two real unknowns $\operatorname{Re}\zeta$ and $\operatorname{Im}\zeta$.

5. The Rows of B are Linearly Independent

Corresponding to the complex operator A of previous chapters,

$$A \equiv \begin{pmatrix} e^* \\ e^*D \\ \vdots \\ e^*D^{m-1} \end{pmatrix},$$

it was necessary in Section 2 to define the real operator B which maps R^n to R^{2m} by

$$B = \begin{pmatrix} \operatorname{Re} A \\ \operatorname{Im} A \end{pmatrix} = \left.\begin{pmatrix} \operatorname{Re} e^* \\ \vdots \\ \operatorname{Re} e^*D^{m-1} \\ \operatorname{Im} e^* \\ \vdots \\ \operatorname{Im} e^*D^{m-1} \end{pmatrix}\right\} 2m$$

$$\underbrace{\qquad\qquad}_{n}$$

Proposition. If $\operatorname{Im} \zeta \neq 0$ then the rows of B are linearly independent.

Corollary. $BW^{-1}B^T$ is invertible.

Proof of Proposition. We will show that B has full rank 2m by exhibiting a set of real vectors $\{q_{kr},\ 0 \leq k \leq m-1\}$, such that

$$\{Bq_{kr}\} = \left\{ \left\{ \begin{pmatrix} 1 \\ x \\ x \\ x \\ 0 \\ x \\ x \\ x \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ x \\ x \\ 0 \\ 0 \\ x \\ x \end{pmatrix}, \ldots \right\} \begin{matrix} \} m \\ \} m \end{matrix} \right.,$$

and a set $\{q_{km}\}$ such that

$$\{Bq_{km}\} = \left\{ \begin{pmatrix} 0 \\ x \\ x \\ x \\ 1 \\ x \\ x \\ x \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ x \\ x \\ 0 \\ 1 \\ x \\ x \end{pmatrix}, \ldots \right\} .$$

In other words,

$$q_{kr}{}^{(j)}(\alpha) = \begin{cases} 0 , & 0 \le j \le k-1 , \\ 1 , & j = k , \end{cases}$$

and

$$q_{km}{}^{(j)}(\alpha) = \begin{cases} 0 , & 0 \le j \le k-1 , \\ i , & j = k . \end{cases}$$

The existence of $2m$ such real q_k's is equivalent to the linear independence of the rows of B.

Clearly, for either set of q_k,

$$q_k(\tau) = (\tau-\alpha)^k(\tau-\bar{\alpha})^k s(\tau)$$

for some real $s(\tau)$ with $s(\alpha) \neq 0$. Obviously $q_k{}^{(j)}(\alpha) = 0$ for $0 \le j \le k-1$. Furthermore $q_k{}^{(k)}(\alpha) = \phi \neq 0$. If ϕ were real it would suffice to let $s(\tau) = 1/\phi$. But what if ϕ is complex?

It turns out that $s(\tau) = \theta\tau + \eta$ with real θ and η to be determined. To see this we must examine $q_k{}^{(k)}(\alpha)$. First form an expression for $(\tau-\bar{\alpha})^k$:

$$(\tau-\bar{\alpha})^k = (\tau - \alpha + 2i \operatorname{Im}\alpha)^k = \sum_{j=0}^{k} \binom{k}{j} (\tau-\alpha)^j (2i \operatorname{Im}\alpha)^{k-j} ,$$

by the binomial theorem.

Then

$$(\tau-\alpha)^k(\tau-\bar{\alpha})^k = \sum_{j=0}^{k} \binom{k}{j}(\tau-\alpha)^{k+j}(2i\,\mathrm{Im}\,\alpha)^{k-j}$$

and

$$\frac{d^r}{d\tau^r}\{(\tau-\alpha)^k(\tau-\bar{\alpha})^k\} = \sum_{j=\max(0,r-k)}^{k} \binom{k}{j}\frac{(k+j)!}{(k+j-r)!}(\tau-\alpha)^{k+j-r}(2i\,\mathrm{Im}\,\alpha)^{k-j}$$

and

$$\frac{d^r}{d\tau^r}\{(\tau-\alpha)^k(\tau-\bar{\alpha})^k\}\Big|_{\tau=\alpha} = \begin{cases} 0\ , & r < k\ , \\ k!(2\,\mathrm{Im}\,\alpha)^k\cdot i^k\ , & r = k\ . \end{cases}$$

We may now invoke Leibniz' rule,

$$D^k(ps) = \sum_{j=0}^{k} \binom{k}{j}(D^{k-j}p)(D^k s)\ ,$$

to find

$$\frac{d^k}{d\tau^k}\{(\tau-\alpha)^k(\tau-\bar{\alpha})^k s(\tau)\}\Big|_{\tau=\alpha} = k!(2\,\mathrm{Im}\,\alpha)^k\cdot i^k s(\alpha)\ .$$

This expression for $q_k^{(k)}(\alpha)$ shows that it is only necessary to choose an appropriate real s of degree at most 1 to get any desired complex value of $q_k^{(k)}(\alpha)$. If ω is the desired complex value of $s(\alpha)$ then

$$\mathrm{Re}\ s(\alpha) = \mathrm{Re}(\theta\alpha+\eta) = \theta\,\mathrm{Re}\,\alpha + \eta = \mathrm{Re}\ \omega\ ;$$
$$\mathrm{Im}\ s(\alpha) = \theta\,\mathrm{Im}\,\alpha = \mathrm{Im}\ \omega\ .$$

Thus $\theta = \mathrm{Im}\ \omega/\mathrm{Im}\ \alpha$, $\eta = \mathrm{Re}\ \omega - \theta\,\mathrm{Re}\,\alpha$, so we can construct s and therefore each q_{kr} and q_{km}. So the rows of B are linearly independent as claimed.

6. The Last Lagrange Multiplier is Zero

Section 2 demonstrates that there are two kinds of stationary points for $\nu = q^*Wq$, q real, namely those for which the last Lagrange multiplier vanishes, and those for which the multiplicity is greater than anticipated, so that $(p+q)^{(m)}(\zeta) = 0$.

Proposition. Let ζ represent a stationary point for $\|q\|$ that is locally minimal with respect to complex perturbations of ζ. Then the last Lagrange multiplier $\ell_{m-1} = 0$.

Proof. Since $\nu = q^*Wq$, $\dot{\nu} = 2\,\mathrm{Re}\,(q^*W\dot{q})$. But $Wq = \mathrm{Re}(A^*\ell)$ for a complex vector ℓ of Lagrange multipliers. Therefore

$$\dot{\nu} = 2\,\mathrm{Re}\,(\ell^*A\dot{q}) = -2\,\mathrm{Re}\,(\ell^*(\tilde{A}\tilde{D}p + ADq)\dot{\zeta})$$

because of the constraint $\tilde{A}p + Aq = 0$. Then

$$\ddot{\nu} = -2\,\mathrm{Re}\,\{\dot{\ell}^*(\tilde{A}\tilde{D}p+ADq)\dot{\zeta} + \ell^*(\tilde{A}\tilde{D}^2p+AD^2q)\dot{\zeta}^2 + \ell^*AD\dot{q}\dot{\zeta}\} \quad . \tag{6.1}$$

Assume now that we are at one of the stationary points with $\tilde{A}\tilde{D}p + ADq = 0$. Our next task is to obtain expressions for $\dot{q}$ and $\dot{\ell}$. Differentiating the constraint reveals that

$$(\tilde{A}\tilde{D}p + ADq)\dot{\zeta} + A\dot{q} = 0 \ , \qquad \text{so} \qquad A\dot{q} = 0 \ ,$$

while differentiating the stationarity condition $Wq = A^*\ell$ yields

$$W\dot{q} = \mathrm{Re}\{A^*\dot{\ell} + D^*A^*\ell\bar{\dot{\zeta}}\}$$

so

$$A\dot{q} = AW^{-1}\,\mathrm{Re}\,\{A^*\dot{\ell} + D^*A^*\ell\bar{\dot{\zeta}}\} = 0 \ .$$

From

$$AW^{-1}\,\mathrm{Re}\,(A^*\dot{\ell}) = -\,AW^{-1}\,\mathrm{Re}\,(D^*A^*\ell\dot{\bar{\xi}})\ ,$$

deduce that

$$BW^{-1}B^T\begin{pmatrix}\mathrm{Re}\ \dot{\ell}\\ \mathrm{Im}\ \dot{\ell}\end{pmatrix} = -B\,\mathrm{Re}\,(W^{-1}D^*A^*\ell\dot{\bar{\xi}})\ ,$$

where

$$B = \begin{pmatrix}\mathrm{Re}\ A\\ \mathrm{Im}\ A\end{pmatrix}$$

as in previous sections. Since the rows of B are linearly independent, $BW^{-1}B^T$ is positive definite and

$$\begin{pmatrix}\mathrm{Re}\ \dot{\ell}\\ \mathrm{Im}\ \dot{\ell}\end{pmatrix} = -\,(BW^{-1}B^T)^{-1}B\,\mathrm{Re}\,(W^{-1}D^*A^*\ell\dot{\bar{\xi}})\ , \tag{6.2}$$

$$\dot{q} = W^{-1}\,\mathrm{Re}\,(A^*\dot{\ell} + D^*A^*\ell\dot{\bar{\xi}})\ .$$

Then

$$\mathrm{Re}(A^*\dot{\ell}) = -\,B^T(BW^{-1}B^T)^{-1}B\,\mathrm{Re}\,(W^{-1}D^*A^*\ell\dot{\bar{\xi}})$$

and

$$\dot{q} = W^{-1}(W - B^T(BW^{-1}B^T)^{-1}B)W^{-1}D^*\,\mathrm{Re}\,(A^*\ell\dot{\bar{\xi}})\ . \tag{6.3}$$

Recall that

$$\ddot{v} = -2\,\mathrm{Re}\,(\ell^*(\tilde{A}\tilde{D}^2p + AD^2q)\dot{\zeta}^2) - 2\,\mathrm{Re}\,(\ell^*AD\dot{q}\dot{\zeta})\ .$$

We may write

$$\mathrm{Re}(\ell^*AD\dot{q}\dot{\zeta}) = \mathrm{Re}(\dot{\zeta}\ell^*A)DW^{-1}(W - B^T(BW^{-1}B^T)^{-1}B)W^{-1}D^*\,\mathrm{Re}\,(A^*\ell\dot{\bar{\xi}})\ . \tag{6.4}$$

The matrix $(W - B^T(BW^{-1}B^T)^{-1}B)$ is positive semidefinite so both sides are real and ≥ 0.

As in the previous chapter we may write

$$\ddot{v} = -2(\operatorname{Re}\dot{\zeta}\ \operatorname{Im}\dot{\zeta})\begin{pmatrix}\operatorname{Re}\phi + (\operatorname{Re} b)^T M(\operatorname{Re} b) & -\operatorname{Im}\phi + (\operatorname{Im} b)^T M(\operatorname{Re} b)\\ -\operatorname{Im}\phi + (\operatorname{Im} b)^T M(\operatorname{Re} b) & -\operatorname{Re}\phi + (\operatorname{Im} b)^T M(\operatorname{Im} b)\end{pmatrix}\begin{pmatrix}\operatorname{Re}\dot{\zeta}\\ \operatorname{Im}\dot{\zeta}\end{pmatrix}$$

where

$$b = D^*A^*\ell\ ,$$

$$M = W^{-1/2}(1 - (W^{-1/2}B)[(W^{-1/2}B)^T(W^{-1/2}B)]^{-1}(W^{-1/2}B)^T)W^{-1/2}\ ,$$

and $\phi = \ell^*(\tilde{A}\tilde{D}^2p + AD^2q)$.

Then a tedious but straightforward argument paralleling that of Section III.9 shows that $\ddot{v} \geq 0$ for all $\dot{\zeta}$ implies $\phi = 0$.

Alternatively we may recognize that for a suitable $\dot{\zeta}$,

$$\ddot{v} = -2\{|\ell^*(\tilde{A}\tilde{D}^2p + AD^2q)\dot{\zeta}^2| + \operatorname{Re}(\ell^*AD\dot{q}\dot{\zeta})\}\ .$$

At a local minimum $\ddot{v} \geq 0$ for all $\dot{\zeta}$; recall (6.4) to see that $\ell^*(\tilde{A}\tilde{D}^2p + AD^2q) = 0$ and also $\ell^*AD\dot{q} = 0$.

Thus by either argument, at a stationary point which is also a minimum, $\ell_{m-1} = 0$ or $(p+q)^{(m+1)}(\zeta) = 0$. In the first case we are finished. The second case implies that $n \geq 2(m+2)$.

Furthermore, $\ell^*AD\dot{q} = 0$ and (6.3) tell us that

$$\operatorname{Re}(\ell^*A)DW^{-1}\{W - B^T(BW^{-1}B^T)^{-1}B\}W^{-1}D^*\operatorname{Re}(A^*\ell) = 0\ ,$$

Since the matrix in brackets is positive semidefinite,

$$\{W - B^T(BW^{-1}B^T)^{-1}B\}W^{-1}D^*\operatorname{Re}(A^*\ell) = 0$$

and

$$\operatorname{Re}(\ell^*A)D = s^TB$$

where

$$s^T = \mathrm{Re}(\ell^* A) D W^{-1} B^T (B W^{-1} B^T)^{-1}$$

so s^T is real.

Our next goal is to construct a matrix like B, but augmented by two more rows, from which we can conclude the result. Partition ℓ^* and s^T as follows:

$$\ell^* = (\hat{\ell}, \lambda)^* ,$$
$$s^T = (\mu \;\; u \;\; \theta \;\; v)^T .$$

λ, μ, and θ are scalars. Then

$$\mathrm{Re}(\ell^* A) = (\mathrm{Re}\ \ell)^T \mathrm{Re}\, A - (\mathrm{Im}\ \ell)^T \mathrm{Im}\, A$$

and

$$\mathrm{Re}(\ell^* A) D = (\mathrm{Re}\,\hat{\ell} \;\; \mathrm{Re}\,\lambda \;\; -\mathrm{Im}\,\hat{\ell} \;\; -\mathrm{Im}\,\lambda)^T \begin{pmatrix} \mathrm{Re}\ AD \\ \mathrm{Im}\ AD \end{pmatrix} .$$

Finally let

$$\hat{A} = \begin{pmatrix} e^* D \\ \vdots \\ e^* D^{m-1} \end{pmatrix}$$

so

$$A = \begin{pmatrix} e^* \\ \hat{A} \end{pmatrix}$$

and

$$AD = \begin{pmatrix} \hat{A} \\ e^* D^m \end{pmatrix} .$$

Then

$$\mathrm{Re}(\ell^* A) D - s^T B = 0$$

may be written

$$(\mu, \operatorname{Re}\hat{\ell}^T - u^T, \operatorname{Re}\lambda, \theta, -\operatorname{Im}\hat{\ell}^T - v^T, -\operatorname{Im}\lambda) \begin{pmatrix} \operatorname{Re} e^* \\ \operatorname{Re} \hat{A} \\ \operatorname{Re} e^* D^m \\ \operatorname{Im} e^* \\ \operatorname{Im} \hat{A} \\ \operatorname{Im} e^* D^m \end{pmatrix} = 0 .$$

The matrix on the right is just a B matrix, but for m augmented by 1. Since $n \geq 2(m+2)$, the augmented matrix has at most n rows which are linearly independent. Consequently $\hat{\ell} = u - iv$, $\theta = 0$, $\mu = 0$, and $\lambda = 0$. But this λ is just the last Lagrange multiplier ℓ_{m-1}, concluding the proof.

We learned in the previous chapter that to find the nearest real polynomial with a real double zero, it might be necessary to solve equations for a real double zero and equations for a real triple zero. But in this chapter we have the more satisfactory result that to find the nearest real polynomial with a complex conjugate pair of double zeros, we need solve only one set of equations; it is not necessary to look for the nearest real polynomial with a complex conjugate pair of triple zeros.

CHAPTER V

FINDING THE NEAREST POLYNOMIAL WITH MORE THAN ONE MULTIPLE ZERO

1. Introduction

Previous chapters have exhibited the equations to be solved to find the nearest polynomial with one multiple zero or one pair of complex conjugate multiple zeros. Now we turn to the more general problem of finding the nearest polynomial with a specified configuration of multiple zeros. We shall see that despite some complications the theory bears a family resemblance to what has gone before. We shall find that, in the complex case, the equations to be solved for the multiple zeros assume forms simpler than what might have been expected, because certain Lagrange multipliers vanish. However there is some doubt, in general, as to which of these simpler equations should be solved for the multiple zeros. Fortunately when all the zeros are double the equations to solve are fairly obvious.

Unfortunately, just as in the case of the complex conjugate multiple zeros, the equations we solve become much more complicated when divided differences are taken in order to inhibit unwanted coalescence of the multiple zeros. These equations are given in full detail for the case of several double zeros, and especially for the case of two double zeros.

2. The Nearest Polynomial with Several Multiple Zeros

Given a complex polynomial $p(\tau)$ we seek the nearest polynomial $(p+q)(\tau)$ such that $p+q$ has k complex multiple zeros ζ_i. Each ζ_i has a multiplicity $m_i \geq 2$, and $\sum_i m_i \leq n$. Corresponding to the operator A of previous chapters we define A_i by

$$A_i = \underbrace{\left.\begin{bmatrix} e_i^* \\ e_i^* D \\ \vdots \\ e_i^* D^{m_i-1} \end{bmatrix}\right\} m_i}_{n}$$

e_i^* is the evaluation functional for ζ_i. The m_i by $n+1$ operator $\tilde{A}_i$ is defined analogously with $\tilde{e}_i^*$ replacing e_i^*. Then the equation

$$\tilde{A}_i p + A_i q = 0$$

expresses the constraint that $p+q$ has an m-tuple zero ζ_i.

We also define the operator

$$S = \underbrace{\left.\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_k \end{bmatrix}\right\} \sum m_i}_{n}$$

which may be seen to be somewhat like the B of the previous chapter; it will be used for similar purposes.

Proposition. If $\zeta_i \neq \zeta_j$ when $i \neq j$ then the rows of S are linearly independent.

Corollary. $SW^{-1}S^*$ is invertible.

Proof of Proposition. We will show that S has full rank by displaying $\sum m_i$ linearly independent vectors

$$Sq_{j,r}\ , \quad 1 \le j \le k, \quad 0 \le r \le m_j - 1\ .$$

The $q_{j,r}$ are defined by their corresponding polynomials as

$$q_{j,r}(\tau) = (\tau-\zeta_j)^r \prod_{i\neq j} (\tau-\zeta_i)^{m_i}$$

and the conclusion follows immediately.

Our goal is to minimize $\nu = q^*Wq$ subject to $\tilde{A}_i p + A_i q = 0$, $1 \le i \le k$. Let the raised dot ($\dot{}$) represent differentiation in a particular direction of a specific ζ_j: $\zeta_j(\theta) = \zeta_j(0) + \theta\dot{\zeta}_j$. Then as usual

$$\dot{\nu} = \frac{d\nu}{d\theta} = 2\,\mathrm{Re}\,(q^*W\dot{q})\ .$$

Differentiate the j'th constraint to find

$$(\tilde{A}_j\tilde{D}p + A_jDq)\dot{\zeta}_j + A_j\dot{q} = 0\ ,$$

but differentiate the other $k-1$ constraints to find

$$A_i\dot{q} = 0\ , \quad i \neq j\ ,$$

because A_i is independent of ζ_j.

By applying the Lagrange multiplier theorem of Appendix 6 at a stationary point, discover in the usual way that

$$q^*W = \sum_i \ell_i^* A_i = \hat{\ell}^* S\ . \tag{2.1}$$

There are k vectors ℓ_i^* of Lagrange multipliers and $\hat{\ell}^*$ is their concatenation. Furthermore

$$\ell_j^*(\tilde{A}_j\tilde{D}p + A_jDq) = 0 \tag{2.2}$$

for $1 \leq j \leq k$.

Thus at a stationary point, for each j, either its last Lagrange multiplier ℓ^*_{j,m_j-1} vanishes or ζ_j has multiplicity one greater than expected. In the next section we will see how the techniques of previous chapters can be applied to show that the minima of ν always have $\ell^*_{j,m_j-1} = 0$.

Now when we substitute in the constraints we find

$$\tilde{A}_ip + A_iW^{-1}S^*\hat{\ell} = 0 \ , \quad 1 \leq i \leq k \ ,$$

or

$$SW^{-1}S^*\hat{\ell} = -\tilde{S}p \ . \tag{2.3}$$

Since the rows of S are linearly independent, $SW^{-1}S^*$ is positive definite symmetric and therefore invertible. But we may assume that k elements of $\hat{\ell}$ vanish, so we have $\sum m_i$ linear equations in $(\sum m_i) - k$ unknowns. The attempt to solve such a system by Gaussian elimination yields k expressions which must vanish. The corresponding k non-linear equations in the ζ_i may in principle be solved for the ζ_i. In subsequent sections we will display equations for the case that all $m_i = 2$.

3. The Last Lagrange Multipliers are Zero

From the previous section we may deduce that

$$\dot{\nu} = 2\ \mathrm{Re}(q^*W\dot{q}) = 2\ \mathrm{Re}(\hat{\ell}^*S\dot{q}) = -2\ \mathrm{Re}(\ell_j^*(\tilde{A}_j\tilde{D}p+A_jDq)\dot{\zeta}_j)\ .$$

When ν is stationary, then for each j, either its last Lagrange multiplier vanishes or the multiplicity of ζ_j is one greater than expected.

Proposition. Assume $i \neq j \Rightarrow \zeta_i \neq \zeta_j$. Then at a stationary point at which ν is minimal with respect to complex perturbations in ζ_j, the last Lagrange multiplier in ℓ_j^* vanishes.

Proof. Continue to differentiate the expression for $\dot{\nu}$ above:

$$\ddot{\nu} = -2\ \mathrm{Re}\{\dot{\ell}_j^*(\tilde{A}_j\tilde{D}p+A_jDq)\dot{\zeta}_j + \ell_j^*(\tilde{A}_j\tilde{D}^2p+A_jD^2q)\dot{\zeta}_j^2 + \ell_j^*A_jD\dot{q}\dot{\zeta}\}\ .$$

Assume that $\tilde{A}_j\tilde{D}p + A_jDq = 0$ at a stationary point, which simplifies the expression for $\ddot{\nu}$ above. Furthermore the assumption means that $\sum m_i < n$ because $k \geq 2$ and all ζ_i's are distinct.

From (2.1),

$$\dot{q} = W^{-1}\dot{S}^*\hat{\ell} + W^{-1}S^*\dot{\hat{\ell}}\ ,$$

and from the constraint and the assumption,

$$S\dot{q} = 0\ .$$

Therefore

$$SW^{-1}S^*\dot{\hat{\ell}} = -\ SW^{-1}\dot{S}^*\hat{\ell}\ .$$

But

$$\dot{S}^*\hat{\ell} = \sum\dot{A}_i^*\ell_i = D^*A_j^*\ell_j\dot{\bar{\zeta}}_j\ ,$$

so

$$\dot{\hat{\ell}} = - (SW^{-1}S^*)^{-1}SW^{-1}D^*A_j^*\ell_j\dot{\bar{\zeta}}_j ,$$
$$\dot{q} = W^{-1}D^*A_j^*\ell_j\dot{\bar{\zeta}}_j - W^{-1}S^*(SW^{-1}S^*)^{-1}SW^{-1}D^*A_j^*\ell_j\dot{\bar{\zeta}}_j ,$$

and

$$\ell_j^*A_jD\dot{q}\dot{\zeta} = \ell_j^*A_jDW^{-1/2}\{1 - (W^{-1/2}S^*)(W^{-1/2}S^*)^\dagger\}W^{-1/2}D^*A_j^*\ell_j|\dot{\zeta}_j|^2 .$$

$(1 - MM^\dagger)$ is positive semidefinite for any M so

$$\ddot{v} = - 2\,\mathrm{Re}\{\ell_j^*(\tilde{A}_j\tilde{D}^2p + A_jD^2q)\dot{\zeta}_j^2\}$$
$$- 2(\ell_j^*A_jDW^{-1/2}\{1 - (W^{-1/2}S^*)(W^{-1/2}S^*)^\dagger\}W^{-1/2}D^*A_j^*\ell_j)|\dot{\zeta}_j|^2 .$$

If v is to have a local minimum then $\ddot{v} \geq 0$ for any $\dot{\zeta}_j$, yet by apt choice of $\dot{\zeta}_j$ we may arrange for both terms to be real and negative, so they both must vanish:

$$\lambda_j^*(p+q)^{(m_j+1)}(\zeta_j) = 0$$

and

$$\ell_j^*A_jDW^{-1/2}\{1 - (W^{-1/2}S^*)(W^{-1/2}S^*)^\dagger\} = 0 . \tag{3.1}$$

From this point we follow the argument of III.8 to show that λ_j^*, the last element of ℓ_j^*, vanishes. From (3.1) we find

$$\ell_j^*A_jD = v^*S$$

where $v^* = \ell_j^*A_jDW^{-1/2}((W^{-1/2}S^*)^\dagger)^*$. Now partition v^* conformally with S so

$$v^*S = \sum v_i^*A_i .$$

Introduce an augmented operator

$$\hat{S} \equiv \underbrace{\left.\begin{bmatrix} A_1 \\ \vdots \\ A_j \\ e_j^* D^{m_j} \\ A_{j+1} \\ \vdots \\ A_k \end{bmatrix}\right\} \textstyle\sum m_i + 1}_{n} \quad .$$

Then we may rewrite the equation

$$v^*S - \ell_j^* A_j D = 0$$

as

$$(v_1^* \cdots v_{j-1}^* \; \hat{v}_j^* \; v_{j+1}^* \cdots v_k^*)\hat{S} = 0 \; , \tag{3.2}$$

where $\hat{v}_j^* = (v_j^* \; 0) - (0 \; \ell_j^*)$. Since $\sum m_i < n$, the rows of $\hat{S}$ are linearly independent, so the vector in (3.2) vanishes. In particular, the last element of $\hat{v}_j^*$, which is $-\lambda_j^*$, the last Lagrange multipler, vanishes as claimed, completing the proof.

As in Chapter III, the present result applies when complex perturbations are considered. In the case of real perturbations of a real polynomial, the result is known to be false in general for $k = 1$ and counterexamples could probably be constructed for larger k. It seems likely, however, that in most practical problems satisfactory results may be obtained by assuming that the last Lagrange multipliers vanish.

4. Equations for k Real Double Zeros

The nearest polynomial with k real double zeros is of interest in studying polynomials like Wilkinson's (Chapter X). The formulas we shall derive have not been treated by means of divided differences. Section 6 contains formulas for the case $k = 2$ derived with the aid of divided differences.

The equation we wish to solve is (2.3);

$$SW^{-1}S^*\hat{\ell} = -\tilde{S}p .$$

We know that the last elements vanish for each ℓ_i, a subvector of $\hat{\ell}$. Therefore we may define the vector Λ by letting Λ_i be the first element of ℓ_i. Then

$$S^*\hat{\ell} = \sum A_i^*\ell_i = \sum \Lambda_i e_i .$$

Recall that e_i^* is the evaluation functional for ζ_i.

Having eliminated some of the unknowns we are left with $2k$ equations in the $2k$ variables $\Lambda_1,\Lambda_2,\ldots,\Lambda_k$ and $\zeta_1,\zeta_2,\ldots,\zeta_k$. Since the equations are linear in the Λ_i's we can easily eliminate them, leaving k non-linear equations in the ζ's. To do this divide the equation (2.3) into two pieces:

$$S_0W^{-1}S_0^*\Lambda = -\tilde{S}_0p$$

and

$$S_1W^{-1}S_0^*\Lambda = -\tilde{S}_1p \tag{4.1}$$

where

$$S_0 = \begin{pmatrix} \vdots \\ e_i^* \\ \vdots \end{pmatrix} \quad \text{and} \quad S_1 = \begin{pmatrix} \vdots \\ e_i^* D \\ \vdots \end{pmatrix} .$$

To simplify matters later multiply (4.1) by the matrix $Z = \operatorname{diag}(\zeta_1,\ldots,\zeta_k)$. Then if we define

$$T_0 = S_0 W^{-1} S_0^* ,$$

$$T_1 = Z S_1 W^{-1} S_0^* ,$$

$$v_0 = \tilde{S}_0 p = \begin{pmatrix} p(\zeta_1) \\ \vdots \\ p(\zeta_k) \end{pmatrix} ,$$

$$v_1 = Z\tilde{S}_1 p = \begin{pmatrix} \zeta_1 p'(\zeta_1) \\ \vdots \\ \zeta_k p'(\zeta_k) \end{pmatrix} ,$$

where

$$(T_0)_{ij} = e_i^* W^{-1} e_j \quad \text{and} \quad (T_1)_{ij} = \zeta_i e_i^* D W^{-1} e_j ,$$

then we may eliminate Λ and try to find zeros of the function

$$F(z) = \Lambda_0 - \Lambda_1 = T_0^{-1} v_0 - T_1^{-1} v_1 , \tag{4.2}$$

where $z = (\zeta_1,\ldots,\zeta_k)$ and F are k-vectors.

To keep the following computational details simple, we restrict attention to real ζ_i. We wish to solve (4.2) by Newton's method; to get the necessary derivatives let $(\dot{})$ represent $\frac{d}{d\zeta_j}$ and recall that $(\dot{M^{-1}}) = - M^{-1}\dot{M}M^{-1}$ for invertible matrices M. Thus

$$\begin{aligned} \dot{F}(z) &= (\dot{T_0^{-1}})v_0 + T_0^{-1}\dot{v}_0 - (\dot{T_1^{-1}})v_1 - T_1^{-1}\dot{v}_1 \\ &= T_0^{-1}(\dot{v}_0 - \dot{T}_0\Lambda_0) - T_1^{-1}(\dot{v}_1 - \dot{T}_1\Lambda_1) . \end{aligned} \tag{4.3}$$

Now

$$\dot{T}_0 = \dot{S}_0 W^{-1} S_0^* + S_0 W^{-1} \dot{S}_0^*$$

$$= \begin{pmatrix} & 0 & \\ \cdots & e_j^* D W^{-1} e_i & \cdots \\ & 0 & \end{pmatrix} + \begin{pmatrix} & \vdots & \\ 0 & e_i^* W^{-1} D^* e_j & 0 \\ & \vdots & \end{pmatrix} .$$

The non-zero entries are contained in the j'th row and the j'th column respectively. Continuing,

$$\dot{T}_0 \Lambda_0 = e_j^* D W^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sum \Lambda_{0,i} e_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \text{j'th entry non-zero} + \Lambda_{0,j} S_0 W^{-1} D^* e_j ,$$

(4.4)

$$\dot{T}_1 \Lambda_1 = (e_j^* + \zeta_j e_j^* D) W^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sum \Lambda_{1,i} e_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \Lambda_{1,j} Z S_1 W^{-1} D^* e_j .$$

By use of formulas (4.4) in (4.3) we may compute the j'th row of the Jacobian matrix appropriate for use with Newton's method to find solutions of (4.2).

In terms of our familiar diagonal norms,

$$(T_0)_{ij} = \sum_{r=1}^{n} (\zeta_i \zeta_j^*)^{n-r} / w_r ,$$

$$(T_1)_{ij} = \sum (n-r)(\zeta_i \zeta_j^*)^{n-r} / w_r ,$$

$$(\dot{T}_0\Lambda_0)_i = \Lambda_{0,j}\zeta_i \sum (n-r)(\zeta_i\zeta_j^*)^{n-r-1}/w_r \ , \quad i \neq j \ ,$$

$$(\dot{T}_0\Lambda_0)_j = \sum_i \Lambda_{0,i}\zeta_i^* \sum (n-r)(\zeta_j\zeta_i^*)^{n-r-1}/w_r$$
$$+ \Lambda_{0,j}\zeta_i \sum (n-r)(\zeta_i\zeta_j^*)^{n-r-1}/w_r \ ,$$

$$(\dot{T}_1\Lambda_1)_i = \Lambda_{1,j}\zeta_i \sum (n-r)^2(\zeta_i\zeta_j^*)^{n-r-1}/w_r \ , \quad i \neq j \ ,$$

$$(\dot{T}_1\Lambda_1)_j = \sum_i \Lambda_{1,j}\zeta_i^* \sum (n-r)^2(\zeta_j\zeta_i^*)^{n-r-1}/w_r$$
$$+ \Lambda_{1,j}\zeta_i \sum (n-r)^2(\zeta_i\zeta_j^*)^{n-r-1}/w_r \ .$$

5. Deflation for Several Double Zeros

When solving equation (4.2) for polynomials of degrees only modestly larger than $2k$, one often finds that zeros of F are quite abundant. In order to prevent reconvergence to zeros already found, some sort of deflation is required.

Unless further steps are taken, moreover, convergence will occur to solutions in which some of the ostensibly distinct ζ_i have coalesced. This behavior must also be suppressed; we shall do so numerically.

A workable approach is to find the zeros of G, rather than F, where

$$G(z) \equiv F(z)/\Delta$$

$$\Delta \equiv \Delta_1 \Delta_2 ,$$

$$\Delta_1 \equiv \prod_{i>r} (\zeta_i - \zeta_r)^2$$

for elements ζ_i and ζ_r of z, and

$$\Delta_2 \equiv \prod_{s} \|z - z^s\|_2^2$$

for known zeros z^s of F.

If we let $(\dot{}) = \frac{d}{d\zeta_j}$ then

$$\dot{G} = \dot{F}/\Delta - (\dot{\Delta}/\Delta)G .$$

We know that

$$(\dot{\Delta}/\Delta) = (\dot{\Delta}_1/\Delta_1) + (\dot{\Delta}_2/\Delta_2)$$

and we find that

$$\dot{\Delta}_1 = 2\Delta_1 \sum_{i \neq j} 1/(\zeta_j - \zeta_i) , \qquad \dot{\Delta}_2 = 2\Delta_2 \sum_{s} \frac{(\zeta_j - z_j^2)}{\|z - z^s\|_2^2} .$$

6. The Equations for Two Real Double Zeros

For the case when only two real double zeros ζ_1 and ζ_2 are sought, divided differences may be exploited to reduce the probability of coalescence of ζ_1 and ζ_2 to the same double zero.

Recall our equation:

$$F(z) = \Lambda_0 - \Lambda_1 = T_0^{-1}v_0 - T_1^{-1}v_1 \,. \tag{4.2}$$

Here

$$T_0 = \begin{pmatrix} e_1^*W^{-1}e_1 & e_1^*W^{-1}e_2 \\ e_2^*W^{-1}e_1 & e_2^*W^{-1}e_2 \end{pmatrix},$$

$$T_1 = \begin{pmatrix} \zeta_1 e_1^*DW^{-1}e_1 & \zeta_1 e_1^*DW^{-1}e_2 \\ \zeta_2 e_2^*DW^{-1}e_1 & \zeta_2 e_2^*DW^{-1}e_2 \end{pmatrix},$$

$$v_0 = \begin{pmatrix} p(\zeta_1) \\ p(\zeta_2) \end{pmatrix}, \qquad v_1 = \begin{pmatrix} \zeta_1 p'(\zeta_1) \\ \zeta_2 p'(\zeta_2) \end{pmatrix}.$$

Then

$$T_0^{-1} = \frac{1}{\det(T_0)} \begin{pmatrix} e_2^*W^{-1}e_2 & -e_1^*W^{-1}e_2 \\ -e_2^*W^{-1}e_1 & e_1^*W^{-1}e_1 \end{pmatrix} \equiv \frac{1}{\delta_0} T_0^{\ddagger}$$

$$T_1^{-1} = \frac{1}{\det(T_1)} \begin{pmatrix} \zeta_2 e_2^*DW^{-1}e_2 & -\zeta_1 e_1^*DW^{-1}e_2 \\ -\zeta_2 e_2^*DW^{-1}e_1 & \zeta_1 e_1^*DW^{-1}e_1 \end{pmatrix} = \frac{1}{\delta_1} T_1^{\ddagger} \,.$$

The symbol ‡ denotes the adjoint matrix: $A^{\ddagger} = \det(A)A^{-1}$.

Revert to the usual diagonal norm to obtain the following expression for δ_0:

$$\delta_0 = \sum_{j=1}^{n} \frac{1}{w_j} \Big[\sum_{k=1}^{n} \frac{1}{w_k} \{ |\zeta_1^2|^{n-j} |\zeta_2^2|^{n-k} - (\zeta_1\zeta_2^*)^{n-j} (\zeta_2\zeta_1^*)^{n-k} \} \Big]$$

$$(6.1)\qquad \hat{\delta}_0 \equiv \frac{\delta_0}{|\zeta_1-\zeta_2|^2} = \sum_{j=1}^{n-1} \frac{1}{w_j} \sum_{k=j+1}^{n} \frac{1}{w_k} (|\zeta_1\zeta_2|^2)^{n-k} |\Delta_{k-j}|^2 .$$

In the last expression, the divided difference $\Delta_i \equiv (\zeta_1^i-\zeta_2^i)/(\zeta_1-\zeta_2)$ is a polynomial in ζ_1 and ζ_2 for any $i \geq 0$. The corresponding result for δ_1 is

$$\hat{\delta}_1 = \sum_{j=1}^{n-1} \left(\frac{n-j}{w_j}\right) \sum_{k=j+1}^{n} \left(\frac{n-k}{w_k}\right) (|\zeta_1\zeta_2|^2)^{n-k} |\Delta_{k-j}|^2 .$$

To apply Newton's method the derivatives will be required; assume ζ_1 and ζ_2 are real:

$$(6.2)\qquad \frac{\partial \hat{\delta}_0}{\partial \zeta_1} = \frac{2}{w_n} \sum_{j=1}^{n-1} \frac{\Delta_{n-j}}{w_j} \frac{\partial \Delta_{n-j}}{\partial \zeta_1}$$
$$+ 2[\zeta_1\zeta_2^2 \sum_{j=1}^{n-2} \frac{1}{w_j} \sum_{k=j+1}^{n-1} \frac{1}{w_k} \Delta_{k-j} (\zeta_1^2\zeta_2^2)^{n-k-1} \{(n-k)\Delta_{k-j} + \zeta_1 \frac{\partial \Delta_{k-j}}{\partial \zeta_1}\}]$$

Since ζ_1 and ζ_2 are symmetric in (6.1), $\frac{\partial \delta_0}{\partial \zeta_2}$ may be obtained by interchanging the roles of ζ_1 and ζ_2 in (6.2). Similarly $\frac{\partial \delta_1}{\partial \zeta_1}$ may be obtained by substituting $(n-j)/w_j$ for $1/w_j$ and $(n-k)/w_k$ for $1/w_k$.

When finding zeros we will need to compute θ_0, the first element of the vector $T_0^{\dagger} v_0$, and θ_1, the first element of the vector $T_1^{\dagger} v_1$. Then

$$(6.3)\qquad \hat{\theta}_0 \equiv \theta_0/(\zeta_1-\zeta_2) = \sum_{j=1}^{n} (\zeta_2^*)^{n-j} \Delta_{p,n-j}/w_j .$$

Now

$$\Delta_{p,n-j} \equiv \frac{\zeta_2^{n-j}p(\zeta_1) - \zeta_1^{n-j}p(\zeta_2)}{\zeta_1 - \zeta_2} .$$

$\Delta_{p,n-j}$ is a polynomial in ζ_1 and ζ_2; the details of its construction are given in Appendix 5. Similarly

$$(6.4) \qquad \hat{\theta}_1 \equiv \theta_1/(\zeta_1-\zeta_2) = \zeta_1|\zeta_2|^2 \sum_{j=1}^{n-1}(n-j)(\zeta_2^*)^{n-j-1}\Delta_{p',n-j-1}/w_j$$

where

$$\Delta_{p',n-j} \equiv \frac{\zeta_2^{n-j}p'(\zeta_1) - \zeta_1^{n-j}p'(\zeta_2)}{\zeta_1 - \zeta_2} .$$

The derivatives of the $\hat{\theta}$'s will also be needed. In the real case they are

$$(6.5) \quad \begin{aligned}
\frac{\partial\hat{\theta}_0}{\partial\zeta_1} &= \frac{1}{w_n}\frac{\partial\Delta_{p,0}}{\partial\zeta_1} + \zeta_2\sum_{j=1}^{n-1}\zeta_2^{n-1-j}\frac{\partial\Delta_{p,n-j}}{\partial\zeta_1}/w_j , \\
\frac{\partial\hat{\theta}_0}{\partial\zeta_2} &= \frac{1}{w_n}\frac{\partial\Delta_{p,0}}{\partial\zeta_2} + \sum\zeta_2^{n-j-1}\{\zeta_2\frac{\partial\Delta_{p,n-j}}{\partial\zeta_2} + (n-j)\Delta_{p,n-j}\}/w_j , \\
\frac{\partial\hat{\theta}_1}{\partial\zeta_1} &= \zeta_2^2\sum(n-j)\zeta_2^{n-j-1}\{\zeta_1\frac{\partial\Delta_{p',n-j-1}}{\partial\zeta_1} + \Delta_{p',n-j-1}\}/w_j , \\
\frac{\partial\hat{\theta}_1}{\partial\zeta_2} &= \zeta_1\zeta_2\sum(n-j)\zeta_2^{n-j-1}\{(n-j+1)\Delta_{p',n-j-1} + \zeta_2\frac{\partial\Delta_{p',n-j-1}}{\partial\zeta_2}\}/w_j .
\end{aligned}$$

We could find zeros of the function

$$\begin{aligned}
F(z) &= T_0^{-1}v_0 - T_1^{-1}v_1 \\
&= \frac{1}{\delta_0}T_0^{\ddagger}v_0 - \frac{1}{\delta_1}T_1^{\ddagger}v_1
\end{aligned}$$

but for simplicity we will instead find the zeros of

$$\hat{F}(z) = \frac{\delta_0\delta_1}{|\zeta_1-\zeta_2|^2(\zeta_1-\zeta_2)}F(z) = \frac{\hat{\delta}_1}{(\zeta_1-\zeta_2)}T_0^{\ddagger}v_0 - \frac{\hat{\delta}_0}{(\zeta_1-\zeta_2)}T_1^{\ddagger}v_1 \ .$$

$\hat{F}(z) = 0$ is a system of two equations. The first one is

$$\hat{\delta}_1\hat{\theta}_0 - \hat{\delta}_0\hat{\theta}_1 = 0 \ . \tag{6.6}$$

The second equation may be obtained from (6.6) by reversing all occurrences of ζ_1 and ζ_2 in the expressions for the $\hat{\delta}$'s and $\hat{\theta}$'s. The appropriate derivatives may be computed similarly.

Now that a specific equation, (6.6), is ready to be solved, methods for computing the various divided differences that appear in it will be required; these methods are in Appendix 5. We turn now to the question: what happens when $\zeta_1 \to \zeta_2$?

The original function (4.2) is undefined when $\zeta_1 = \zeta_2$. The modified equation

$$\delta_1 T_0^{\ddagger}v_0 - \delta_0 T_1^{\ddagger}v_1 = 0$$

turns out to be satisfied whenever $\zeta_1 = \zeta_2$. But the divided difference version, (6.6), is not so easily satisfied; let us examine what happens to its terms as $\zeta_1 \to \zeta_2$.

We discover that

$$\lim_{\zeta_1,\zeta_2\to\zeta} \Delta_k = \frac{d}{d\zeta}(\zeta^k) = k\zeta^{k-1} \ ,$$

$$\lim_{\zeta_1,\zeta_2\to\zeta} \Delta_{p,k} = \zeta^k p'(\zeta) - k\zeta^{k-1}p(\zeta) \ ,$$

$$\lim_{\zeta_1,\zeta_2\to\zeta} \Delta_{p',k} = \zeta^k p''(\zeta) - k\zeta^{k-1}p'(\zeta) \ .$$

Substituting these expressions in (6.6) and simplifying leads eventually to the equation to be solved for the nearest triple zero (III.7.1). Recall that the case of a complex conjugate pair also reduced to a triple zero when the divided differences became confluent. Just as in that case, numerical methods will be required to inhibit convergence to the triple zero solutions we wish to avoid.

Both the method of this section and the method for $k > 1$ double zeros may be used when two double zeros are required. Both methods seem to work satisfactorily for polynomials of low degree, but the general method for k double zeros worked better for Wilkinson's polynomial of degree 20 discussed in Chapter X. The equations described in this section seem to have a much greater propensity for causing Newton's method to dawdle aimlessly without converging. It may be that the divided differences warp the geometry of the function whose zeros are sought in a way that tends to conceal the zeros. There is some compensation in the fact that those divided differences help prevent coalescence of the zeros much more effectively than numerical means alone.

CHAPTER VI

LOCATION THEORY FOR NEAREST POLYNOMIALS WITH A DOUBLE ZERO

1. Introduction

In this chapter may be found some clues to the answer to the question: Given a polynomial p, all of whose zeros are simple, where should we look to find the nearest polynomial $p+q$ with a double zero ζ? That ζ which minimizes $\|q\|$ globally is one of the solutions of the equation

$$F(\zeta) \equiv \sigma_1 p(\zeta) - \sigma_0 \zeta p'(\zeta) = 0 \; ; \tag{1.1}$$

but there are usually many other solutions, most of which represent local minima.

Remember that the real non-analytic functions σ_0 and σ_1 are defined as

$$\sigma_0 \equiv \sum_{j=1}^{n} |\zeta^2|^{n-j}/w_j \quad ,$$

$$\sigma_1 \equiv \sum_{j=1}^{n-1} |\zeta^2|^{n-j}(n-j)/w_j \; .$$

Thus we are considering only the norms derived from diagonal Hermitian quadratic forms. Most of the results to follow, moreover, only apply to real polynomials p.

The purpose of attempting to develop a theory of location is to make our numerical solution procedures more efficient. Equation (1.1) is typically solved by Newton's method from some starting point. An ideal starting point would have the property that Newton's method would always converge to the global minimum corresponding to the

nearest polynomial with a double zero. A satisfactory starting point would always converge to a local minimum that is nearly globally minimal. The ad hoc starting procedures discussed in Chapter VIII usually seem to be satisfactory but the known theory is insufficient to account for their success.

The results in the following sections seem far from optimal. One might hope that a theory could be developed comparable to the elegant theory of the location of zeros of polynomials discussed by Marden [21] and Householder [12]. But much of the theory for polynomials hinges on the entire analytic nature of polynomial functions. Certain of the examples to follow effectively counter some of the conjectures that might be made by analogy with the polynomial case.

We can make a few preliminary observations about (1.1). Among its solutions are the global minimum we seek, numerous other local minima, a few non-minimal stationary points, and the solution $\zeta = 0$. This solution $\zeta = 0$ is an artifact of the way we wrote the equation. We could just as well divide by ζ and write

$$\zeta^*\Big(\sum_{j=1}^{n-1}((n-j)/w_j)|\zeta^2|^{n-j-1}\Big)p(\zeta) - \sigma_1 p'(\zeta) = 0 \ . \tag{1.2}$$

Then $\zeta = 0$ is a solution of this equation only if $p'(0) = 0$; that is, only if the next to last coefficient $p_{n-1} = 0$. An examination of the stationary condition $q^*W = \ell^*A$ tells us that $q_{n-1} = 0$ while the constraint $Ap + Aq = 0$ tells us that $q_{n-1} = -p_{n-1}$. Therefore $\zeta = 0$ is a stationary point for $\|q\|$ if and only if $p_{n-1} = 0$. Even then $\zeta = 0$ need not represent a minimum.

Since the factor ζ does not seem to contribute any information, why not leave it out in our subsequent analyses? We keep it for a reason which becomes apparent when we write (1.1) in yet a third form:

$$\frac{\zeta p'(\zeta)}{p(\zeta)} = \frac{\sigma_1(\zeta)}{\sigma_0(\zeta)} \equiv R(\zeta) \ . \tag{1.3}$$

Now

$$R(\zeta) = (\sum(|\zeta^2|^{n-j}/w_j)\cdot(n-j))/(\sum(|\zeta^2|^{n-j}/w_j))$$

may be thought of as a weighted average of the quantities $(n-j)$. If we do so then we realize that

$$0 \leq R(\zeta) < n-1$$

for

$$0 \leq |\zeta| < \infty \ .$$

Thus (1.3) equates a meromorphic function of the complex variable ζ to a bounded positive real function of $|\zeta|$, which is in fact analytic when regarded as a real function of a real variable. If the factor of ζ were removed from (1.3) it would lose its attractive form. We will exploit that form later.

A typical result in this theory is the following.

Proposition. Let p be real with two real zeros α_1 and α_2, $\alpha_1 \leq \alpha_2$. Then

$$F(\zeta) = \sigma_1 p(\zeta) - \sigma_0 \zeta p'(\zeta) = 0$$

has a solution ζ such that $\alpha_1 \leq \zeta \leq \alpha_2$.

Proof. If α_1 and α_2 have opposite signs or if either is zero, then $\zeta = 0$ satisfies the assertion. Then without loss of generality assume that $0 < \alpha_1 < \alpha_2$ and that (α_1,α_2) contains no real zero of p. Then

$$F(\alpha_1)F(\alpha_2) = \sigma_0(\alpha_1)\sigma_0(\alpha_2)\alpha_1\alpha_2 p'(\alpha_1)p'(\alpha_2) .$$

If that product is zero or negative then a zero of F lies in $[\alpha_1,\alpha_2]$ by the intermediate value theorem. But if that product is positive then $p'(\alpha_1)p'(\alpha_2) > 0$. Considering Taylor series, we see that

$$p(\alpha_1+\delta) \doteqdot \delta p'(\alpha_1) ,$$
$$p(\alpha_2-\delta) \doteqdot -\delta p'(\alpha_2) ,$$

for small enough $\delta > 0$. Thus

$$p(\alpha_1+\delta)p(\alpha_2-\delta) \doteqdot -\delta^2 p'(\alpha_1)p'(\alpha_2) < 0$$

so the p must have another zero in $[\alpha_1,\alpha_2]$, contrary to assumption. The contradiction implies $F(\alpha_1)F(\alpha_2) \leq 0$ and concludes the proof.

2. No Complex Solutions for Certain Real Polynomials

Wilkinson's polynomial of chapter X has the property that all its zeros are real and have the same sign. When solving (1.1) for Wilkinson's polynomial we need not search for complex zeros because of the following.

Proposition. Let $p(\tau) = \Pi(\tau^m-\alpha_j)$ be a complex polynomial in τ^m. If all the numbers α_j are either zero or have the same argument θ then the non-zero solutions ζ of (1.1) may only have arguments $(\theta+k\pi)/m$, $0 \le k \le 2m-1$.

Corollary. If a real polynomial $p(\tau) = \Pi(\tau-\alpha_j)$ has all real zeros α_j all of the same sign, then all its ζ's are real.

Corollary. If an even real polynomial $p(\tau) = \Pi(\tau^2-\alpha_j^2)$ has all zeros $\pm\alpha_j$ real, then all its ζ's are either real or pure imaginary.

Proof of Proposition. Rewrite (1.1) in the form of (1.3):

$$\zeta p'(\zeta)/p(\zeta) = R(|\zeta|) \ .$$

Remember R is a real function of $|\zeta|$ and $0 \le R < n-1$. Suppose first the special case that all $\alpha_j = 0$ so $p(\tau) = \tau^n$. Then (1.1) reduces to $n = R(\zeta)$, so the only solution of (1.1) is the universal solution $\zeta = 0$.

Otherwise we may assume that at least one $\alpha_j \neq 0$. Recall that

$$\begin{aligned} p'(\zeta)/p(\zeta) &= m\zeta^{m-1} \sum 1/(\zeta^m-\alpha_j) \\ &= m\zeta^{m-1} \sum (\bar{\zeta}^m-\bar{\alpha}_j)/|\zeta^m-\alpha_j|^2 \ ; \end{aligned}$$

take imaginary parts of (1.3) to find

$$0 = \mathrm{Im}(\zeta p'/p) ,$$
$$0 = \mathrm{Im}(\zeta^m \sum \bar{\alpha}_j/|\zeta^m-\alpha_j|^2) ,$$
$$0 = \mathrm{Im}(\zeta^m e^{-i\theta})\cdot\sum|\alpha_j|/|\zeta^m-\alpha_j|^2 .$$

Since at least one α_j is non-zero the sum $\sum$ of positive quantities may not vanish. Then if $\emptyset$ denotes the argument of a non-zero ζ we have

$$\mathrm{Im}(\exp(i(m\emptyset-\theta))) = 0$$

from which the result follows. Q.E.D.

Note the two resulting equations for $|\zeta|$ are

$$R = m|\zeta|^m \sum 1/(|\zeta|^m \pm |\alpha_j|)$$

which could be expressed as two real polynomials of degree $3n-2$ in $|\zeta|$. However, for polynomials in τ^m it might be reasonable to restrict perturbations to polynomials in τ^m by causing appropriate weights in the norm to become infinite. Then $R(|\zeta|)$ becomes $R(|\zeta|^m)$ and the resulting polynomials are of degree $(3n-2)/m$ in $|\zeta|^m$.

3. Counterexample

The previous proposition might lead one to hope that polynomials with all zeros real would not have complex solutions to (1.1). The following counterexample, produced by W. Kahan, eliminates such hopes:

Example. Let $n = 2$ and $p(\tau) = (\tau-1)(\tau+1)$. If $2w_1 < w_2$, then (1.1) has a complex solution

$$\zeta = \pm i\sqrt{1 - (2w_1/w_2)} .$$

Comments. Some other surprising facts may be learned from this one example. We start by deriving all the solutions of (1.1). Let $\omega = (w_1/w_2) > 0$. Then (1.1) is

$$|\zeta|^2(\zeta^2-1) - (|\zeta|^2+\omega)\zeta(2\zeta) = 0$$

or, dividing by the solution $\zeta = 0$,

$$\zeta|\zeta|^2 + 2\zeta\omega + \zeta^* = 0 ;$$

then

$$(\mathrm{Re}\ \zeta)(|\zeta|^2 + 2\omega + 1) = 0$$

and

$$(\mathrm{Im}\ \zeta)(|\zeta|^2 + 2\omega - 1) = 0 .$$

By considering the various possibilities we conclude that the only solutions of these equations are just $\zeta = 0$ and, if $\omega < \frac{1}{2}$, $\zeta = \pm i(1-2\omega)^{1/2}$. The norm of the corresponding q's may be calculated to be

$$\|q\|^2 = w_2 , \qquad \text{for a double zero at } 0,$$
$$\|q\|^2 = 4\omega(1-\omega)w_2 , \qquad \text{for a double zero at } \pm i(1-2\omega)^{1/2} .$$

So for $0 < \omega < \frac{1}{2}$, the global minima are at $= \pm i(1-2\omega)^{1/2}$, not at $\zeta = 0$. In this case, 0 represents a saddle point; it is where the global minimum occurs if only real ζ are considered. But on the imaginary axis, the minima occur elsewhere, and a local maximum occurs at $\zeta = 0$ if only pure imaginary ζ are considered.

Of course, there are other real polynomials with all zeros real which have solutions of (1.1) which are complex but not pure imaginary. It is perhaps surprising that an even real polynomial with some zeros real and some pure imaginary may have solutions ζ of (1.1) which are neither real nor pure imaginary. For instance, by appropriate choice of weights so that the $R(|\zeta|)$ of (1.3) has the value 2 when $|\zeta| = 1$, we find that some solutions ζ for the polynomial

$$p(\tau) = \tau^4 - 1$$

are $\zeta = 0$ and $\zeta = (\pm 1 \pm i)/\sqrt{2}$. We may further restrict the weights so that these are the only ζ's.

Returning to Kahan's counterexample, recall the Lucas theorem: the convex hull of the zeros of a polynomial contains all the zeros of its derivative. The present example shows that no such simple statement may be made about the geometrical relationship between the zeros of a polynomial and the solutions of (1.1). Some early experimental results suggested that the convex hull of the origin and the zeros of the polynomial always contained the global minimum. But the counterexample shows that this is not always the case.

Yet the solutions of (1.1) do behave somewhat like the zeros of the derivative of the corresponding polynomial. Consider these symmetry

Facts:

1) If p is real then F of (1.1) is real;

2) if p is odd then so is F;

3) if p is even then so is F;

4) if all the zeros of p are multiplied by a constant phase factor $\exp(i\theta)$ then so are the zeros of F. Thus there is no essential difference between a real polynomial and a complex one whose zeros are symmetric about a line through the origin.

In contrast, consider this invariance of polynomials under scaling: if the zeros of p are all multiplied by a scale factor, then all the zeros of all the derivatives are scaled by the same factor. But if the weights in the σ's of (1.1) are regarded as fixed, then scaling the zeros of p does not introduce a corresponding scaling of the solutions of (1.1), which change in a complicated way. One could regard the weights as depending on the scaling factor, however. If, for instance,

$$w_j = c_j \cdot (\mu^2)^{n-j}$$

where c_j is fixed and μ is the modulus of the zero of p of largest modulus, then a scaling change in the zeros of p will produce a corresponding scaling of the solutions of (1.1). One could go further and imagine that $\mu = |\zeta|$, a function of the ostensibly unknown ζ. Then the σ's are constant and the F of (1.1) takes an especially simple form: it becomes a polynomial. In some of the sections to follow this analytical "swindle" will be exploited.

4. A Bound on the Solutions ζ

We will exploit Theorem (17,2a) of Marden [21] to bound the solutions of (1.1). It is not immediately obvious how large those solutions might be, relative to the zeros of the polynomials.

Marden's theorem concerns the location of the zeros of a linear combination of monic polynomials of degree n. Let $x(\tau) - \lambda y(\tau)$ be that linear combination, and let $C(c,r)$ represent a circle of radius r centered at c. $C_x(c_x,r_x)$ contains all the zeros of x and $C_y(c_y,r_y)$ contains all the zeros of y. The theorem asserts that all the zeros of $x - \lambda y$ lie in the union of the n circles $C_k(\gamma_k,\rho_k)$, $1 \leq k \leq n$, where

$$\gamma_k = (c_y - \omega_k c_x)/(1-\omega_k)$$

and

$$\rho_k = (r_y + |\omega_k| r_x)/|1-\omega_k|$$

and

$$\omega_k = \lambda^{1/n}\varepsilon_k \ .$$

The ε_k are the n n^{th} roots of 1.

Our result is the following.

Corollary. If $|\alpha_{max}|$ is the maximum modulus of the zeros of p, then all the solutions ζ of (1.1) satisfy

$$|\zeta| \leq 2n^2|\alpha_{max}| \ . \tag{4.1}$$

Proof. Rewrite (1.1) in a form appropriate to the theorem:

$$G_R(\zeta) = \left(\frac{\zeta p'(\zeta)}{n}\right) - \left(\frac{R}{n}\right)p(\zeta) = 0 \ .$$

Then if R is held fixed, G_R is in the proper form. Let the circles

C_x and C_y be crudely approximated by $C(0,|\alpha_{max}|)$. This circle certainly contains all the zeros of p, and hence of p', as well as 0. Then $\gamma_k = 0$ so the circles C_k of the theorem are concentric and only the radius of the largest matters:

$$\rho_k = \frac{1 + \sqrt[n]{R/n}}{|1 - \sqrt[n]{R/n}\ \varepsilon_k|} |\alpha_{max}| .$$

Remembering that $0 \leq R < n-1$, it is clear that

$$\begin{aligned} \rho_1 &= \frac{1 + \sqrt[n]{R/n}}{1 - \sqrt[n]{R/n}} |\alpha_{max}| \\ &\leq \frac{2}{1 - \sqrt[n]{(n-1)/n}} |\alpha_{max}| \\ &\leq 2n^2 |\alpha_{max}| . \end{aligned}$$

Since any solution of (1.1) is a zero of G_R for some positive $R < n-1$, the bound is valid for all such solutions. Q.E.D.

The purpose of this crude estimate is just to show that the solutions of (1.1) are bounded. The gross approximations involved might lead one to doubt that the bound is realistic, and indeed for "normal" polynomials the solutions do not seem to exceed $|\alpha_{max}|$.

However Wilkinson's polynomial of degree 20, discussed in chapter X, has a solution for (1.1) at $\zeta \doteq -117.31$; the norm has $w_j = 1/|p_j|^2$ which minimizes relative changes in the coefficients. In this case $|\zeta_{max}|$ exceeds $|\alpha_{max}|$ by a factor of nearly 5. Presumably by appropriate choice of norm that factor could be made even larger -- how much larger is unknown.

One might consider a type of iteration scheme: since the bound (4.1) depends heavily on the maximum value of R, which we bounded by $n-1$, any knowledge that reduces that R_{max} should affect the bound appreciably. But R is monotonic in $|\zeta|$ so R_{max} depends on the bound on $|\zeta|$, which is in turn dependent on R_{max}. Clearly we could reduce the bounds on $|\zeta|$ and R_{max} alternatingly. Unfortunately in practice such an iteration seems to improve the bound so little as to be scarcely worth the trouble.

5. Propositions for Real Quadratic Polynomials

The example of section 3 was a counter to a tempting, but incorrect assertion. That same example could be regarded positively, however, as an example of the propositions of the present section.

Proposition 5.1. Consider a real monic quadratic polynomial

$$p(\tau) = \tau^2 - 2\alpha\tau + \gamma \ .$$

Let μ be the modulus of its largest zero. Then every solution ζ of (1.1) satisfies $|\zeta| \leq \mu$.

Proof. By examination of cases. Equation (1.1) may be written

$$(\zeta^2-2\alpha\zeta+\gamma)(|\zeta|^2/w_1) - \zeta(2\zeta-2\alpha)(|\zeta|^2/w_1 + 1/w_2) = 0 \ .$$

Factor out ζ to remove the uninteresting solution $\zeta = 0$; then letting $\omega = (w_1/w_2) > 0$, and taking real and imaginary parts leaves the equations

$$|\zeta|^2 \operatorname{Re}\zeta + (2\omega-\gamma)\operatorname{Re}\zeta - 2\alpha\omega = 0 \ , \tag{5.1}$$

$$|\zeta|^2 \operatorname{Im}\zeta + (2\omega+\gamma)\operatorname{Im}\zeta = 0 \ . \tag{5.2}$$

The second of these equations is satisfied if $|\zeta|^2 = -(2\omega+\gamma)$ or $\operatorname{Im}\zeta = 0$, providing two cases.

In the first of these cases $\gamma < 0$ so the zeros of p are real and

$$\mu = |\alpha| + (\alpha^2-\gamma)^{1/2} \ .$$

But we may easily verify that

$$|\zeta|^2 = -(2\omega+\gamma) \leq \mu^2$$

as claimed.

Im $\zeta = 0$ in the second case so the solutions ζ are just the real solutions of (5.1), which satisfy

(5.3) $$g(\zeta) \equiv \zeta^3 + (2\omega-\gamma)\zeta - 2\alpha\omega = 0 \ .$$

g may have complex solutions but these do not satisfy (1.1).

We will prove the proposition by showing that $g(-\mu) < 0$, $g(+\mu) > 0$, and the real critical points where $g'(\zeta)$ vanishes are contained in $[-\mu,+\mu]$. Thus the real zeros of g are bracketed in $[-\mu,+\mu]$ whether they be 1, 2, or 3 in number. The details, however, depend on whether the zeros of p are real or complex.

Suppose first that $\alpha^2 < \gamma$ so the zeros of p are complex and $\mu = \gamma^{1/2}$. Then

$$g(-\mu) = -2\omega(\gamma^{1/2}+\alpha) < 0$$

and

$$g(+\mu) = +2\omega(\gamma^{1/2}-\alpha) > 0 \ .$$

Furthermore the zeros of g' are $\pm((\gamma-2\omega)/3)^{1/2}$. When these zeros are real they are less than $\gamma^{1/2}$ in modulus since $\omega > 0$.

Now suppose that $\alpha^2 \geq \gamma$ so the zeros of p are real and

$$\mu = |\alpha| + (\alpha^2-\gamma)^{1/2} \ .$$

Then

$$g(-\mu) = -\mu(\mu^2+2\omega-\gamma) - 2\alpha\omega$$
$$g(+\mu) = +\mu(\mu^2+2\omega-\gamma) - 2\alpha\omega \ .$$

It is easy to verify that

$$\mu^2 + 2\omega - \gamma > 0$$

and

$$|\mu(\mu^2+2\omega-\gamma)| \geq |2\alpha\omega|$$

so $g(-\mu) < 0$ and $g(+\mu) > 0$. And finally we may verify that when g' has real zeros $\pm((\gamma-2\omega)/3)^{1/2}$, they do not exceed μ in magnitude. Q.E.D.

Our next result is in a similar vein.

Proposition 5.2. Consider a real monic quadratic polynomial

$$p(\tau) = \tau^2 - 2\alpha\tau + \gamma \ .$$

Then there is a solution ζ of (1.1) in the smallest circle containing both zeros of p.

Proof. The zeros of p are $\alpha \pm (\alpha^2-\gamma)^{1/2}$ and the smallest circle containing them has center α and radius $|\alpha^2-\gamma|^{1/2}$. Therefore the assertion is that there is a solution ζ such that

$$|\zeta-\alpha| \leq |\alpha^2-\gamma|^{1/2} \ .$$

The solution $\zeta = 0$ satisfies the proposition if $\gamma \leq 0$ or $\gamma \geq 2\alpha^2$, so assume henceforth that $0 < \gamma < 2\alpha^2$.

Recalling equations (5.1) and (5.2), we find that the only remaining solutions are the real solutions of

$$g(\zeta) \equiv \zeta^3 + (2\omega-\gamma)\zeta - 2\alpha\omega = 0 \ .$$

In the limiting case $R \to 0$, the ζ's approach α and 0. In contrast, as $R \to 1$ the ζ's approach $\pm\gamma^{1/2}$. So, in particular, if $\gamma < 0$, corresponding to the zeros of p being real and opposite in sign, then in the second limit the zeros are pure imaginary. This situation corresponds to the counterexample of section 3.

Two results from the previous section that the limiting cases support are that 1) the magnitude of the ζ's does not exceed that of the larger zero of p, and 2) there is always one ζ in the smallest circle containing both zeros of the quadratic p. These are correct inferences.

Proposition 6.1. Let ζ be any solution of (6.1) when p is a real quadratic polynomial. Then $|\zeta|$ does not exceed the magnitude of the larger zero of p.

Proof. Consider four cases: the zeros of p are equal; the zeros of p are complex; the zeros of p are real as are the ζ; the zeros of p are real but the ζ are complex. The first case is trivial and the other three cases are similar in proof. For the last case, for instance, we have

$$(1-R)^2\alpha^2 + R(2-R)\gamma < 0 \quad \text{and} \quad \alpha^2 > \gamma \ .$$

Obviously $\gamma < 0$. We wish to compare $|\zeta|$ with μ, the modulus of the larger zero of p:

$$|\zeta| = \sqrt{\left(\frac{R}{2-R}\right)(-\gamma)} \ ,$$

$$\mu = |\alpha| + \sqrt{\alpha^2 - \gamma} \ .$$

Thus

$$\mu^2 - |\zeta|^2 = 2\alpha^2 + 2|\alpha|\sqrt{\alpha^2-\gamma} - 2\gamma(\tfrac{1-R}{2-R})$$

which is a sum of non-negative terms, since $\gamma < 0$ and $R < 1$. The last term is positive so $|\zeta| < \mu$. Q.E.D.

Proposition 6.2. The smallest circle containing both zeros of a real quadratic p contains a solution of (6.1).

Proof. As in the previous proposition there are four cases. Below we sketch the proof of the case in which both zeros of p are complex. Then $\alpha^2 < \gamma$, $\gamma > 0$, and both ζ's are real. We wish to show that $|\zeta-\alpha| \leq (\gamma-\alpha^2)^{1/2}$ for one of the ζ's.

Now

$$\zeta - \alpha = (\tfrac{-1}{2-R})\alpha \pm \Delta^{1/2}$$

where

$$\Delta \equiv (\tfrac{1-R}{2-R})^2\alpha^2 + \tfrac{R}{2-R}\gamma \ .$$

Then

$$|\zeta-\alpha|^2 = (\tfrac{1}{2-R})^2\alpha^2 + (\tfrac{1-R}{2-R})^2\alpha^2 + (\tfrac{R}{2-R})\gamma \mp (\tfrac{2}{2-R})\alpha\Delta^{1/2} \ ,$$

and we want to show that for either $+$ or $-$,

$$\mp \alpha\Delta^{1/2} \leq (1-R)(\gamma-\alpha^2) - \frac{\alpha^2}{2-R} \ .$$

If we choose the sign that makes $\mp\alpha$ negative we find that the last inequality is equivalent to $\gamma \geq \alpha^2$, which is what we assumed.

The proofs of the other cases are similar. Q.E.D.

As a tool for analysis the swindle does not seem to help much in the quadratic case. All of the propositions about quadratics are

Thus we must show that there is a solution ζ in $[\eta,\theta]$ where

$$\eta = \alpha - |\alpha^2-\gamma|^{1/2}, \qquad \theta = \alpha + |\alpha^2-\gamma|^{1/2}.$$

We do so by demonstrating that $g(\eta)\cdot g(\theta) \leq 0$.

Now

$$g(\eta)g(\theta) = \alpha^2(3|\alpha^2-\gamma| + \alpha^2 - \gamma)^2 - |\alpha^2-\gamma|(|\alpha^2-\gamma| + 3\alpha^2 - \gamma + 2\omega)^2 .$$

Suppose first that $\alpha^2 \geq \gamma$. Then

$$g(\eta)g(\theta) = -4(\alpha^2-\gamma)(\gamma^2 + \omega^2 + 2\omega(2\alpha^2-\gamma)) .$$

But the last factor is easily seen to be positive.

Suppose that $\alpha^2 < \gamma$. Then

$$g(\eta)g(\theta) = -4(\gamma-\alpha^2)(\omega^2 + 2\alpha^2\omega + \alpha^2(2\alpha^2-\gamma)) .$$

But at the outset we restricted $\gamma < 2\alpha^2$. Q.E.D.

This last proposition might lead one to suppose that for any polynomial p of degree $n \geq 2$, equation (1.1) has a solution in the smallest circle containing two zeros of p. In section 7 this supposition will be shown to be incorrect, and a weaker conjecture will be proposed.

6. Swindle Results for Real Quadratic Polynomials

A method for evading certain problems arising from the non-analyticity of (1.1) was briefly mentioned in section 3. Namely, each weight in the norm was defined to be

$$w_j = c_j |\zeta^2|^{n-j} .$$

Thus σ_0 and σ_1 are constant and therefore so is R of (1.3). This amounts to an analytical swindle since the dependence of the w_j on ζ was not incorporated into the derivation of (1.1). None the less any solution of (1.1) is also a solution of

$$s(\zeta) \equiv \zeta p'(\zeta) - Rp(\zeta) = 0 \tag{6.1}$$

for some fixed R; the R depends on $|\zeta|$ in general, but not in the swindle case. In either case $0 \leq R < n-1$.

It is useful to study the solutions of (6.1) for fixed R to see what light they shed on the original problem.

We start by noting that (6.1) has a solution $\zeta = 0$ only if $p(0) = 0$. So the part of the previous theory that depends on a solution at $\zeta = 0$ may not necessarily be true.

Write the quadratic p as

$$p(\tau) = \tau^2 - 2\alpha\tau + \gamma$$

so α is the arithmetic mean of the zeros of p and γ is their product. Then the zeros of s are

$$\zeta = \left(\frac{1-R}{2-R}\right)\alpha \pm \sqrt{\left(\frac{1-R}{2-R}\right)^2\alpha^2 + \left(\frac{R}{2-R}\right)\gamma} .$$

proved just as easily without the swindle. It is never the less helpful to verify the similarity of the theories in the quadratic case, since it is difficult to extend any results to higher degrees without the swindle.

7. The Smallest Circle Containing Two Zeros Need Not Contain a ζ

In sections 1 and 5 we learned that 1) there is a real ζ between any two real zeros of a real polynomial p, 2) a corresponding result holds for complex polynomials symmetric about a line through the origin, and 3) the smallest circle containing the two zeros of a real quadratic polynomial contains a ζ. Furthermore, when α is a complex zero of a real p with $|\text{Re } \alpha| < |\text{Im } \alpha|$, then $\zeta = 0$ is contained in the smallest circle containing α and its conjugate. In section 8 we will see that when a polynomial with a double zero is subjected to a small perturbation causing the double zero to split, the smallest circle containing the split zeros contains a ζ. From these facts we might conclude that the smallest circle containing two zeros of any polynomial p contains a ζ.

This conclusion is supported by all the experimental results reported in chapters IX and X, using norms which measure absolute or relative changes in the coefficients of p. But an investigation to settle this specific question turned up a counterexample, given below, and led to a further conjecture which is not yet resolved.

The counterexample was discovered by computationally exploiting the analytic swindle described in section 6. A crude optimization program varied the zeros of a real cubic polynomial and the fixed constant R in order to make the ζ's lie as far as possible from the center of the smallest circle containing the two complex zeros of the polynomial. A polynomial $p(\tau)$ was found with zeros α at 1.0 and $.224 \pm .174i$. When $R = 1.987$ the zeros of $s(\tau)$, as in equation (6.1), were $-.830$ and $.424 \pm .099i$; see Figure VI.1. Thus the complex ζ's are just outside the circle containing the complex zeros α.

The swindle was used because the polynomial equation $s(\tau) = 0$ may be solved equickly. Our real interest, of course, is in finding an example without using the swindle. So another crude optimization program was run with $p(\tau)$ fixed but with the norm weights allowed to vary in such a way that $\zeta = .424 \pm .174i$ remained a solution of (1.1). Surprisingly enough, the program quickly converged to a suitable counterexample: Let the weights be 1, 1000, and 10000. Then (1.1) has no solutions inside the smallest circle containing the α's $.224 \pm .174i$. The closest ζ's are at $.4245 \pm .0993i$ and 0. See Figure VI.2.

Thus we must discard the conjecture that the smallest circle containing two zeros of a polynomial contains a ζ. That should come as no surprise, however, for the corresponding conjecture about derivatives is not true either: the smallest circle containing two zeros of a polynomial need not contain a zero of the derivative. Rather the following is known:

Proposition. Let a circle of radius ρ contain m zeros of a polynomial p of degree n. Then there is a zero of the $m\text{-}1^{th}$ derivative of p in the concentric circle of radius

$$\rho \csc((\pi/2)/(n+1-m)) .$$

This proposition is stated in a stronger form and proved by Kahan [17]. The proposition suggests the following revised

Conjecture. Let a circle of radius ρ contain m zeros of a polynomial p of degree n. Then there is a solution of the appropriate equation for the nearest polynomial with an m-tuple zero within

the concentric circle of radius

$$\rho \csc((\pi/2)/(n+1-m)) .$$

Thus real cubic polynomials that have a complex conjugate pair of zeros α should have a solution ζ for a double zero such that $|\zeta - \mathrm{Re}\,\alpha| \leq \sqrt{2}\,|\mathrm{Im}\,\alpha|$. None of the examples we have encountered or constructed have violated this revised conjecture.

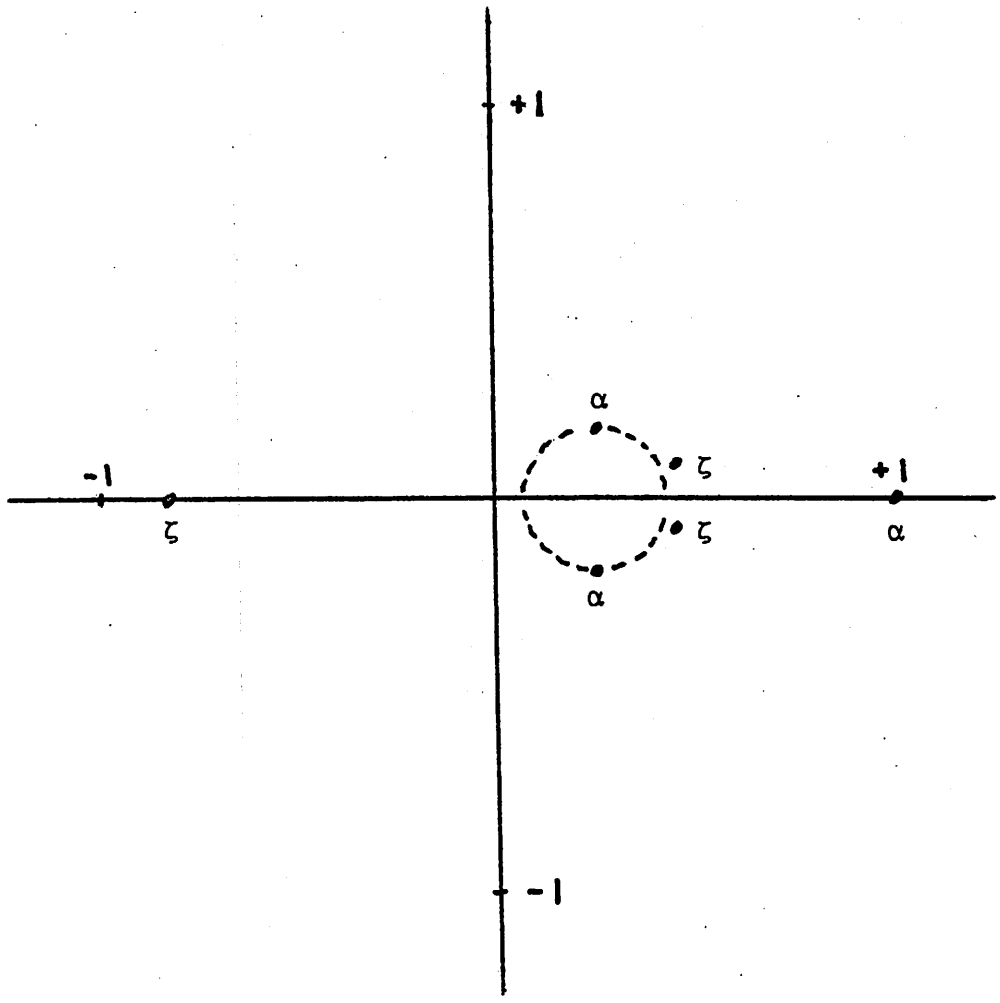

Figure VI.1. Counterexample based on swindle.
No ζ lies inside circle.
$p(\alpha) = 0$, $s(\zeta) = 0$, $R = 1.987$.

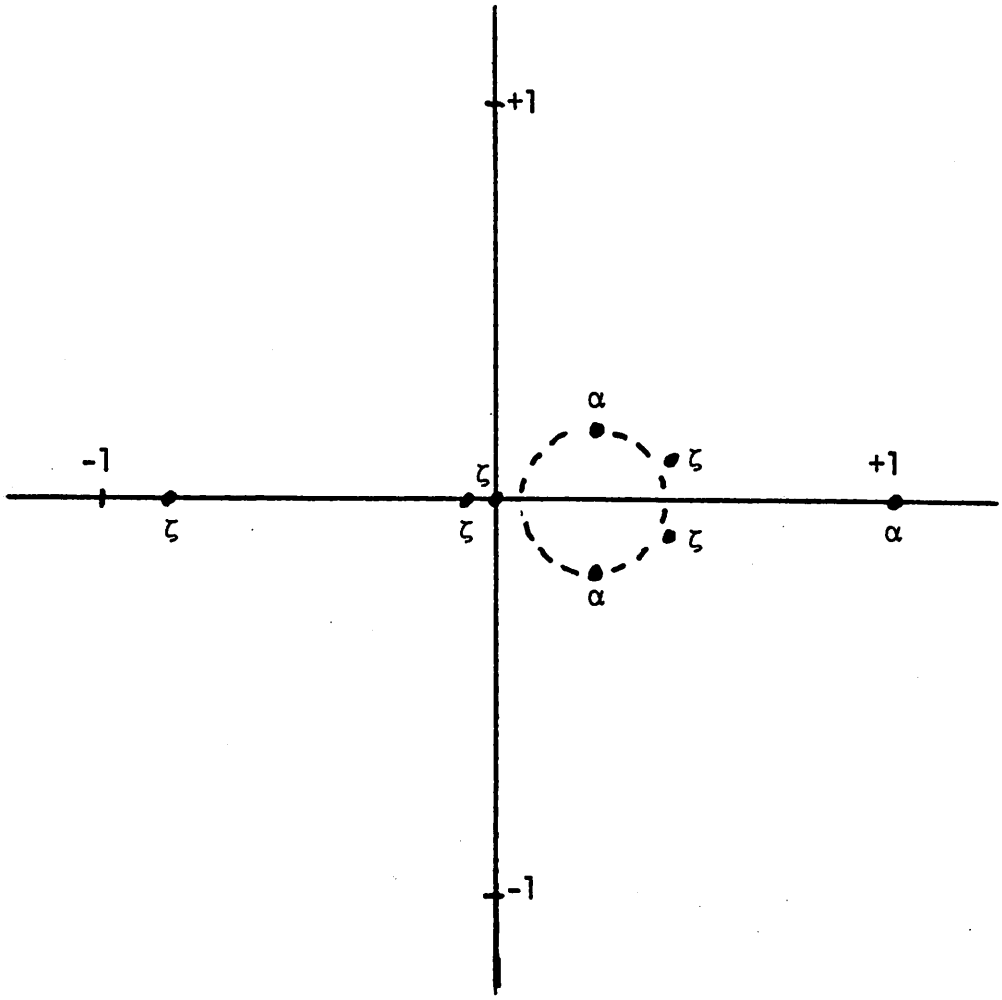

Figure VI.2. Counterexample without swindle.
No ζ lies inside circle. $p(\alpha) = 0$,
$F(\zeta) = 0$, other ζ's are farther away.
$w_1 = 1$, $w_2 = 1000$, $w_3 = 10000$.

8. Infinitesimal Location Theory

This section provides a bridge between the location theory of previous sections and the perturbation theory of the next chapter. In this section we seek to answer the question: "Where do the solutions ζ of (1.1) go when a polynomial with a double zero is perturbed infinitesimally?"

Recall that if α is a double zero of a polynomial p then it is a solution of equations (1.1) and (6.1) -- as would be expected, since a place where no perturbation is required to get a double zero is obviously a critical point for norms of such perturbations. Most perturbations of a polynomial with a multiple zero will break that multiple zero into ill conditioned simple zeros, but we shall see that the solution of (1.1) only moves in a well conditioned manner when subject to such a perturbation.

Let

$$p(\tau) = (\tau-\alpha)^2 q(\tau) , \qquad q(\alpha) \neq 0 ,$$

be our starting polynomial with a double zero and a solution of (1.1) at α. Let

$$\check{p}(\tau) = p(\tau) + \delta\varepsilon h(\tau) , \qquad h(\alpha) \neq 0 ,$$

be p subject to a perturbation which is a linear function of $\delta\varepsilon$. Also $\alpha+\delta\alpha$ will represent a zero of $\check{p}$ perturbed from α. Then expanding in Taylor series,

$$0 = \check{p}(\alpha+\delta\alpha) \doteq \tfrac{1}{2}(\delta\alpha)^2 p''(\alpha) + \delta\varepsilon(h(\alpha) + \delta\alpha h'(\alpha)) .$$

Simplifying, we find

$$\delta\alpha \doteq \pm \left((-h(\alpha)/q(\alpha))\delta\varepsilon\right)^{1/2} , \tag{8.1}$$

the classical result that a double zero tends to divide into two simple zeros according to a fractional power of the perturbation.

α is also a zero of

$$f(\zeta) = R(\zeta)p(\zeta) - \zeta p'(\zeta) .$$

Let $\alpha + \delta\zeta$ be the perturbed solution when p is perturbed to $\check{p}$. We wish to find a Taylor series expansion for $\delta\zeta$ in terms of $\delta\varepsilon$. R is not analytic in ζ, so we must use the fact that it is an analytic real function of the real variables $\operatorname{Re}\zeta$ and $\operatorname{Im}\zeta$. Eventually we find that

$$\delta\zeta = \{(R(\alpha)h(\alpha)-\alpha h'(\alpha))/(2\alpha q(\alpha))\}\delta\varepsilon + O(\delta\varepsilon^2) \tag{8.2}$$

provided

$$\alpha \neq 0$$

and

$$R(\alpha)h(\alpha) - \alpha h'(\alpha) \neq 0 .$$

The last condition represents a kind of "orthogonal" perturbation h which does not affect the solution ζ of (1.1) to first order.

Comparing (8.2) and (8.1) we see that for a typical perturbation h, the zeros of p move away from α much faster than the zero of f. Since those ill conditioned zeros of p are moving in opposite directions, the smallest circle containing them will also contain a solution of (1.1) whenever p is close enough to the manifold of polynomials with double zeros.

For further comparison, consider the change in the zero of the derivative of p. If $\alpha + \delta\theta$ denotes the zero of $\check{p}'$, we find that

$$\delta\theta = \left(-h'(\alpha)/2q(\alpha)\right)\delta\varepsilon$$

provided $h'(\alpha) \neq 0$. So the zero of the derivative also changes linearly with $\delta\varepsilon$. If $\left(R(\alpha)h(\alpha)/\alpha h'(\alpha)\right)$ is sufficiently small -- as must occur if α is sufficiently close to zero -- then $\delta\zeta$ and $\delta\theta$ are nearly the same. Unfortunately $\delta\zeta$ and $\delta\theta$ are quite different in general so $\delta\theta$ may not serve well as an estimate of $\delta\zeta$.

CHAPTER VII

PERTURBATION THEORY FOR MULTIPLE ZEROS OF POLYNOMIALS

1. Introduction

In this chapter we will recall the standard theory of perturbations of multiple zeros of polynomials, discern its limitations, and propose a more satisfactory theory which reflects the insights gained from the research described in previous chapters.

To recall the classical theory, start with a polynomial with multiple zero α:

$$p(\tau) = (\tau-\alpha)^m q(\tau) , \quad q(\alpha) \neq 0 .$$

The condition $q(\alpha) \neq 0$ means that the multiplicity of α is precisely m. We wish to see how an arbitrary perturbation of p affects α. In general α will tend to split up into m distinct zeros.

Apply a perturbing polynomial $\varepsilon r(\tau)$ of degree at most $n-1$ to get

$$\check{p}(\tau) = (\tau-\alpha)^m q(\tau) + \varepsilon r(\tau) .$$

If $(\tau-\alpha)^m$ divided $r(\tau)$ then the problem would be uninteresting since the m-tuple zero α would retain its identity regardless of the perturbation. Similarly if $(\tau-\alpha)^k$ divided $r(\tau)$, $1 \leq k \leq m-1$, then the k-tuple zero α would persist after perturbation and the only interesting problem would be the fate of the zeros of $(\tau-\alpha)^{m-k} q(\tau) + (r(\tau)/(\tau-\alpha)^k)$. Thus we may assume without loss of generality that $(\tau-\alpha)$ does not divide $r(\tau)$, i.e. $r(\alpha) \neq 0$.

For our purposes the degree of p is presumed to be known and fixed. Since we are only interested in the zeros of p, there is no

essential loss of generality in restricting the degree of r to be no greater than $n-1$, because a small perturbation $\varepsilon\tilde{r}$ of degree n would be equivalent to some other small perturbation εr of smaller degree.

Let $\alpha+\eta$ represent a zero of the perturbed polynomial $\check{p}$:

$$\check{p}(\alpha+\eta) = 0 = \eta^m q(\alpha+\eta) + \varepsilon r(\alpha+\eta) \ . \tag{1.1}$$

Thus

$$\varepsilon = \eta^m[-q(\alpha+\eta)/r(\alpha+\eta)] \ .$$

However our interest is in expressing η in terms of ε. Since r and q are polynomials they may be expanded easily in a Taylor series about α; thus

$$\varepsilon = -\eta^m[q(\alpha)/r(\alpha)] + \text{higher order terms} \ .$$

Then

$$\eta = [(-r(\alpha)/q(\alpha))\varepsilon]^{1/m} + \text{higher order terms} \ .$$

The m different m^{th} roots define the different perturbations η corresponding to the m zeros of $\check{p}$ derived from the m-tuple zero α of p.

Thus we seem to have a series in fractional powers of ε when $m > 1$. In the next section we will indicate a rigorous justification for this result and explain a constructive method for the higher order terms.

Our overall goal is to find series that converge rapidly, since we do not want to calculate more than one or two terms. Consequently we want series that converge over the largest possible region so that

convergence will be fast in the region of interest. If the region of convergence is not much larger than the region of interest, convergence is so slow there that the series "fails" in the sense that it is not practically useful. A worse failure arises when the region of convergence does not contain all of the region of interest.

2. Classical Theory of Expansions of Algebraic Functions

In the previous section we indicated how to solve

$$f(\varepsilon,\eta) = \eta^m q(\alpha+\eta) + \varepsilon r(\alpha+\eta) = 0 \ , \tag{2.1}$$

subject to

$$\begin{cases} \deg q = n - m < n \ , \\ \deg r \leq n - 1 \ , \\ r(\alpha) \neq 0 \ , \\ q(\alpha) \neq 0 \ , \end{cases} \tag{2.2}$$

for η in terms of a series in fractional powers of ε. Now we will cite the classical results which justify our approach and explain how to construct that series.

$f(\varepsilon,\eta) = 0$ is an example of an algebraic equation defining algebraic functions ε or η in terms of the other. It is easy to get ε as a function of η; our goal is to construct η as a function of ε. We will recall certain results from standard texts, changing the notation to suit our problem, and omitting hypotheses which duplicate our assumptions (2.2).

The first result is

Weierstrass' Preparation Theorem [22, p. 105]: There is a neighborhood

$$|\varepsilon| < \rho_1 \ , \qquad |\eta| < \rho_2 \ ,$$

such that

$$f(\varepsilon,\eta) = [E_0(\varepsilon) + E_1(\varepsilon)\eta + \cdots + E_{m-1}(\varepsilon)\eta^{m-1} + \eta^m]g(\varepsilon,\eta)$$

for functions $E_0, E_1, \ldots, E_{m-1}$ which are analytic in that neighborhood and g which is analytic and never vanishes in that neighborhood. $E_0(0) = E_1(0) = \cdots = E_{m-1}(0) = 0$.

Expansions of Simple Zeros

Consider first the case of expansions of a simple zero. The next result is a consequence of the preparation theorem:

Implicit Function Theorem [22, p. 109]: When $m = 1$, then there is a neighborhood

$$|\varepsilon| < \rho_1 , \quad |\eta| < \rho_2 ,$$

such that $f(\varepsilon,\eta) = 0$ has a unique root $\eta = \eta(\varepsilon)$ for any ε in the neighborhood. $\eta(\varepsilon)$ is single valued and analytic in the neighborhood and $\eta(0) = 0$.

In other words, in the vicinity of a simple zero α, η may be expressed as a Taylor series in ε. The theorem says nothing about the size of that vicinity -- it may be quite small.

If all the zeros of p are simple, then there is a neighborhood in which the n zeros of $p(\tau) + \varepsilon r(\tau)$ are all simple and they may be expressed as n Taylor series in ε, defining n analytic functions of ε.

Given a function $\eta(\varepsilon)$ defined by the polynomial equation $f(\varepsilon,\eta) = 0$, a singular point ε_0 may be defined for our purpose as one for which the discriminant of $f(\varepsilon_0,\eta)$ vanishes. The discriminant of a polynomial with n zeros $\alpha_1, \alpha_2, \ldots, \alpha_n$ may be defined [10, p. 115] to be

$$D(\varepsilon) = \prod_{1 \le i < j \le n} (\alpha_i - \alpha_j)^2 .$$

D is a function of ε because the zeros α_i are. $D(\varepsilon)$ may also be expressed [12, p. 39] as a polynomial in the η-coefficients of $f(\varepsilon,\eta)$.

Then at a singular point ε_0, $p(\tau) + \varepsilon_0 r(\tau)$ has at least one multiple zero. Bliss [1, p. 29] shows that the radii of convergence of the n Taylor series for perturbed simple zeros are at least as large as the distance to the nearest singular point. Thus when perturbing $p(\tau)$, with all zeros simple, in the direction $r(\tau)$, the expansions in powers of ε converge for $|\varepsilon|$ at least as large as $|\varepsilon_0|$ in the nearest polynomial $p(\tau) + \varepsilon_0 r(\tau)$ on the manifold of polynomials with double zeros. When p and r are real we must remember that complex ε must be considered when computing radii of convergence.

It is usually the case, moreover, that the radius of convergence is exactly the least $|\varepsilon|$ such that $p(\tau) + \varepsilon r(\tau)$ has a double zero. Of course if p and r have some zero in common then the "series" for that zero will converge everywhere. But in the usual case when the zeros of p and r are distinct, the Taylor series which coalesce to a multiple zero of $p + \varepsilon_0 r$ can not converge for $|\varepsilon| > |\varepsilon_0|$.

Expansions from a Singular Point

What if we start from a singular point, where $p(\tau)$ has a multiple zero? The answer is contained in

Puiseux's Theorem [10, p. 118]: Let $m \geq 1$ in (2.1). Then there is a neighborhood

$$|\varepsilon| < \rho_1 , \qquad |\eta| < \rho_2 ,$$

and an integer k such that η is an analytic function of θ, where $\theta^k = \varepsilon$. The k values of θ determine k analytic functions.

Since we require that $r(\alpha) \neq 0$ we will find that there are $k = m$ distinct branches, defining m Puiseux fractional power series. As before, the radius of convergence depends on the distance to the next singular point in any of the directions εr as ε takes on complex values.

Newton's polygons may be used to transform f into a form from which it is convenient to construct the actual expansions. For details the curious may consult Bliss [1, p. 35] or Kung and Traub [40] for a modern algorithmic account; the process involves expanding $f(\varepsilon,\eta)$ in a Taylor series in both variables ε and η, and then plotting points corresponding to the terms with non-zero coefficients. Thus

$$f(\varepsilon,\eta) = q(\alpha)\varepsilon^0\eta^m + r(\alpha)\varepsilon^1\eta^0 + \text{other terms} . \tag{2.3}$$

Because our discussion is based on the constraints (2.2) the Newton polygon has the especially simple form shown in Figure VII.1. Bliss shows how to use the Newton polygon to discover the substitutions

$$\varepsilon = \theta^m \quad \text{and} \quad \eta = \theta\phi$$

which transform (2.1) to

$$\phi^m q(\alpha+\theta\phi) + r(\alpha+\theta\phi) = 0 . \tag{2.4}$$

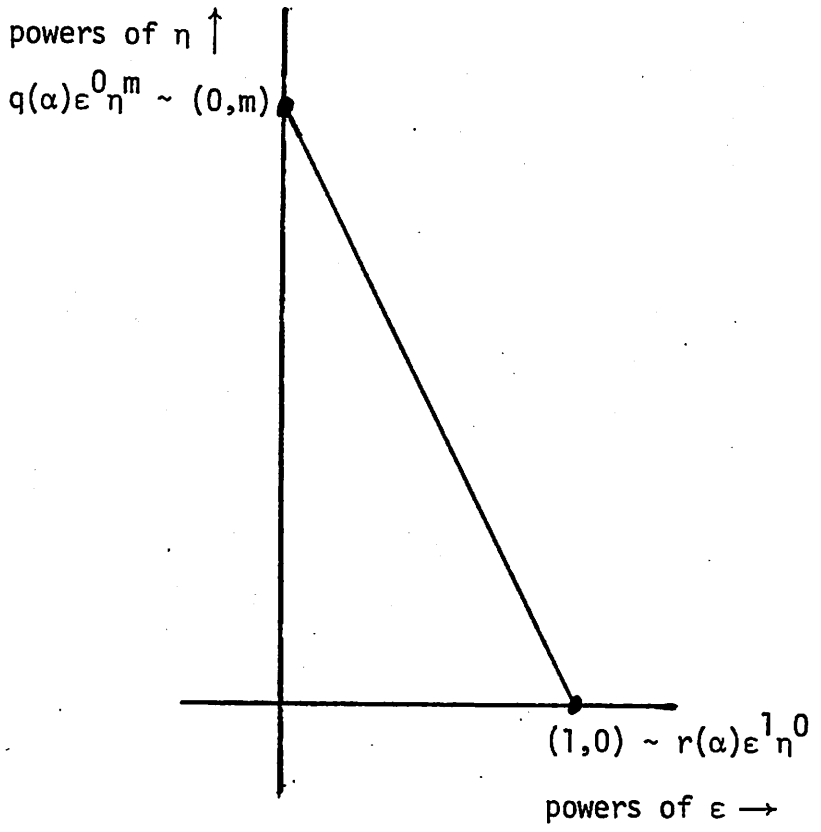

Figure VII.1. Newton's polygon for
$f(\varepsilon,\eta) = \eta^m q(\alpha+\eta) + \varepsilon r(\alpha+\eta)$,
$q(\alpha) \neq 0$, $r(\alpha) \neq 0$.

Bliss shows that all the expansions of interest are obtained from (2.4), which may be solved easily by the method of substitution or by faster methods [40] to express ϕ as a Taylor series in θ.

Define

$$x(\tau) = -r(\tau)/q(\tau)$$

and suppose

$$\phi = A + B\theta + C\theta^2 + O(\theta^3) \ ;$$

then we find that

$$\begin{aligned} A^m &= x(\alpha) \ , \\ B &= (A^2/m)(x'(\alpha)/x(\alpha)) \ , \\ C &= (A^3/2m)\{\frac{x''(\alpha)}{x(\alpha)} + (\frac{3-m}{m})(\frac{x'(\alpha)}{x(\alpha)})^2\} \ . \end{aligned}$$

It does not matter whether we use one value of A and m values of θ or vice versa. Higher order terms are tedious to derive for general m.

For $m = 1$ the expressions become

$$\eta = A\varepsilon + B\varepsilon^2 + C\varepsilon^3 + O(\varepsilon^4)$$

where

$$(2.5) \qquad \begin{cases} A = x(\alpha) \ ; \\ B = Ax'(\alpha) \ ; \\ C = A((x'(\alpha))^2 + \frac{1}{2}Ax''(\alpha) \ . \end{cases}$$

For $m = 2$, however,

$$\eta = A\varepsilon^{1/2} + B\varepsilon + C\varepsilon^{3/2} + O(\varepsilon^2)$$

where

$$(2.6)\qquad \begin{cases} A^2 = x(\alpha)\ , \\ B = \frac{1}{2}x'(\alpha)\ , \\ C = \frac{1}{4}A(x''(\alpha) + (x'(\alpha))^2/(2x(\alpha)))\ . \end{cases}$$

3. Failure of Classical Taylor and Puiseux Series Expansions

Suppose we consider perturbing the quadratic polynomial $(\tau-1)^2$ in the direction toward $(\tau-0)^2$, i.e.

$$\check{p}(\tau) = (\tau-1)^2 + \varepsilon(2\tau-1) .$$

Then the zeros of $\check{p}$ are

$$1 - \varepsilon \pm \sqrt{\varepsilon^2-\varepsilon} = 1 - \varepsilon + i\varepsilon^{1/2}(1-\varepsilon)^{1/2} .$$

We could expand $(1-\varepsilon)^{1/2}$ in a Taylor series $1 - \frac{1}{2}\varepsilon - \frac{1}{8}\varepsilon^2 \ldots$, yielding Puiseux fractional power series for the zeros; those series can not converge outside a circle of radius equal to the distance to the nearest singularity of $(1-\varepsilon)^{1/2}$. That singularity is the branch point at $\varepsilon = 1$.

Thus when we consider perturbations of p from one point on the manifold of quadratic polynomials with a double zero toward another point on that manifold, the fractional power series expansions of the perturbed double zero fail to converge rapidly as that manifold is approached. The same slow convergence occurs whenever we attempt expansions from one point on the manifold toward another point on the manifold. For practical purposes, a power series that converges slowly is worth little more than one that does not converge at all.

Figure VII.2 represents the space of monic real quadratic polynomials. Each point in the plane corresponds to such a polynomial. The coordinates of a point corresponding to

$$p(\tau) = \tau^2 + p_1\tau + p_2$$

are the coefficients p_1 and p_2. The curve is the manifold of

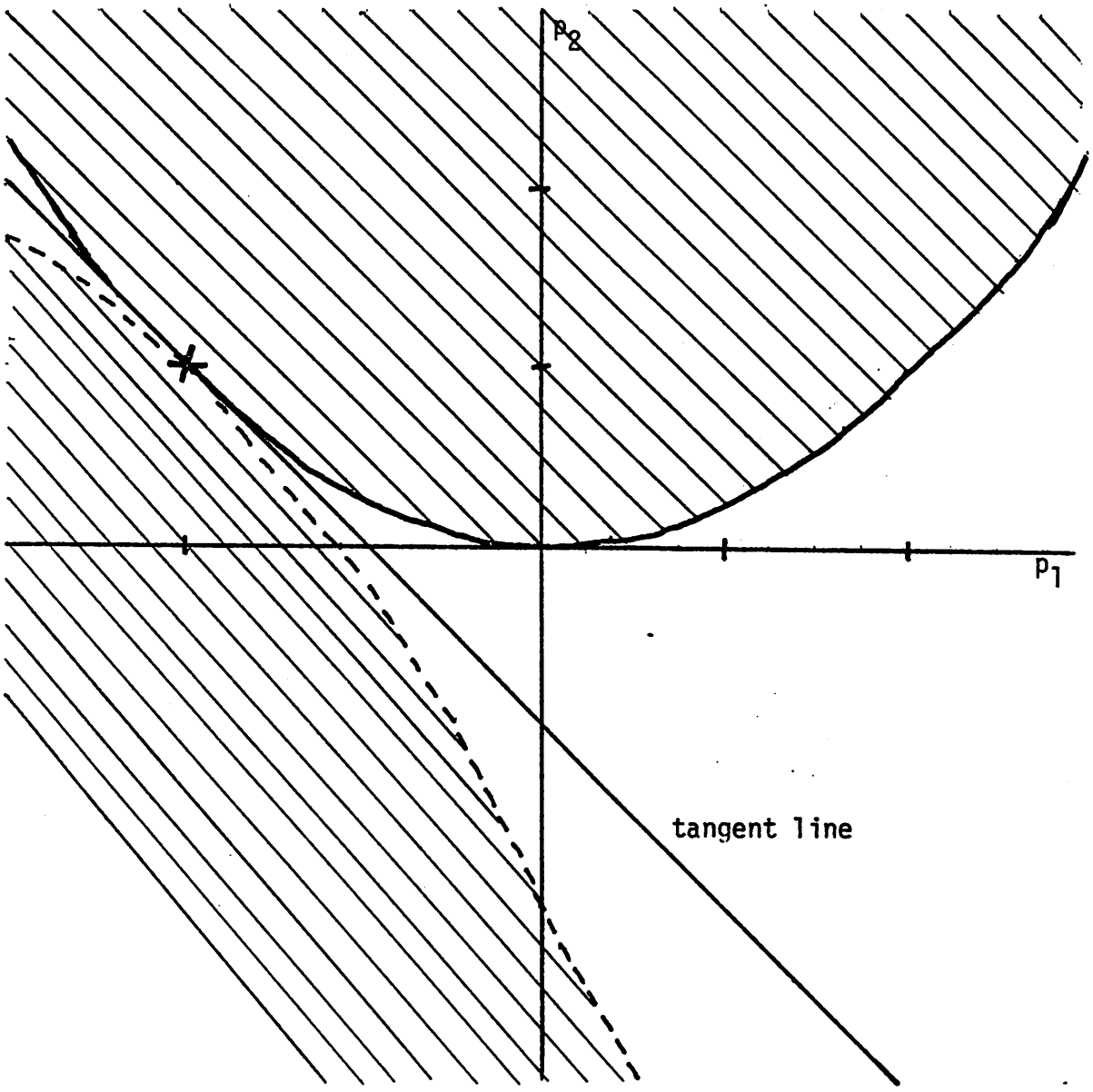

Figure VII.2. The zeros of polynomials in the shaded region may be represented by convergent Puiseux fractional power series from *.
The zeros of polynomials on the tangent line may be represented by convergent finite integral power series from *.

polynomials with double zeros; its equation is $p_1^2 = 4p_2$.

The * marks the polynomial $p(\tau) = (\tau-1)^2$ whose coordinates are $p_1 = -2$, $p_2 = 1$. We can imagine perturbing p to any other polynomial $\check{p}$ in the space; then we may ask: can the zeros of $\check{p}$ be obtained from the zeros of p by convergent Puiseux fractional power series in $\varepsilon(\check{p}-p)$? The shaded region in Figure VII.2 is the region of points $\check{p}$ for which those fractional power series do converge. That region is bounded by the union of the parabola $p_1^2 = 4p_2$ and another parabola, $p_1^2 + 8p_1 + 8 = -4p_2$, which is congruent and osculatory to the first. Puiseux fractional power series expansions from * will not converge to any point outside the shaded region. The shaded regions were determined by considering real perturbations in real directions; that turns out to be sufficient for this special case of a real quadratic with a double zero. For more general polynomials it would also be necessary to consider <u>complex</u> perturbations in order to properly delimit the shaded region.

What happens on the indicated line tangent to the manifold at *? That line represents polynomials one of whose zeros is always 1. Then the appropriate "expansions" for the two zeros of

$$(\tau-1)^2$$

when perturbed in the direction

$$\tau^2 + \rho\tau - \rho - 1$$

are 1 and $1 - \varepsilon(\rho+2)$. This finite expansion converges everywhere on the tangent line.

Notice that there are polynomials arbitrarily close to * such as

$$(\tau-1)^2 - \delta(\tau+\tfrac{1}{8}\delta - 1)$$

whose zeros can not be represented by convergent Puiseux fractional power series from *.

In contrast to the case of starting on the manifold, suppose now that we start off it, but near it. Then the regions where convergence of conventional Taylor series may occur are circumscribed indeed; see Figure VII.3 for examples.

In conclusion, we see that the classical Taylor and Puiseux series approaches for expressing changes of zeros in terms of a parameter of the perturbations is limited in applicability since neither series will converge beyond the nearest singularity of the function they represent. In our case singularities amount to double zeros. In the next section we will see how to alleviate this problem.

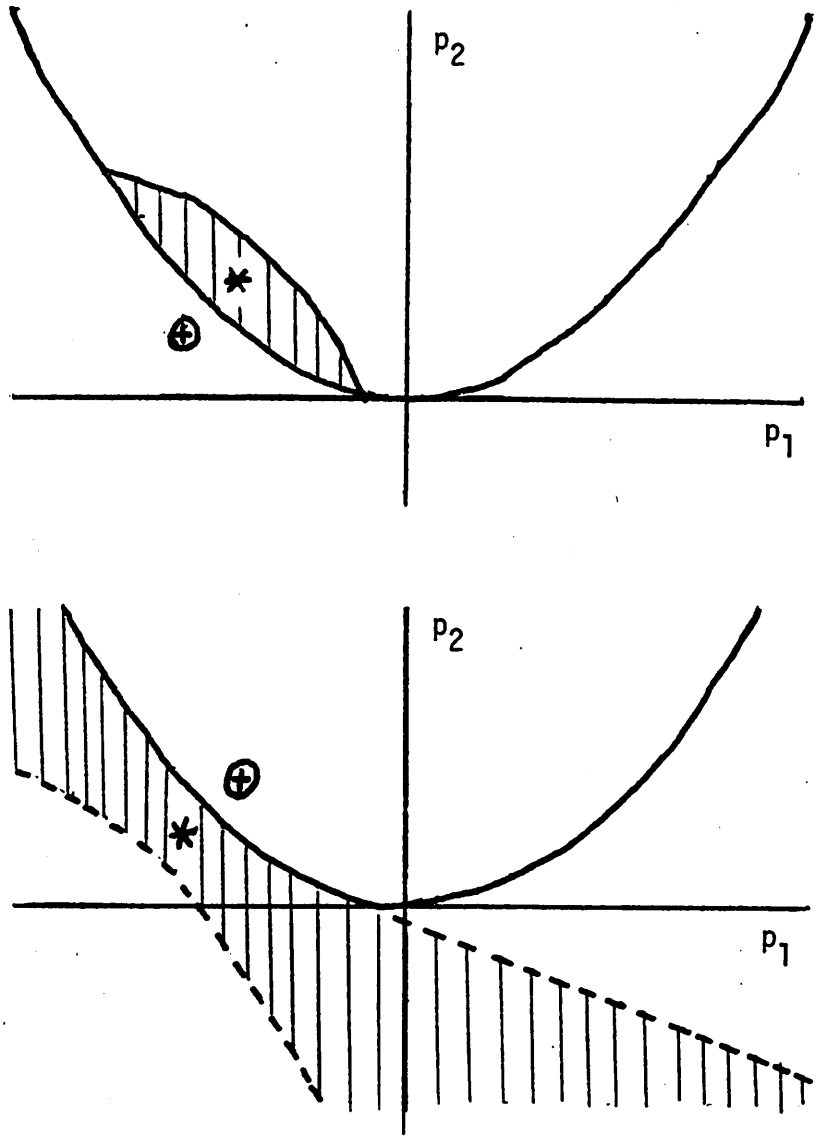

Figure VII.3. The zeros of polynomials outside the shaded regions can not be represented by convergent Taylor series from *. ⊕ marks a polynomial close to * whose zeros can not be so represented.

4. Why Find the Nearest Polynomial with a Multiple Zero?

Suppose that the output of a physical system may be modeled by the zeros of a polynomial $\hat{p}$ whose somewhat uncertain coefficients may be computed from experimental data. Suppose furthermore that polynomials with multiple zeros lie within the region of uncertainty.

We may desire to determine how the zeros of the polynomial can vary as the coefficients vary within their uncertainty. A natural way to do this is with a Taylor series expansion of the type described in section 2, but such an approach is doomed to fail when $\hat{p}$ is near a pejorative manifold. Such expansions are not valid across the manifolds of polynomials with multiple zeros. Thus we can not study the variation of the zeros of $\hat{p}$ subject to all perturbations that interest us if the ball representing our uncertainty intersects a manifold. Furthermore the convergence rate of the expansions we do have becomes unacceptable as they approach their radius of convergence. Thus we would like to find an expansion process that is convergent in a ball that is much larger than the uncertainty in $\hat{p}$. Then only 1 or 2 terms of an expansion would be needed in order to bound the variation in the zeros as $\hat{p}$ moves within its ball of uncertainty. See Figure VII.4.

In the rest of this chapter we will describe a new method for bounding variations of zeros that may be used in situations like that of Figure VII.4. This technique is based on finding a polynomial $p = \hat{p} + \delta\hat{p}$ which is close to $\hat{p}$ and has as high a multiplicity configuration as any in the ball of uncertainty. All its zeros are well conditioned, reflecting the fact that it is far from the next higher manifold. p would usually be found by one of the methods

described in chapters III-V. When such a p is found, the technique to be described exploits the manifold on which p lies to obtain bounds applicable over the entire region of interest.

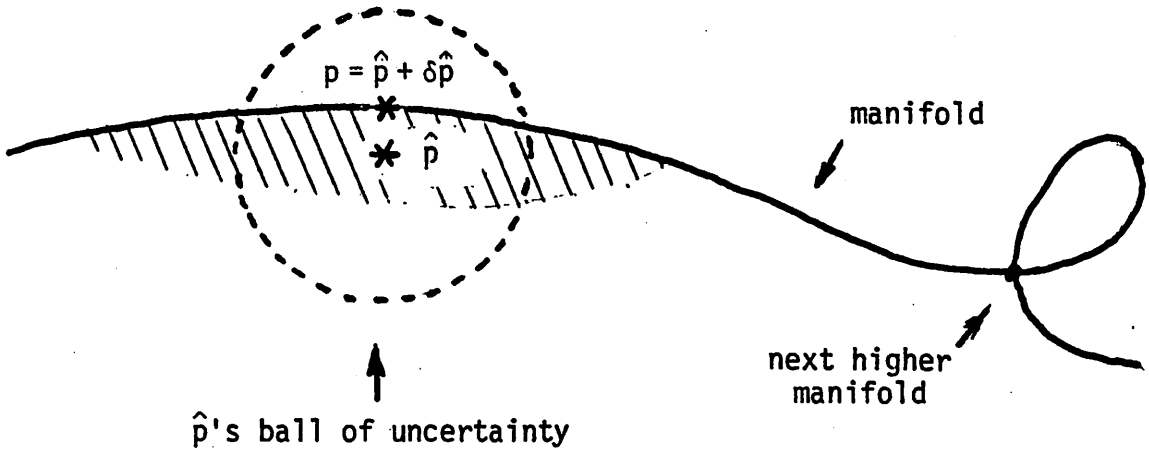

Figure VII.4. Moving to a manifold to improve the region of convergence. Taylor series expansions from $\hat{p}$ converge only in the shaded region. Puiseux fractional power series expansions from $p = \hat{p} + \delta\hat{p}$ converge in a large region as in Figure VII.2 which however omits points arbitrarily close to p. The new expansions from p converge in a region extending to the next higher manifold and including all of $\hat{p}$'s ball of uncertainty.

5. Resolving Expansions into Components

Our task now is to find a simpler method for describing the changes in the zeros of a polynomial due to perturbations.

First consider a polynomial on the manifold of polynomials with one m-tuple zero:

$$p(\tau) = (\tau-\alpha)^m q(\tau) , \qquad q(\alpha) \neq 0 .$$

We want to perturb p to another polynomial on that same manifold:

$$\tilde{p}(\tau) = (\tau-\tilde{\alpha})^m \tilde{q}(\tau) , \qquad \tilde{q}(\tilde{\alpha}) \neq 0 .$$

The classical fractional Puiseux series approach of the previous section attempts (and fails) to get from p to $\tilde{p}$ along a straight line in the space of polynomials of degree n:

$$\mathring{p}(\tau) = (\tau-\alpha)^m q(\tau) + \varepsilon[(\tau-\tilde{\alpha})^m \tilde{q}(\tau) - (\tau-\alpha)^m q(\tau)] .$$

See Figure VII.5.

We will instead move along the manifold, regarding it as a convenience rather than a barrier:

$$\dot{p}(\tau) = [\tau - (\alpha + \varepsilon(\tilde{\alpha}-\alpha))]^m [q(\tau) + \varepsilon(\tilde{q}(\tau)-q(\tau))] .$$

Now the multiple zero stays multiple, and the change in the multiple zero may be easily expressed as a function of ε. If the multiple zero is $\alpha+\eta$ then

$$\eta = (\tilde{\alpha}-\alpha)\varepsilon$$

which is certainly convergent for <u>all</u> ε. The changes in the other zeros are described by Taylor series in the classical manner. These

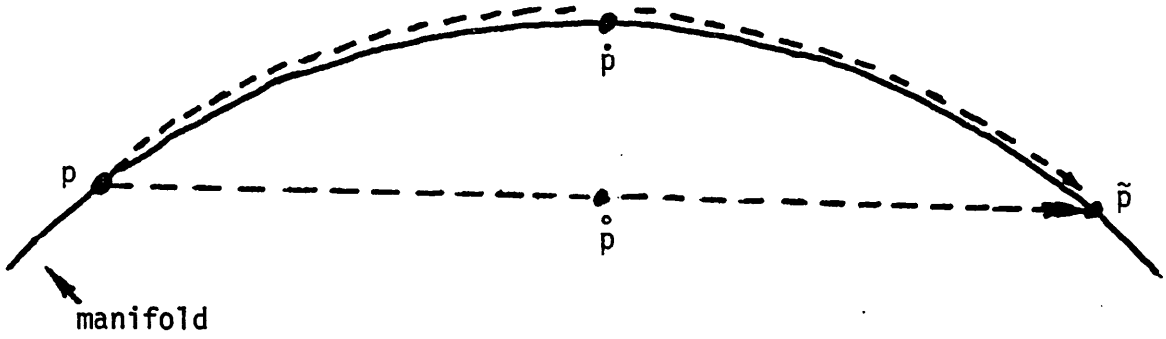

Figure VII.5. Two ways to get from p to $\tilde{p}$. The classical Puiseux expansion goes directly via $\mathring{p}$. The new expansion goes along the manifold via $\dot{p}$.

Taylor series will converge in some region in the space of polynomials of degree n-m. That region is determined by the locations of manifolds of polynomials with multiple zeros in the n-m dimensional space. These manifolds correspond to manifolds of polynomials with more than one multiple zero in the original n dimensional space.

For a specific example, if we start with a polynomial with a double zero, so $m = 2$, we can expand the zeros along the manifold until we reach a submanifold containing polynomials with two double zeros, or one quadruple zero, or some other configuration that implies a multiple zero in $q+\varepsilon(\tilde{q}-q)$. A submanifold of polynomials with a single <u>triple</u> zero, however, would have no effect on the expansion, for a triple zero in $\dot{p}$ implies only a simple zero in $q+\varepsilon(\tilde{q}-q)$.

Obviously this approach can be extended to polynomials with several multiple zeros. To get from

$$p(\tau) = (\prod_i(\tau-\alpha_i)^{m_i})q(\tau)$$

to

$$\tilde{p}(\tau) = (\prod_i(\tau-\tilde{\alpha}_i)^{m_i})\tilde{q}(\tau)$$

just let

$$\dot{p}(\tau) = (\prod_i(\tau - (\alpha_i + \varepsilon(\tilde{\alpha}_i-\alpha_i)))^{m_i})\cdot(q(\tau) + \varepsilon(\tilde{q}(\tau)-q(\tau))) \ .$$

Suppose now that we wish to expand from a polynomial on a manifold to a polynomial off that manifold. As we saw in the previous section, a straight Taylor series expansion may be limited in applicability by the presence of the same or other manifolds. From our present vantage point it appears that the procedure most likely to

succeed would be to expand along the manifold to get as close as possible to the off-manifold polynomial we seek, and then expand "orthogonally" directly from the manifold to that point with Taylor series. We would thus minimize the effect of nearby manifolds on the convergence of the Taylor series. Figure VII.6 illustrates the notion.

There may still be no reasonable way to expand from p to <u>every</u> polynomial of degree n. For instance consider the situation in Figure VII.7. A self-intersection singularity, corresponding to a polynomial with two double zeros, means that it is impossible to expand from p to $\tilde{p}$. If our problem were, however, to expand from y to $\check{p}$, it might be possible to do so by finding a $\tilde{p}$ on y's manifold of polynomials with two double zeros.

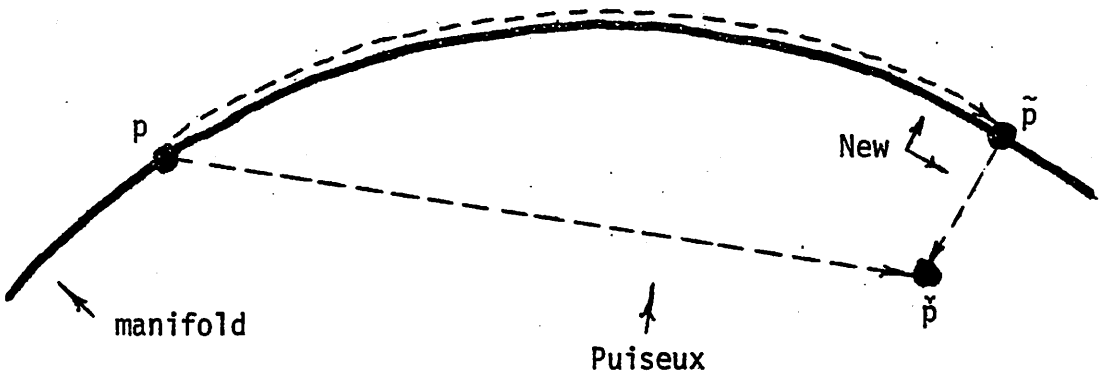

Figure VII.6. Two ways to get from p to $\check{p}$.

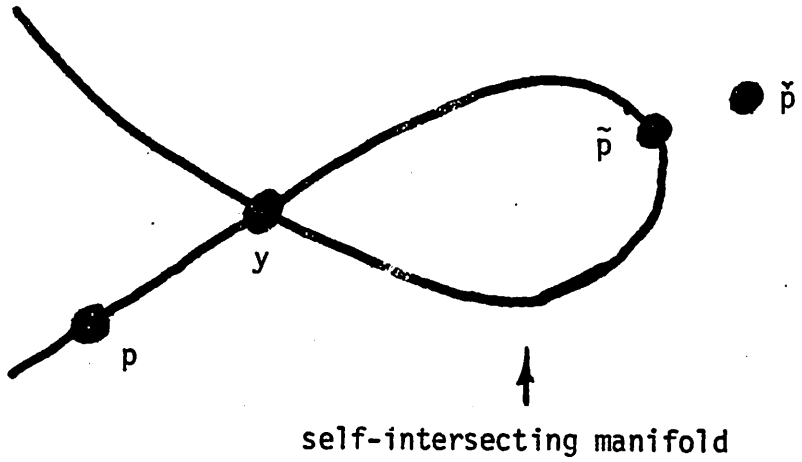

Figure VII.7. There is no reasonable way to expand from p to $\check{p}$, or even to $\tilde{p}$.

6. A Practical Technique for Bounding Changes in Zeros

In the previous section we introduced the notion of expanding along a manifold before resorting to conventional Taylor or Puiseux series techniques. In order to have a technique usable for bounding changes in zeros as coefficients vary, we need to overcome two problems:

1) Apparently it is necessary to solve the problem of finding $\tilde{p}$, the nearest point on the manifold, for every $\check{p}$ for which we want an expansion. As we have seen this is a difficult numerical problem that is even more intractable symbolically.

2) Our expansions have always been defined in terms of a direction $r(\tau)$ and a size parameter ε. We would like to state the expansion directly in terms of the perturbing polynomial without introducing the additional parameter ε.

The second problem may be solved fairly easily by letting ε go to 1 at the end or by ignoring ε altogether. We find that the term that was attached to the k^{th} power of ε contains powers of r that are always greater than or equal to k, and thus we can construct a series in r -- whether r is represented by its coefficients, its zeros, or the value of r and its derivatives at some point. The next section contains examples of such series.

As for the first problem, we might settle for $\tilde{s}$, an approximation to $\tilde{p}$ that can be expressed symbolically. $\tilde{s}$ should be a satisfactory substitute in regions where the manifold is not too wild.

Figure VII.8 illustrates the approximation. Instead of $\tilde{p}$ we could compute a projection $\hat{s}$ of $\check{p}$ on a tangent surface and map $\hat{s}$ to a polynomial $\tilde{s}$ on the manifold. We hope that $\tilde{s}$ is reasonably close to $\tilde{p}$.

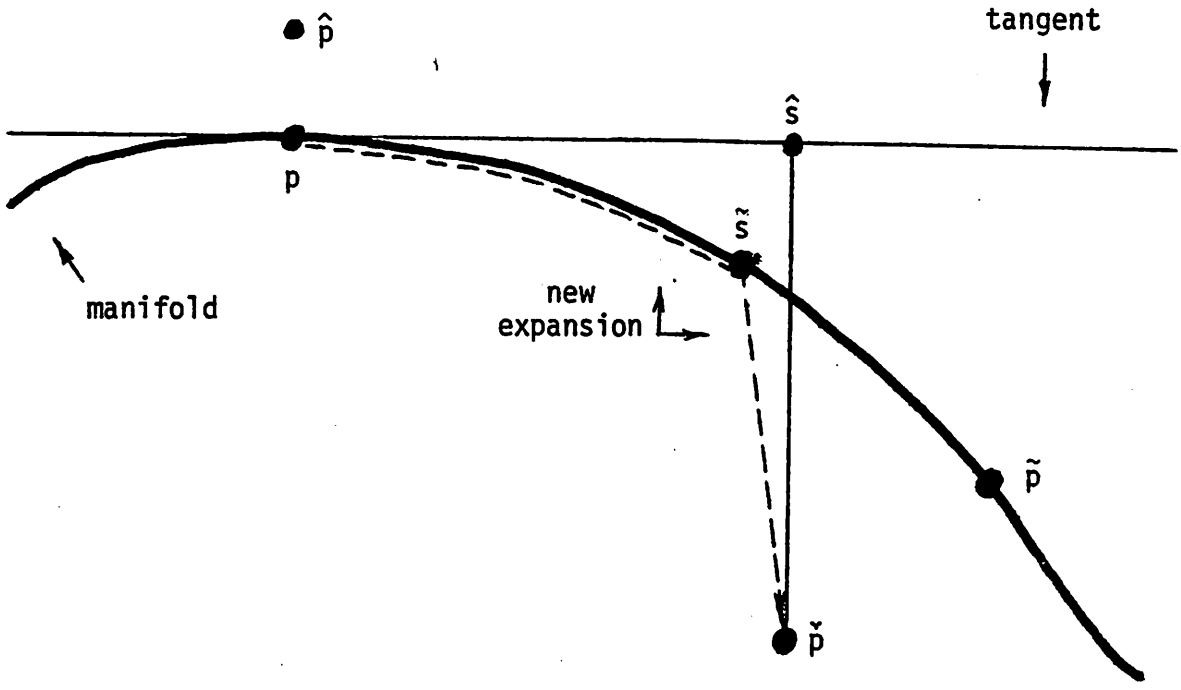

Figure VII.8. As a practical matter, the new expansion must get from p to $\check{p}$ via $\tilde{s}$ rather than $\tilde{p}$. $\hat{p}$ is a polynomial for which p is the closest polynomial on the manifold.

Given $\check{p}$ and p, $\hat{s}$ is uniquely determined by the norm, but there are many possible ways of mapping from the tangent surface to the manifold. Unfortunately there is no simple way of insuring that $\tilde{s} = \check{p}$ when $\check{p}$ is already on the manifold. Any discrepancy in this case is intolerable because it leads to the situation in Figure VII.9 with its familiar problem of short radii of convergence.

Any expansion technique for arbitrary $\check{p}$ must somehow recognize when $\check{p}$ is on the manifold. A vanishing discriminant is an example of a condition characterizing polynomials on the manifold. But such characterizations are too complicated to be useful.

The notion of expanding along the manifold may still be put to good use, however, if we only seek bounds on changes in zeros rather than explicit expansions in terms of a perturbation. Thus given p with zeros θ_i of various multiplicities, we may ask for bounds on

$$|\theta_i - \check{\theta}_i|$$

for zeros $\check{\theta}_i$ of polynomials $\check{p}$ such that $\|p-\check{p}\| \leq \Delta$. See Figure VII.10. The variation of $\check{\theta}_i$ with respect to θ_i can be thought of as having two components, one due to motion on the manifold and one due to motion orthogonal to the manifold. If we can bound these changes separately and independently then we can add the bounds to get the overall variation.

Taking a closer look at the components of $\check{p} - p$, recall that

$$p(\tau) = (\tau-\alpha)^m q(\tau) ,$$
$$\tilde{p}(\tau) = (\tau-\tilde{\alpha})^m \tilde{q}(\tau) ,$$

where

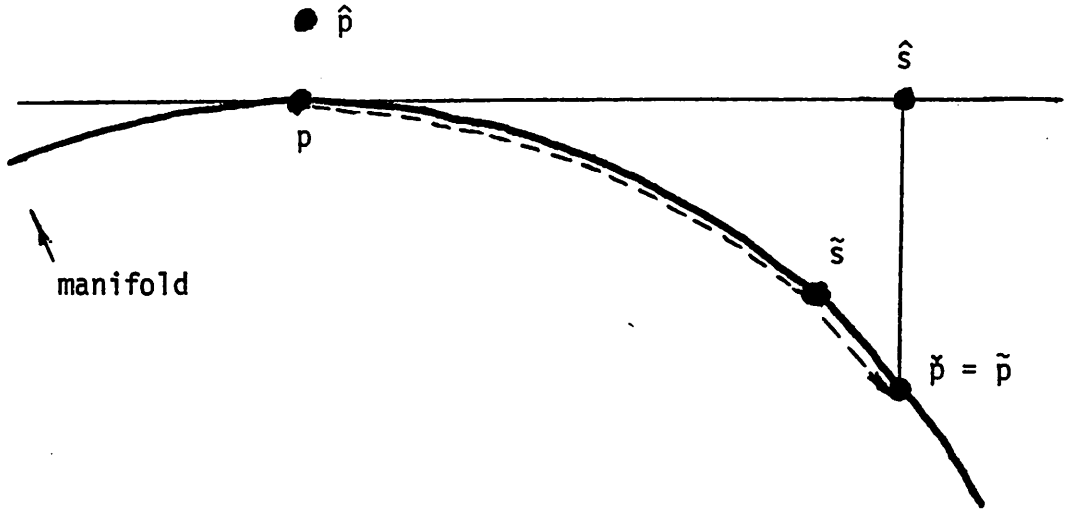

Figure VII.9. Shortcoming in revised expansion method when $\check{p}$ is on the manifold. The Puiseux expansion from $\tilde{s}$ to $\check{p}$ is doomed to have a short radius of convergence.

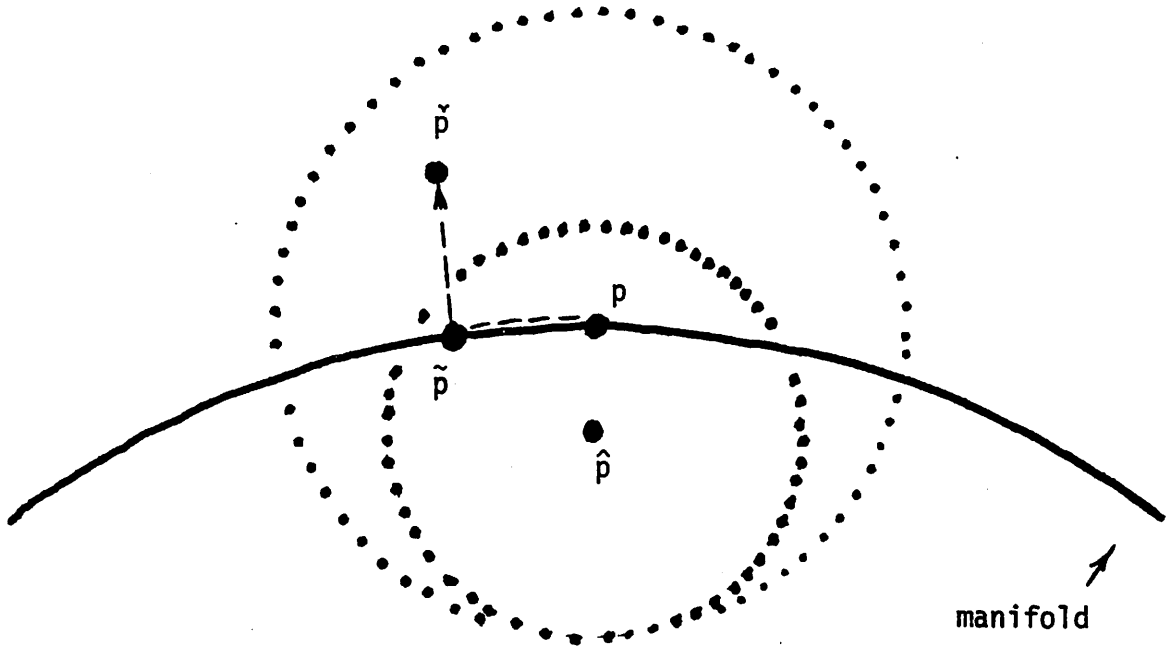

Figure VII.10. How do the zeros vary as $\check{p}$ varies within the small ball centered on $\hat{p}$? A bound may be computed by studying the variation in the zeros as $\check{p}$ varies within the larger ball centered on p.

$$\tilde{\alpha} = \alpha + \delta\alpha$$

$$\tilde{q} = q + \delta q \ .$$

q is a monic polynomial of degree $n-m$; δq is not monic and is of degree at most $n-m-1$. Then

$$\tilde{p} - p = (\tau-\alpha)^m \delta q(\tau) + \sum_{j=1}^{m} \binom{m}{j} (\tau-\alpha)^{m-j} (q+\delta q)(\tau)(-\delta\alpha)^j$$

where $\binom{m}{j} = m!/(j!(m-j)!)$. We will mostly be interested in the infinitesimal case for which we need not be concerned about the higher order terms.

Summary of the New Technique

Before looking at details we summarize the new technique.

We are given a polynomial $\hat{p}$ with a norm and a bound on the uncertainty in $\hat{p}$. We want a bound on the corresponding uncertainty in the zeros of $\hat{p}$.

The ball representing polynomials practically indistinguishable from $\hat{p}$ contains some polynomials $\tilde{p}$ with multiple zeros. By the numerical means discussed in chapters III to V, we locate the polynomial p nearest to $\hat{p}$ with all zeros well conditioned; some are therefore multiple. Then we may determine a ball about p that contains the original ball about $\hat{p}$ and which is usually only slightly larger. Then we may bound the variation in the zeros of polynomials $\check{p}$ in this second ball.

To do so we first construct symbolic expansions for the changes in the zeros of p due to moving to another polynomial $\tilde{p}$ on the

same manifold but within the second ball (Figure VII.10). For the multiple zeros $\tilde{\alpha}$ these expansions from α have only two terms but for the simple zeros $\tilde{\beta}$ these expansions from β are Taylor series in the perturbation δq.

Now we compute expansions from $\tilde{p}$ to points $\check{p}$ which lie on the planes normal to the manifold at $\tilde{p}$. These symbolic expansions are Puiseux fractional power series to get zeros $\check{\alpha}$ from the multiple zeros $\tilde{\alpha}$ and Taylor series to get zeros $\check{\beta}$ from simple zeros $\tilde{\beta}$. The series are in $\check{p} - \tilde{p}$ which is orthogonal to the manifold at $\tilde{p}$.

Then we substitute, again symbolically, the series for $\tilde{\alpha}$ and $\tilde{\beta}$ in the second sets of series to obtain series for $\check{\alpha}$ and $\check{\beta}$ which do not contain $\tilde{\alpha}$ or $\tilde{\beta}$. Finally we may convert the numerical bound Δ on the size of the second ball into numerical bounds on the terms of the series for $\check{\alpha}$ and $\check{\beta}$.

It is essential to study an example to understand the technique. The example given in the next section is simplified but contains the essential ideas.

The method just described ought to be compared to one based on the results of Brian Smith [42]. Smith uses Gerschgorin circles to obtain bounds for the zeros of a polynomial subject to uncertainty in its coefficients. Smith's bounds are easier to compute than those based on expansions, but they may be unrealistic by a factor that is proportional to the degree of the polynomial. However, they are valid for finite as well as infinitesimal perturbations, unlike the new method. Comparative evaluation of the two bounding methods must be postponed until the new bounds can be computed automatically.

Notation

Recall the vector notation of chapter I. We will represent q by a vector of dimension $n-m+1$ and δq by a vector of dimension $n-m$. Corresponding to polynomial multiplication of δq by $(\tau-\alpha)$ define

$$P_1 = \underbrace{\begin{pmatrix} 1 & & & \\ -\alpha & 1 & & 0 \\ & -\alpha & \ddots & \\ & & \ddots & 1 \\ & 0 & & -\alpha \end{pmatrix}}_{n-m} \Bigg\} n-m+1 \quad .$$

Then corresponding to polynomial multiplication of q or $P_1\delta q$ by $(\tau-\alpha)^{m-1}$ define

$$P_{m-1} = \underbrace{\begin{pmatrix} 1 & & & 0 \\ -(m-1)\alpha & \ddots & & \\ \vdots & \ddots & 1 & \\ \vdots & & -(m-1)\delta & \\ (-\alpha)^{m-1} & & \vdots & \\ & \ddots & \vdots & \\ 0 & & (-\alpha)^{m-1} & \end{pmatrix}}_{n-m+1} \Bigg\} n \quad .$$

Then to first order

$$\tilde{p} - p = P_{m-1}P_1\delta q - mP_{m-1}q\delta\alpha \quad .$$

In chapter VIII we will see that an "orthogonal" perturbation to $\tilde{p}$ has the form

$$\check{p} - \tilde{p} = W^{-1}\tilde{A}^*\delta\ell$$

where $\tilde{A}$ is the m-1 by n matrix

$$\tilde{A} = \begin{pmatrix} \tilde{e}^* \\ \tilde{e}^* D \\ \vdots \\ \tilde{e}^* D^{m-2} \end{pmatrix}$$

$\tilde{e}^* = (\tilde{\alpha}^{n-1}\ \tilde{\alpha}^{n-2} \cdots \tilde{\alpha}\ 1)$, which depends on $\tilde{\alpha}$, hence the $\tilde{}$ in $\tilde{A}$. This $\tilde{A}$ should not be confused with the m by n+1 matrix $\tilde{A}$ of chapters III, IV, and V. A or e without $\tilde{}$ means $\tilde{\alpha} = \alpha$. $\delta\ell$ is an m-1 vector which is infinitesimal like δq and $\delta\alpha$. To first order $W^{-1}\tilde{A}^*\delta\ell = W^{-1}A^*\delta\ell$, so

$$p = \check{p} - p \doteq W^{-1}A^*\delta\ell + P_{m-1}P_1\delta q - mP_{m-1}q\delta\alpha$$

$$= \begin{pmatrix} W^{-1}A^* \\ P_{m-1}P_1 \\ -mP_{m-1}q \end{pmatrix} \begin{pmatrix} \delta\ell \\ \delta q \\ \delta\alpha \end{pmatrix} \equiv M\delta h \ .$$

The matrix operator M is n by n and invertible so a specific infinitesimal perturbation δp may be mapped into $\delta\ell$, δq, and $\delta\alpha$, the components of δh.

We would like to define a region in δh-space whose image, mapped into δp-space, is the ball $\|\delta p\|_W \leq \Delta$. Obviously that region is just

$$\{\delta h \mid \|\delta h\|_H \leq \Delta\}$$

where $\|\delta h\|_H \equiv \|M\delta h\|_W$. For infinitesimals with quadratic norms this approach is practical.

Best Possible Bounds for Changes in Zeros Due to Variations Over an Infinitesimal Ball

To see how to get the infinitesimal bounds in a series expansion, let

$$\|\delta p\|_W^2 = \delta p^* W \delta p = \delta h^* M^* W M \delta h = \delta h^* H \delta h = \|\delta h\|_H^2 \tag{6.1}$$

where

$$H \equiv \begin{pmatrix} AW^{-1}A^* & 0 & 0 \\ 0 & P_1^* X P_1 & -mP_1^* X q \\ 0 & -mq^* X P_1 & m^2 q^* X q \end{pmatrix}$$

and

$$X \equiv P_{m-1}^* W P_{m-1} .$$

The zero entries in H arise because $AP_{m-1} = 0$.

Suppose we want to compute the first two terms of an infinitesimal bound for the zeros $\check{\alpha}$ of

$$\check{p}(\tau) = p(\tau) + \delta p(\tau) = (\tau-\alpha)^2 q(\tau) + \delta p(\tau) .$$

The change due to the move from p to $\tilde{p}$ is just $\delta\alpha$. The orthogonal direction is $W^{-1}A^*\delta\ell = W^{-1}e\delta\lambda$ where e^* is the evaluation functional for α and $\delta\lambda$ is a scalar. Then using (2.6),

$$x(\tau) = \delta\lambda \frac{W^{-1}e(\tau)}{q(\tau)} ,$$

$$\check{\alpha} - \tilde{\alpha} = \sqrt{x(\alpha)} + \frac{1}{2}x'(\alpha) + \cdots .$$

But $x(\alpha) = \delta\lambda(W^{-1}e(\alpha))/q(\alpha)$ which is just a constant γ_1 times $\delta\lambda$. Likewise $x'(\alpha)$ is just a different constant γ_2 times $\delta\lambda$. Thus

$$\check{\alpha} - \alpha = \sqrt{\gamma_1}(\delta\lambda)^{1/2} + (\gamma_2\delta\lambda+\delta\alpha) + \cdots .$$

How large can these terms become, given that $\|\delta h\|_H \leq \Delta$? The maximum value of $|\delta\lambda|^2$ is $\Delta^2/(e_\alpha^* W^{-1} e_\alpha)$ so for the first term,

$$|\sqrt{\gamma_1}(\delta\lambda)^{1/2}| \leq \sqrt{\frac{|\gamma_1|}{(e_\alpha^* W^{-1} e_\alpha)^{1/2}}}\,\Delta^{1/2} .$$

As for the second term,

$$|(\gamma_2\ 0\ 1)\begin{pmatrix}\delta\lambda\\ \delta q\\ \delta\alpha\end{pmatrix}| \leq \|(\gamma_2\ 0\ 1)\|_H\Delta = \sqrt{(\gamma_2\ 0\ 1)H^{-1}\begin{pmatrix}\gamma_2^*\\ 0\\ 1\end{pmatrix}}\,\Delta .$$

Such bounds are achievable by δh satisfying $\|\delta h\|_H \leq \Delta$ and so are best possible.

A Region Circumscribing an Infinitesimal Ball

The method just outlined is best possible for perturbations that are infinitesimal or essentially so. Sometimes we may be content with bounds that are not optimal but hopefully are realistic.

To that end rewrite (6.1) as

$$\|\delta p\|_W^2 = \delta g^* V \delta g$$

where

$$V = \begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & v\\ 0 & v^* & 1\end{pmatrix},$$

$$v = -\frac{(P_1^* X P_1)^{-1/2} P_1^* X q}{(q^* X q)^{1/2}},$$

and

$$\delta g = \begin{pmatrix} W^{-1/2}A^*\delta\ell \\ (P_1^*XP_1)^{1/2}\delta q \\ m(q^*Xq)^{1/2}\delta\alpha \end{pmatrix} .$$

Then we might let

$$\begin{aligned} \|W^{-1/2}A^*\delta\ell\|_2 &= \|W^{-1}A^*\delta\ell\|_W \le \Delta , \\ \|(P_1^*XP_1)^{1/2}\delta q\|_2 &= \|P_{m-1}P_1\delta q\|_W \le \Delta , \\ \|m(q^*Xq)^{1/2}\delta\alpha\|_2 &= m\|P_{m-1}q\|_W|\delta\alpha| \le \Delta ; \end{aligned}$$

but depending on v, we might find that the image of the region so defined does not contain the entire ball $\|\delta p\|_W \le \Delta$. If $q = X^{-1}e_\alpha$ then $v = 0$ and the image is just the ball, while if $q = P_1 u$ then $\|v\| = 1$ and the image is not an n-dimensional ball or ellipsoid but something of lower dimension which can not possibly contain the ball.

To see what is going on, suppose $\|\delta p\|_W = \Delta$ exactly and $\delta\ell = 0$. How large can $\delta\alpha$ and δq become? We have

$$\Delta^2 = \begin{pmatrix} (P_1^*XP_1)^{1/2}\delta q \\ m(q^*Xq)^{1/2}\delta\alpha \end{pmatrix}^* \begin{pmatrix} 1 & v \\ v^* & 1 \end{pmatrix} \begin{pmatrix} (P_1^*XP_1)^{1/2}\delta q \\ m(q^*Xq)^{1/2}\delta\alpha \end{pmatrix}$$

so

$$\delta q^*(P_1^*XP_1)\delta q + m^2 q^*Xq|\delta\alpha|^2 = \Delta^2/\text{minev}$$

where "minev" means the smallest eigenvalue of

$$\begin{pmatrix} 1 & v \\ v^* & 1 \end{pmatrix} .$$

But the eigenvalues of that matrix are just 1, of multiplicity n-2, $1 - \|v\|_2$, and $1 + \|v\|_2$. So at worst

$$|\delta\alpha|^2 \le \frac{\Delta^2}{m^2 q^* X q (1-\|v\|_2)}$$

$$\|P_{m-1}P_1\delta q\|_W^2 \le \frac{\Delta^2}{1-\|v\|_2} ,$$

where

$$\|v\|_2^2 = \frac{q^*XP_1(P_1^*XP_1)^{-1}P_1^*Xq}{q^*Xq} .$$

Therefore our constraints should read

$$(6.2)\quad \begin{cases} \|\delta\ell\|_L = \|W^{-1}A^*\delta\ell\|_W \le \Delta_\ell \equiv \Delta , \\ \|\delta q\|_Q = \|P_{m-1}P_1\delta q\|_W \le \Delta_q \equiv \Delta/(1-\|v\|_2)^{1/2} , \\ |\delta\alpha| \le \Delta_\alpha \equiv \Delta/(m\|P_{m-1}q\|_W(1-\|v\|_2)^{1/2}) . \end{cases}$$

The image of such an infinitesimal region does indeed contain the ball $\|\delta p\|_W \le \Delta$, and in fact circumscribes it; the question remains: how much larger is the image than the ball? If $\delta\ell$, δq, and $\delta\alpha$ have bounds Δ_ℓ, Δ_q, and Δ_α in the proper norms, then

$$\|M\delta h\|_W \le \Delta\{1 + 2\frac{1+\|v\|_2}{1-\|v\|_2}\}^{1/2} .$$

Thus bounds based on (6.2) will be realistic if and only if $\|v\|_2 << 1$.

It turns out that $\|v\|_2 << 1$ if and only if $P_{m-1}q$, which has an m-1-tuple zero α, is far from the nearest polynomial $P_{m-1}P_1u$ with an m-tuple zero α. To see this, solve the least squares problem "find u to minimize $\|P_{m-1}q - P_{m-1}P_1u\|_W$" to get

$$u = (W^{1/2}P_{m-1}P_1)^+W^{1/2}P_{m-1}q$$

so

$$\begin{aligned}\|P_{m-1}q - P_{m-1}P_1u\|_W^2 &= q^*P_{m-1}{}^*W^{1/2}(1 - W^{1/2}P_{m-1}P_1(W^{1/2}P_{m-1}P_1)^\dagger)W^{1/2}P_{m-1}{}^*q \\ &= q^*Xq - q^*XP_1(P_1{}^*XP_1)^{-1}P_1{}^*Xq \\ &= q^*Xq(1 - \|v\|_2^2)\end{aligned}$$

and

$$\|v\|_2^2 = 1 - \frac{\|P_{m-1}q - P_{m-1}P_1u\|_W^2}{\|P_{m-1}q\|_W^2} \quad .$$

Recall from section II.3 that the condition number γ of the multiple zero α is inversely related to the distance to the next higher manifold. In fact, from the definition of condition number in II.4 we know

$$\gamma \geq \frac{1}{m}\frac{1}{|q(\alpha)|}\frac{|y(\alpha)|}{\|P_{m-1}y\|_W}$$

for any y of degree $n-m$ or less. Take $y = q - P_1u$ in particular to see

$$\gamma \geq 1/(m\|P_{m-1}q - P_{m-1}P_1u\|_W)$$

whence

$$\begin{aligned}1/(1 - \|v\|_2) &\leq m^2\|P_{m-1}q\|_W^2(1 + \|v\|_2)\gamma^2 \\ &= 2m^2\|P_{m-1}q\|_W^2\gamma^2 \quad .\end{aligned}$$

Thus we have demonstrated the

Proposition. If the condition number of α is small then the image of the infinitesimal region defined by (6.2) is not much larger then the infinitesimal ball $\|\delta p\|_W \leq \Delta$.

Bounds for Changes in Zeros Due to Variation Over a Region Circumscribing a Ball

When it is inconvenient to bound the changes in the zeros by use of (6.1) we can resort to (6.2). If the zero α is well conditioned and the ball is not too big then we have confidence that the error bounds we derive are not much larger than necessary.

So suppose that Δ_ℓ, Δ_α, and Δ_q bound $\delta\ell$, $\delta\alpha$, and δq. How can the zeros of p vary subject to these bounds? Let α be the multiple zero and β a simple zero of q. First consider possible changes due to motion along the manifold. Let $\tilde{\alpha}$ and $\tilde{\beta}$ denote corresponding zeros of a polynomial $\tilde{p}$ along the manifold. Trivially

$$|\tilde{\alpha}-\alpha| \leq \Delta_\alpha .$$

To get $\tilde{\beta}$ it is necessary to construct a Taylor series expansion. β is a simple zero of q; $\tilde{\beta}$ a simple zero of $q+\delta q$. Let $q(\tau) = (\tau-\beta)q_\beta(\tau)$ and

$$x(\tau) \equiv -\delta q(\tau)/q_\beta(\beta)$$

as in (2.5). Then

$$\tilde{\beta} - \beta = x(\beta) + x(\beta)x'(\beta) + \cdots$$

$$|\tilde{\beta}-\beta| \leq |x(\beta)| + |x(\beta)||x'(\beta)| + \cdots ,$$

We can use $\|\delta q\|_Q \leq \Delta_q$ to obtain bounds for these terms. For instance,

$$\delta q(\beta) = e^*_\beta \delta q$$

where e^*_β is the functional that evaluates a polynomial at β. Then

$$|\delta q(\beta)| \le \|e_\beta^*\|_Q \|\delta q\|_Q \le \|e_\beta^*\|_Q \Delta_q \ .$$

Now

$$\|e_\beta^*\|_Q = \|e_\beta^*(P_1^* P_{m-1}^* W P_{m-1} P_1)^{-1} e_\beta\|_2$$

which is a constant that may be evaluated. So

$$|x(\beta)| \le \frac{\|e_\beta^*\|_Q}{|q_\beta(\beta)|} \Delta_q$$

and succeeding terms may be calculated in the same way. The bounds can be calculated with just a few terms if $q_\beta(\beta)$ is not too small. Thus we may bound the change in α and β due to movements along the manifold.

Next to consider are changes due to movements orthogonal to the manifold. Suppose we are at

$$\tilde{p}(\tau) = (\tau-\tilde{\alpha})^m \tilde{q}(\tau) \ ,$$

and $\tilde{\beta}$ is a zero of $\tilde{q}$. Then an orthogonal perturbation is $W^{-1}\tilde{A}^*\delta\ell$.

To see what happens to $\tilde{\alpha}$, use a formula such as (2.6). First define

$$x(\tau) = -(W^{-1}\tilde{A}^*\delta\ell)(\tau)/\tilde{q}(\tau) \ ;$$

then for $\check{\alpha}$, a zero of $\check{p} = \tilde{p} + W^{-1}\tilde{A}^*\delta\ell$,

$$\check{\alpha} - \tilde{\alpha} = (x(\tilde{\alpha}))^{1/m} + (x(\tilde{\alpha}))^{2/m} \cdot x'(\alpha)/(mx(\alpha)) + \cdots \ .$$

Now

$$x(\tilde{\alpha}) = -\tilde{e}^* W^{-1}\tilde{A}^*\delta\ell/\tilde{q}(\tilde{\alpha}) \ ,$$

$$|x(\tilde{\alpha})| \le |\tilde{e}^* W^{-1}\tilde{A}^*\delta\ell|/|\tilde{q}(\tilde{\alpha})| \ .$$

If $\tilde{\beta}_i$ are the zeros of q, then $|\tilde{q}(\tilde{\alpha})| = \Pi|\tilde{\alpha}-\tilde{\beta}_i|$.

A lower bound may be calculated by using

$$|\tilde{\alpha}-\tilde{\beta}_i| \geq |\alpha-\beta_i| - \Delta_\alpha - \Delta\beta_i$$

where $\Delta\beta_i$ is the bound for $|\tilde{\beta}_i-\beta_i|$ computed previously.

As for the other term,

$$|\tilde{e}^*W^{-1}\tilde{A}^*\delta\ell| \leq \|\tilde{e}^*W^{-1}\tilde{A}^*\|_L \|\delta\ell\|_L \leq \|\tilde{e}^*W^{-1}\tilde{A}^*\|_L \cdot \Delta_\ell \ ;$$

$$\|\tilde{e}^*W^{-1}\tilde{A}^*\|_L = \|\tilde{e}^*W^{-1}\tilde{A}^*(AW^{-1}A^*)^{-1}\tilde{A}W^{-1}\tilde{e}\|_2 \ .$$

Since $|\alpha| - \Delta_\alpha \leq |\tilde{\alpha}| \leq |\alpha| + \Delta_\alpha$ we can compute a bound for $|x(\tilde{\alpha})|$ and for the other terms of $|\check{\alpha}-\tilde{\alpha}|$.

Similarly we can compute a bound for $|\check{\beta}-\tilde{\beta}|$ for $\tilde{\beta}$, one of the other zeros of $\tilde{q}$. The process is similar to that for $|\tilde{\beta}-\beta|$.

Obviously these derivations would be much less tedious if a suitable algebraic manipulation system were available to do part of the work.

So far it may not be apparent that the process described is much of an improvement. A simple example in the next section shows that the payoff can be substantial.

7. An Example of Expansions

We will apply both the classical and the new expansion techniques to an example. It will become evident that the new expansion technique is very much dependent on a symbolic manipulation system like MACSYMA or REDUCE [38] for its successful implementation. Even though the example we provide is somewhat contrived, the amount of algebra required is substantial.

We will study the zeros of polynomials in the neighborhood of the real cubic

$$\hat{p}(\tau) = \tau^3 - (1+\delta)\tau^2 - (1+\delta)\tau + (1-\delta) ,$$

with δ = 1E-6. Its three simple zeros are

$$\hat{\beta} = -.99999975 ,$$
$$\hat{\alpha}_1 = .99877563 ,$$
$$\hat{\alpha}_2 = 1.00122512 .$$

The last two of these are somewhat ill conditioned. We will use the uniform norm in which all weights are 1; then the condition numbers of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are about 350; the condition number of $\hat{\beta}$ is about .43.

The results are given in Tables VII.1 and VII.2. $\hat{p}$ is the original polynomial with zeros $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\beta}$. p is the nearest polynomial with a double zero:

$$p(\tau) = (\tau-\alpha)^2(\tau-\beta)$$

where $\alpha = 1$ and $\beta = -1$. Finally $\check{\beta}$ and $\check{\alpha}$ represent zeros of an arbitrary polynomial $\check{p}$ such that $r = \check{p} - p$ with $\|r\| \leq \Delta$, or $\hat{r} = \check{p} - \hat{p}$ with $\|\hat{r}\| \leq \Delta$.

Table VII.1. Expansions to $\check{p}$

Classical Taylor series

From $\hat{\beta} = -.99999975$, a simple zero of $\hat{p}$:

$$\check{\beta} = \hat{\beta} - .25\hat{r}(\hat{\beta}) + .25\hat{r}(\hat{\beta})\{.25\hat{r}(\hat{\beta}) + .25\hat{r}'(\hat{\beta})\} + O(\hat{r}^3) .$$

From $\beta = -1$, a simple zero of p:

$$\check{\beta} = \beta - \tfrac{1}{4}r(\beta) + \tfrac{1}{4}r(\beta)\{\tfrac{1}{4}r(\beta) + \tfrac{1}{4}r'(\beta)\} + O(r^3) .$$

From $\hat{\alpha}_1 = .99877563$ or $\hat{\alpha}_2 = 1.00122512$, simple zeros of $\hat{p}$:

$$\check{\alpha}_1 = \hat{\alpha}_1 + 204\hat{r}(\hat{\alpha}_1) + 204\hat{r}(\hat{\alpha}_1)\{83282\hat{r}(\hat{\alpha}_1) + 204\hat{r}'(\hat{\alpha}_1)\} + O(\hat{r}^3)$$
$$\check{\alpha}_2 = \hat{\alpha}_2 - 204\hat{r}(\hat{\alpha}_2) - 204\hat{r}(\hat{\alpha}_2)\{83386\hat{r}(\hat{\alpha}_2) - 204\hat{r}'(\hat{\alpha}_2)\} + O(\hat{r}^3)$$

Classical Puiseux fractional power series

From $\alpha = 1$, a double zero of p:

$$\check{\alpha} = \alpha + \sqrt{-\tfrac{1}{2}r(\alpha)} + \tfrac{1}{4}\{\tfrac{1}{2}r(\alpha) - r'(\alpha)\}$$
$$+ \tfrac{1}{4}\sqrt{-\tfrac{1}{2}r(\alpha)}\{-\tfrac{1}{4}r(\alpha) + \tfrac{1}{2}r'(\alpha) - \tfrac{1}{2}r''(\alpha) - \frac{(\frac{1}{2}r(\alpha) - r'(\alpha))^2}{4r(\alpha)}\}$$
$$+ O(r^2)$$

"Expansions" based on the new technique

From $\beta = -1$, a simple zero of p:

$$\tilde{\beta} = \beta - \delta q$$
$$\check{\beta} = \tilde{\beta} + x_{\tilde{\beta}}(\tilde{\beta}) + x_{\tilde{\beta}}(\tilde{\beta})x'_{\tilde{\beta}}(\tilde{\beta}) + O(x^3_{\tilde{\beta}})$$

From $\alpha = 1$, a double zero of p:

$$\tilde{\alpha} = \alpha + \delta\alpha$$

$$\check{\alpha} = \tilde{\alpha} + \sqrt{x_{\tilde{\alpha}}(\tilde{\alpha})} + \frac{1}{2}x'_{\tilde{\alpha}}(\tilde{\alpha}) + \frac{1}{4}\sqrt{x_{\tilde{\alpha}}(\tilde{\alpha})}\{x''_{\tilde{\alpha}}(\tilde{\alpha}) + \frac{(x'_{\tilde{\alpha}}(\tilde{\alpha}))^2}{2x_{\tilde{\alpha}}(\tilde{\alpha})}\} + O(x_{\tilde{\alpha}}{}^2)$$

Table VII.2. Bounds on Zeros

Crude bounds based on classical expansions

$$|\check{\beta}-\hat{\beta}| \leq .43\Delta + .43\Delta^2 + O(\Delta^3)$$
$$|\check{\beta}-\beta| \leq .43\Delta + .43\Delta^2 + O(\Delta^3)$$
$$|\check{\alpha}_2-\hat{\alpha}_2| \leq 353\Delta + 5.1E7\Delta^2 + O(\Delta^3)$$
$$|\check{\alpha}-\alpha| \leq .93\Delta^{1/2} + .78\Delta + \left(.60 + \frac{.43}{\sqrt{|r(\alpha)|/\Delta}}\right)\Delta^{3/2} + O(\Delta^2)$$

Crude bounds based on the new technique

$$|\check{\beta}-\beta| \leq .84\Delta + .38\Delta^2 + O(\Delta^3)$$
$$|\check{\alpha}-\alpha| \leq .93\Delta^{1/2} + 1.00\Delta + .66\Delta^{3/2} + O(\Delta^2)$$

Best possible bounds based on classical expansions

$$|\check{\beta}-\hat{\beta}| \leq .43\Delta + .078\Delta^2 + O(\Delta^3)$$
$$|\check{\beta}-\beta| \leq .43\Delta + .078\Delta^2 + O(\Delta^3)$$
$$|\check{\alpha}_2-\hat{\alpha}_2| \leq 353\Delta + 5.1E7\Delta^2 + O(\Delta^3)$$
$$|\check{\alpha}-\alpha| \leq .93\Delta^{1/2} + .42\Delta + \left(.13 + \frac{.22}{\sqrt{|r(\alpha)|/\Delta}}\right)\Delta^{3/2} + O(\Delta^2)$$

Best possible bounds based on the new technique

$$|\check{\beta}-\beta| \leq .43\Delta + O(\Delta^3)$$
$$|\check{\alpha}-\alpha| \leq .93\Delta^{1/2} + .42\Delta + .0084\Delta^{3/2} + O(\Delta^2)$$

Table VII.1 represents expansions to $\check{p}$ from $\hat{p}$ and p. There is little difference in the expansions for $\check{\beta}$, but the difference for $\check{\alpha}$ is remarkable. Starting from the ill conditioned zeros $\hat{\alpha}$, the Taylor series terms have huge coefficients reflecting short radii of convergence. In contrast, the fractional power series expansion from the double zero at α has modest coefficients but exhibits a different kind of shortcoming: in certain directions the fractional power series does not exist at all, namely those directions, tangent to the manifold, such that $r(\alpha) = 0$. Then the coefficient of the third term becomes infinite because its denominator contains $(r(\alpha))^{1/2}$. As we have seen, in this direction the proper series expansions consist of a trivial one $\check{\alpha} = \alpha$ and a Taylor series in integral powers of r. It is easy enough to bound changes in that special direction; the severe problem is that when $r(\alpha)$ is not zero but is small compared to $\|r\|$, the terms in which $r(\alpha)^{-1}$ appears have huge coefficients.

"Expansions" are also given in the form produced by the new technique. These expansions are not useful until converted into bounds, since they are not in terms of a perturbation r but rather depend on the unknowns $\tilde{\alpha}$ or $\tilde{\beta}$, and on x, which is defined below in terms of an orthogonal perturbation.

Table VII.2 shows bounds for the changes in the zeros based on the expansions. The table gives both "crude" bounds, which reflect the simplest approximations that come to mind, and "best possible" bounds which reflect a finer analysis. An automatic symbol manipulator might produce rather crude bounds while the best possible bounds would likely be produced by a human analyst.

The bounds for $\check{\beta}$ are not of much interest. The bounds for $\check{\alpha}$ reflect the same difficulties as the Taylor or Puiseux series from which they were derived. The interesting part of Table VII.2 shows bounds for small Δ based on the revised expansion techniques discussed in the previous section. The important improvement is that the bound for $|\check{\alpha}-\alpha|$ is now independent of the direction of r and all the coefficients are of modest size. Furthermore the first two terms are the same as the best classical bound. The new technique may be used for bounding until Δ becomes comparable to $|\alpha-\beta|$.

Thus this example vindicates the approach advocated in the previous section. The rest of the current section provides the details of computing Tables VII.1 and VII.2. Those details provide convincing evidence that practical exploitation of the new expansion technique requires a sophisticated symbol manipulation system.

The bounds computed by Smith's method [42] are somewhat larger than those in Table VII.2. In particular, that method indicates

$$|\check{\alpha}-\alpha| \leq 1.32\Delta^{1/2} + O(\Delta) .$$

Details of Expansions

We first construct the expansion from $\hat{p}$. If we consider a perturbation $\varepsilon\hat{r}(\tau)$ to $(\tau-\hat{\alpha}_i)\hat{q}_i(\tau)$ we find, according to (2.5), that the perturbed zero

$$\check{\alpha}_i = \hat{\alpha}_i + x(\hat{\alpha}_i)\varepsilon + x(\hat{\alpha}_i)x'(\hat{\alpha}_i)\varepsilon^2 + \cdots$$

where

$$x(\tau) \equiv -\hat{r}(\tau)/\hat{q}_i(\tau) .$$

Thus if $i = 1$ then $\hat{q}_1(\tau) = (\tau-\hat{\alpha}_2)(\tau-\hat{\alpha}_3)$; $\hat{\alpha}_3 \equiv \hat{\beta}$. Also

$$x(\hat{\alpha}_1) = -\hat{r}(\hat{\alpha}_1)/(\hat{\alpha}_1-\hat{\alpha}_2)(\hat{\alpha}_1-\hat{\alpha}_3) ,$$

$$x'(\hat{\alpha}_1) = \frac{\hat{r}(\hat{\alpha}_1)(2\hat{\alpha}_1-\hat{\alpha}_2-\hat{\alpha}_3)}{((\hat{\alpha}_1-\hat{\alpha}_2)(\hat{\alpha}_1-\hat{\alpha}_3))^2} - \frac{\hat{r}'(\hat{\alpha}_1)}{(\hat{\alpha}_1-\hat{\alpha}_2)(\hat{\alpha}_1-\hat{\alpha}_3)} .$$

We may represent the polynomial $\hat{r}$ by the value of $\hat{r}$ and its derivatives at $\hat{\alpha}_1$ or by its coefficients. Using coefficients,

$$\hat{r}(\tau) = \hat{r}_1\tau^2 + \hat{r}_2\tau + \hat{r}_3 ,$$
$$\hat{r}(\hat{\alpha}_1) = \hat{r}_1\hat{\alpha}_1^2 + \hat{r}_2\hat{\alpha}_1 + \hat{r}_3 ,$$
$$\hat{r}'(\hat{\alpha}_1) = 2\hat{r}_1\hat{\alpha}_1 + \hat{r}_2 .$$

Finally let $\varepsilon \to 1$ to obtain a Taylor series in the coefficients of $\hat{r}$. Notice that in the first order term those coefficients appear linearly, in the second order term they appear quadratically, etc. Substituting numerical values yields

$$\check{\beta} = \hat{\beta} - .25\hat{r}(\hat{\beta}) + .25\hat{r}(\hat{\beta})\{.25\hat{r}(\hat{\beta}) + .25\hat{r}'(\hat{\beta})\} + O(\hat{r}^3) ,$$
$$\check{\alpha}_1 = \hat{\alpha}_1 + 204\hat{r}(\hat{\alpha}_1) + 204\hat{r}(\hat{\alpha}_1)\{83282\hat{r}(\hat{\alpha}_1) + 204\hat{r}'(\hat{\alpha}_1)\} + O(\hat{r}^3) ,$$
$$\check{\alpha}_2 = \hat{\alpha}_2 - 204\hat{r}(\hat{\alpha}_2) - 204\hat{r}(\hat{\alpha}_2)\{83386\hat{r}(\hat{\alpha}_2) - 204\hat{r}'(\hat{\alpha}_2)\} + O(\hat{r}^3) .$$

The expansions for $\check{\alpha}_1$ and $\check{\alpha}_2$ look unlikely to converge for other than small $\hat{r}$; in fact there is a polynomial p with a double zero at distance $\|\hat{r}\| \doteq 1.7E\text{-}6$.

We now consider expansions from

$$p(\tau) = (\tau-\alpha)^2(\tau-\beta)$$

with $\alpha = 1$ and $\beta = -1$. We will compute the effect of a perturbation $r(\tau) \equiv \check{p}(\tau) - p(\tau)$ on α and β. For β, following (2.5),

define

$$x(\tau) = -r(\tau)/(\tau-\alpha)^2$$

so

$$x'(\tau) = \frac{2r(\tau)}{(\tau-\alpha)^3} - \frac{r'(\tau)}{(\tau-\alpha)^2} .$$

Then

$$x(\beta) = -\frac{1}{4}r(\beta) ,$$

and

$$x'(\beta) = -\frac{1}{4}r(\beta) - \frac{1}{4}r'(\beta) ,$$

so

$$\check{\beta} = \beta - \frac{1}{4}r(\beta) + \frac{1}{16}r(\beta)\{r(\beta) + r'(\beta)\} + O(r^3) .$$

Following (2.6) in corresponding fashion for α, define

$$x(\tau) = -r(\tau)/(\tau-\beta)$$

so

$$x'(\tau) = \frac{r(\tau)}{(\tau-\beta)^2} - \frac{r'(\tau)}{(\tau-\beta)}$$

and

$$x''(\tau) = -\frac{2r(\tau)}{(\tau-\beta)^3} + \frac{2r'(\tau)}{(\tau-\beta)^2} - \frac{r''(\tau)}{(\tau-\beta)} .$$

Then

$$x(\alpha) = -\frac{1}{2}r(\alpha) ,$$
$$x'(\alpha) = \frac{1}{4}r(\alpha) - \frac{1}{2}r'(\alpha) ,$$
$$x''(\alpha) = -\frac{1}{4}r(\alpha) + \frac{1}{2}r'(\alpha) - \frac{1}{2}r''(\alpha) .$$

Finally

$$\check{\alpha} = \alpha + \sqrt{-\frac{1}{2}r(\alpha)} + \frac{1}{8}r(\alpha) - \frac{1}{4}r'(\alpha) + \frac{1}{4}\sqrt{-\frac{1}{2}r(\alpha)}\; x''(\alpha)$$
$$+ \frac{(\frac{1}{2}r(\alpha) - r'(\alpha))^2}{8\sqrt{-\frac{1}{2}r(\alpha)}} + O(r^2) .$$

Bounds from Expansion

The changes in zeros may be crudely bounded in a straightforward way:

$$|\check{\beta}-\beta| \leq .25|r(\beta)| + \frac{1}{16}|r(\beta)|\{|r(\beta)| + |r'(\beta)|\} + O(r^3) .$$

But $|r(\beta)| \leq \|(\beta^2\ \beta\ 1)\| \cdot \|r\| \leq \sqrt{3}\,\Delta$. Similarly $|r'(\beta)| \leq \|(2\beta\ 1\ 0)\|\|r\| \leq \sqrt{5}\,\Delta$. So

$$|\check{\beta}-\beta| = .433\Delta + .430\Delta^2 + O(\Delta^3) .$$

The bounds for $|\check{\beta}-\hat{\beta}|$ and $|\check{\alpha}-\hat{\alpha}|$ are similarly derived. As we have seen, bounds for $|\check{\alpha}-\alpha|$ independent of r do not exist.

We can improve on these bounds by taking a little care. For instance, the second term in the expansion for $\check{\beta}-\beta$ is

$$r(\beta)\{r(\beta)+r'(\beta)\}/16 .$$

Writing $r(\tau) = r_1\tau^2 + r_2\tau + r_3$ we find that term becomes

$$(r_1-r_2+r_3)(-r_1+r_3)/16 .$$

Then the question is: how large can

$$|(r_1-r_2+r_3)(-r_1+r_3)|/16$$

be, subject to the constraint $\|r\|^2 = |r_1|^2 + |r_2|^2 + |r_3|^2 = \Delta^2$? This problem in non-linear optimization can be solved, for instance with a Lagrange multiplier, to find that the desired maximum is $.0776\Delta^2$.

Similarly the second term in the expansion for $\check{\alpha}-\alpha$ is

$$\frac{1}{4}(\frac{1}{2}r(\alpha) - r'(\alpha)) .$$

While we could bound the term as

$$\frac{1}{4}(\frac{1}{2}|r(\alpha)| + |r'(\alpha)|) \leq \sqrt{3}\Delta/8 + \sqrt{5}\Delta/4 = .776\Delta \quad ,$$

we do better to observe that $r(\alpha) = r_1 + r_2 + r_3$ so we wish to maximize

$$\frac{1}{4}|-1.5r_1 - \frac{1}{2}r_2 + \frac{1}{2}r_3| \leq \frac{1}{4}\|(-1.5 \quad -\frac{1}{2} \quad \frac{1}{2})\|\left\|\begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}\right\|$$

$$\leq (\sqrt{11}/8)\Delta = .415\Delta \ .$$

Bounds from the New Technique

Now consider how the zeros change when subject to perturbations of the form discussed in section 6. First, p is perturbed to

$$\tilde{p}(\tau) = (\tau-\tilde{\alpha})^2(\tau-\tilde{\beta})$$

by movement along the manifold. Then, an orthogonal perturbation

$$\delta\lambda\tilde{e} = \delta\lambda((\tilde{\alpha}^*)^{n-1} \ (\tilde{\alpha}^*)^{n-2} \ \cdots \ \tilde{\alpha}^* \ 1)$$

is applied. The total perturbation should be commensurate with Δ which to simplify matters will be taken to be no larger than 10^{-4}.

Corresponding to the bound $\|r\| \leq \Delta$ for the conventional expansion we have (6.1):

$$\left\|\begin{pmatrix} \delta\lambda \\ \delta q \\ \delta\alpha \end{pmatrix}\right\|_H^2 = \begin{pmatrix} \delta\lambda \\ \delta q \\ \delta\alpha \end{pmatrix}^* H \begin{pmatrix} \delta\lambda \\ \delta q \\ \delta\alpha \end{pmatrix} \leq \Delta^2 \ .$$

To compute the components of H, note

$$AW^{-1}A^* = e_\alpha^* e_\alpha = 3 ,$$

$$P_{m-1} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} ,$$

$$X = P_{m-1}^* W P_{m-1} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} ,$$

$$P_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} ,$$

$$q = \begin{pmatrix} 1 \\ 1 \end{pmatrix} ,$$

$$Xq = \begin{pmatrix} 1 \\ 1 \end{pmatrix} ,$$

$$P_1 Xq = 0 ,$$

$$q^* Xq = 2 ,$$

$$P_1^* X P_1 = 6 .$$

So

$$H = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 8 \end{pmatrix} .$$

We will compute the best possible bounds from H, but for the crude bounds we will use (6.2). Then $v = 0$ so (6.2) becomes

$$|\delta\lambda| \leq (\sqrt{3}/3)\Delta ,$$

$$|\delta q| = |\delta\beta| \leq (\sqrt{6}/6)\Delta ,$$

and

$$|\delta\alpha| \leq (\sqrt{2}/4)\Delta .$$

In the usual case when $\deg q > 1$, $\delta\beta$ is a Taylor series in εq.

The variation in the double zero $\tilde{\alpha}$ and the simple zero $\tilde{\beta}$ is thus easily bounded for movements along the manifold. Now we turn to

the effect of the orthogonal movement in the direction $\delta\lambda\tilde{e}$. The effect on $\tilde{\beta}$ may again be deduced from (2.5); let

$$x(\tau) = -\delta\lambda \frac{\sum(\tilde{\alpha}^*\tau)^{n-j}}{(\tau-\tilde{\alpha})^2}$$

so

$$x'(\tau) = \delta\lambda\{\frac{2\sum(\tilde{\alpha}^*\tau)^{n-j}}{(\tau-\tilde{\alpha})^3} - \frac{\tilde{\alpha}^*\sum(n-j)(\tilde{\alpha}^*\tau)^{n-j-1}}{(\tau-\tilde{\alpha})^2}\} .$$

Then

$$x(\tilde{\beta}) = -\delta\lambda(\sum(\tilde{\alpha}^*\tilde{\beta})^{n-j}/(\tilde{\beta}-\tilde{\alpha})^2)$$

and

$$x'(\tilde{\beta}) = \delta\lambda\{\frac{2\sum(\tilde{\alpha}^*\tilde{\beta})^{n-j}}{(\tilde{\beta}-\tilde{\alpha})^3} - \frac{\tilde{\alpha}^*\sum(n-j)(\tilde{\alpha}^*\tilde{\beta})^{n-j-1}}{(\tilde{\beta}-\tilde{\alpha})^2}\} .$$

Since $|\alpha-\tilde{\alpha}| \leq \Delta$ and $|\beta-\tilde{\beta}| \leq \Delta$ and $\Delta \leq 10^{-4}$, in the bounds that follow no harm is done by substituting α for $\tilde{\alpha}$ and β for $\tilde{\beta}$, since the resulting coefficients will only be given to 3 figures. For larger Δ more care must be taken. In particular, if the perturbation along the manifold is extended far enough to reach the next higher manifold, where $\tilde{\alpha} = \tilde{\beta}$, the bounds below will be utterly wrong.

To get a crude bound, we would use

$$|x(\tilde{\beta})| \leq |\delta\lambda|(\sum|\tilde{\alpha}^*\tilde{\beta}|^{n-j}/|\tilde{\beta}-\tilde{\alpha}|^2) = \frac{3}{4}|\delta\lambda| \leq \frac{\sqrt{3}}{4}\Delta ,$$

$$|x'(\tilde{\beta})| \leq |\delta\lambda|\cdot\frac{3}{2} \leq \frac{\sqrt{3}}{2}\Delta .$$

Then

$$|\check{\beta}-\tilde{\beta}| \leq |x(\tilde{\beta})| + |x(\tilde{\beta})||x'(\tilde{\beta})| + \cdots$$
$$\leq (\sqrt{3}/4)\Delta + 3/8\,\Delta^2 + \cdots .$$

Since

$$|\tilde{\beta}-\beta| = |\delta\beta| \leq (\sqrt{6}/6)\Delta ,$$

we get

$$|\check{\beta}-\beta| \le (\frac{\sqrt{3}}{4}+\frac{\sqrt{6}}{6})\Delta + \frac{3}{8}\Delta^2 + \cdots$$
$$\le .841\Delta + .375\Delta^2 + O(\Delta^3) .$$

For a more refined bound, just be more careful:

$$x(\tilde{\beta}) = -\frac{1}{4}\delta\lambda \ ,$$
$$x'(\tilde{\beta}) = 0 \quad (+ \text{ higher order terms}) .$$

$x'(\tilde{\beta})$ is exactly 0 when $\tilde{\alpha} = \alpha$ and $\tilde{\beta} = \beta$, and has higher order terms otherwise. $x(\tilde{\beta})$ also has second order terms which we have not bothered to extract.

$$\check{\beta} - \beta = \tilde{\beta} - \beta + \check{\beta} - \tilde{\beta} = -\delta q + x(\tilde{\beta}) + x(\tilde{\beta})x'(\tilde{\beta}) + \cdots$$
$$= -\delta q - \frac{1}{4}\delta\lambda + 0 + \cdots .$$

A best possible bound for $|-\delta q - \frac{3}{4}\delta\lambda|$ may be obtained from the condition $\|\delta h\|_H \le \Delta$:

$$\left|(-\tfrac{1}{4}\ -1\ 0)\begin{pmatrix}\delta\lambda\\ \delta q\\ \delta\alpha\end{pmatrix}\right| \le \|(-\tfrac{1}{4}\ -1\ 0\ \|_H \left\|\begin{pmatrix}\delta\lambda\\ \delta q\\ \delta\alpha\end{pmatrix}\right\|_H$$
$$= \sqrt{(-\tfrac{1}{4}\ -1\ 0)H^{-1}\begin{pmatrix}-\frac{1}{4}\\ -1\\ 0\end{pmatrix}}\ \|\delta h\|_H \le \frac{1}{4}\sqrt{3}\,\Delta .$$

The corresponding computation for $\check{\alpha}$ is slightly more complicated:

$$x(\tau) = -\delta\lambda\left(\textstyle\sum(\tilde{\alpha}^*\tau)^{n-j}\right)/(\tau-\tilde{\beta}) ,$$

$$x'(\tau) = \delta\lambda\left\{\frac{\sum(\tilde{\alpha}^*\tau)^{n-j}}{(\tau-\tilde{\beta})^2} - \frac{\tilde{\alpha}^*\sum(n-j)(\tilde{\alpha}^*\tau)^{n-j-1}}{(\tau-\tilde{\beta})}\right\} ,$$

$$x''(\tau) = \delta\lambda\left\{\frac{(\tilde{\alpha}^*)^2\sum(n-j)(n-j-1)(\tilde{\alpha}^*\tau)^{n-j-2}}{(\tau-\tilde{\beta})} - \frac{2\tilde{\alpha}^*\sum(n-j)(\tilde{\alpha}^*\tau)^{n-j-1}}{(\tau-\tilde{\beta})^2} + \frac{2\sum(\tilde{\alpha}^*\tau)^{n-j}}{(\tau-\tilde{\beta})^3}\right\} .$$

Substituting we find

$$x(\tilde{\alpha}) = -\delta\lambda\left(\sum|\tilde{\alpha}^{n-j}|^2/(\tilde{\alpha}-\tilde{\beta})\right) \quad ,$$

$$x'(\tilde{\alpha}) = \delta\lambda\left\{\frac{\sum|\tilde{\alpha}^{n-j}|^2}{(\tilde{\alpha}-\tilde{\beta})^2} - \frac{\tilde{\alpha}^*\sum(n-j)|\tilde{\alpha}^{n-j-1}|^2}{(\tilde{\alpha}-\tilde{\beta})}\right\} \quad ,$$

$$x''(\tilde{\alpha}) = \delta\lambda\left\{\frac{(\tilde{\alpha}^*)^2\sum(n-j)(n-j-1)|\tilde{\alpha}^{n-j-2}|^2}{(\tilde{\alpha}-\tilde{\beta})} - \frac{2\tilde{\alpha}^*\sum(n-j)|\tilde{\alpha}^{n-j-1}|^2}{(\tilde{\alpha}-\tilde{\beta})^2}\right.$$

$$\left. + \frac{2\sum|\tilde{\alpha}^{n-j}|^2}{(\tilde{\alpha}-\tilde{\beta})^3}\right\} .$$

Then to get a crude bound,

$$|x(\tilde{\alpha})| = \frac{3}{2}|\delta\lambda| \leq \left(\frac{\sqrt{3}}{2}\right)\Delta \ ,$$

$$|x'(\tilde{\alpha})| \leq |\delta\lambda|\left(\frac{3}{4}+\frac{3}{2}\right) \leq \frac{3}{4}\sqrt{3}\,\Delta \ ,$$

$$|x''(\tilde{\alpha})| \leq |\delta\lambda|\left(1+\frac{3}{2}+\frac{3}{4}\right) \leq \left(\frac{13\sqrt{3}}{12}\right)\Delta \ .$$

Since

$$\check{\alpha} - \alpha = \delta\alpha + \sqrt{x(\tilde{\alpha})} + \frac{1}{2}x'(\tilde{\alpha}) + \frac{1}{4}\sqrt{x(\tilde{\alpha})}\left\{x''(\tilde{\alpha}) + \frac{(x'(\tilde{\alpha}))^2}{2x(\tilde{\alpha})}\right\} + \cdots$$

then

$$|\check{\alpha}-\alpha| \leq \sqrt{\frac{\sqrt{3}}{2}}\Delta^{1/2} + \left(\frac{3}{8}\sqrt{3}+\frac{\sqrt{2}}{4}\right)\Delta + \frac{1}{4}\sqrt{\frac{\sqrt{3}}{2}}\left(\frac{13\sqrt{3}}{12}+\frac{9\sqrt{3}}{16}\right)\Delta^{3/2} + \cdots$$

or

$$|\check{\alpha}-\alpha| \leq .931\Delta^{1/2} + 1.003\Delta + .663\Delta^{3/2} + O(\Delta^2) \ .$$

To get the corresponding best possible bounds, note that

$$x(\tilde{\alpha}) = -\frac{3}{2}\delta\lambda \ ,$$

$$x'(\tilde{\alpha}) = -\frac{3}{4}\delta\lambda \ ,$$

$$x''(\tilde{\alpha}) = \frac{1}{4}\delta\lambda \quad .$$

Then for the second term $\delta\alpha + \frac{1}{2}x'(\tilde{\alpha})$ we have

$$|\delta\alpha - \frac{3}{8}\delta\lambda| \leq \|(-\frac{3}{8}\ 0\ 1)\|_H \Delta = \frac{\sqrt{11}}{8}\Delta \ .$$

For the third term,

$$\left|\frac{\sqrt{x(\tilde{\alpha})}}{4}(x''(\tilde{\alpha}) + \frac{(x'(\tilde{\alpha}))^2}{2x(\tilde{\alpha})})\right| = \frac{1}{4}\sqrt{\frac{3}{2}}\,|\frac{1}{4} - \frac{3}{16}|\,|\delta\lambda|^{3/2} \leq .0084\Delta^{3/2} \ .$$

CHAPTER VIII

EXPERIMENTAL METHODS

1. Introduction

In the next chapter experimental results will be given which vindicate the theory of previous chapters. After that we will present experimental results for a class of polynomials more difficult to understand.

In the present chapter we describe how the nearest polynomials with given multiplicity configurations were found. Then we explain the tests made to assure the validity of the results. Finally we show how to contrive test problems with known answers.

Experiments were carried out on the CDC 6400 at the University of California, Berkeley. Coding was in the FORTRAN language for the University of Washington RUN compiler. Although most of the codes usually perform satisfactorily in the stated environment they are not presently in a portable form that would work reliably in other environments. Consequently a detailed discussion and listing of these codes is not included here.

2. How the Equations were Solved

Chapters III-V presented various equations to be solved for solutions ζ corresponding to nearest polynomials with one or more multiple zeros. Expressions were usually obtained both for a function and its partial derivatives so that Newton's method could be applied. To use any iterative method, however, starting guesses must be supplied.

Usually the starting point was taken to be a zero of the appropriate derivative. Thus, if the nearest polynomial with a double zero was sought, a starting point would be chosen from among the zeros of the first derivative. One might try to use the zeros of the original polynomial, but the zeros of the derivative seemed more often to lead to faster convergence.

In order to maximize the probability of first finding the globally nearest polynomial with the desired multiplicity configuration, the starting points were tried in a definite order. That order was fixed by computing the distance to the nearest polynomial with that starting point as a double zero. That distance is an upper bound for the distance to the manifold from that starting point. The starting points with the least upper bounds were used first.

The same criterion for choosing among starting points could be used if the starting points were the zeros of the original polynomial. In this case, however, it would be equally appropriate to rank the starting points according to their condition numbers.

Once a starting point was chosen, Newton's method was used in all but one instance. That case exploited the fact that the equation for the nearest polynomial with a double zero always has a real solution

between two real zeros of a real polynomial. Those two real zeros may be used as starting points for a secant-like iteration for ζ; among many such iterations Brent's [2] is a well known recent one. Brent's method was used to quickly locate real solutions whenever appropriate.

In order to terminate the iteration an error bound on the function evaluation was computed. When the function whose zero was sought was reduced below its error bound, the current iterate was accepted as a zero. These error bounds were usually computed with the aid of interval arithmetic [24]. The lack of suitable facilities for interval arithmetic in CDC hardware and software made it necessary to code interval operations as subroutine calls -- making the codes for the functions virtually unreadable, and thereby providing another reason for not publishing those codes here.

If no solution was found after a fixed number of iterations (usually 40) the iteration was terminated and another starting point tried. If a solution was found it was added to the list of known solutions used to deflate the function, as described in one of the appendices.

When all the reasonable starting points had been tried the accumulated solutions were checked for correctness and the corresponding perturbations analyzed.

3. How Do We Know the Answers are Correct?

The methods just described produce one or more solutions ζ corresponding to locally nearest polynomials with a given multiplicity configuration. The next step is to compute each polynomial from its ζ and check that it is indeed an appropriate solution. Because no similar computations suitable for comparison have been published, extra care was necessary to be sure that the numerical results were reliable.

It must be understood from the outset that in general we can not be sure of having obtained the global minimum. With no theoretical information on the size of the second derivative or on the number of local minima that may exist in a region the best that can be done is to obtain as many local minima as possible and examine each. Empirically we have never found more than $n+2$ local minima while searching for the nearest polynomial with a single multiple zero, so that task is not quite hopeless. Furthermore, whenever one might reasonably expect from the nature of a problem that one minimum would clearly be much better than the rest, that minimum has always been found approximately as expected. An example of such a problem is one in which a perturbation is applied to a polynomial having one multiple zero and several simple zeros, all well conditioned in the sense of chapter II. Thus the perturbed polynomial has simple zeros near the simple zeros of the unperturbed polynomial, but the multiple zero has divided into several very ill conditioned zeros. When the computer codes are asked to find the polynomial with an appropriately multiple zero nearest that perturbed polynomial, they have so far always found a locally closest polynomial with a multiple zero near the multiple zero of the

original unperturbed polynomial. In the circumstances described, moreover, none of the other local minima are competitive in distance. Thus it seems highly likely that the best local minimum is really the global minimum.

There is the additional complication that our results are for real polynomials and, as we have seen in chapters III and V, it is sometimes necessary to solve an extra set of equations for higher multiplicity in order to find the global minimum. In our experience with double zeros, only once has a better minimum been found by solving the equation for a triple zero. Thus our overall results are probably not seriously compromised by failing to check for quadruple zeros when searching for triples, or for various higher configurations when searching for two or more doubles.

The reader may wonder why it is so easy to find the ζ's when the starting points are near ill conditioned zeros of a polynomial. After all, ill conditioned zeros themselves are almost by definition difficult to find.

The explanation lies in the form in which polynomials are presented to our codes, namely as a list of their zeros. If the polynomials were represented by their coefficients, as they are represented to a subroutine to find zeros of polynomials, then the solutions ζ to the equations we wish to solve would also be ill conditioned functions of the input data. But since ill conditioned zeros are normally recognizable as a problem requiring amelioration only when those zeros are in hand, the sensible form for representing that ill conditioned polynomial is by its zeros rather than its coefficients. In that form

the polynomial may always be evaluated with low relative error, even near its zeros.

4. Computed Checks on Results

Once a ζ has been found, we can compute the perturbing polynomial $q(\tau)$ by an equation such as (III.6.4). Then $p(\tau)+q(\tau)$ should be locally nearest to $p(\tau)$ and should have a multiple zero ζ of the intended multiplicity m, or several ζ's of appropriate multiplicities if that was what was requested.

Analytical errors, approximation errors, coding errors, and rounding errors could all cause the results to be other than expected, so each assertion about $p+q$ is checked in the codes.

Note that $p+q$ is never represented by computing the coefficients of $p+q$. Since the coefficients of q are usually intended to be small perturbations of the coefficients of p, adding them together would entail severe loss of significance. Therefore to evaluate $(p+q)(\eta)$ at a specific η, compute

$$p(\eta) = \prod_{i=1}^{n} (\eta-\alpha_i)$$

and

$$q(\eta) = \sum_{i=1}^{n} q_i \eta^{n-i}$$

and then add $p(\eta)$ and $q(\eta)$.

Using this evaluation scheme our first task is to check the assertion that ζ is an m-tuple zero of $p+q$, i.e.

$$p^{(k)}(\zeta)+q^{(k)}(\zeta) = 0 , \quad k = 0,\ldots,m-1 .$$

We do not expect that equation to be satisfied exactly on a finite precision computer so we compute error bounds by interval arithmetic

and ask only that

$$|p^{(k)}(\zeta) + q^{(k)}(\zeta)|$$

be within its error bound. That proves that $p+q$ satisfies the constraint of lying on the manifold of polynomials with m-tuple zeros.

The next assertion to be checked is that $p+q$ represents a stationary point on the manifold with respect to $\|q\|$. The analysis of chapter III shows that this is the case if either the last Lagrange multiplier vanishes or the multiplicity of ζ in $p+q$ is at least one greater than requested. For our codes the last Lagrange multiplier is usually forced to be zero in the solution process for ζ and q. If we wish to examine other stationary points which, as we have shown, can not be minimal with respect to complex perturbations, we check that one of the stationarity conditions is satisfied.

After checking stationarity we turn to minimality of $\|q\|$. Minimality may be checked by examining the Hessian matrix of second derivatives of $\|q\|^2$. Given any <u>fixed</u> ζ, there is a unique q closest to p such that $p+q$ has an m-tuple zero ζ. Thus $\|q\|$ could be regarded as a real function of two real variables, Re ζ and Im ζ, for which partial derivatives can be computed to provide a 2 by 2 Hessian matrix. Alternatively the method of section III.10 could be used to compute a Hessian matrix for the coefficients of q and the ζ's which are now regarded as independent except for constraints. To simplify computation only real changes in q and ζ were considered in computing the constrained Hessian of dimension $n+1-m$.

Using either Hessian, minimality could be checked by computing the signature. Actually the complete set of eigenvalues was computed

to ascertain the shape of the minimum. Minimality corresponds to all eigenvalues positive; maximality to all negative; other configurations correspond to saddle points.

After the checks listed above, the other zeros of $p(\tau) + q(\tau)$ were computed, assuming that the m-tuple zero ζ was known. Then the n zeros were used to reconstitute the coefficients of a polynomial whose coefficients should be close to those of $p(\tau) + q(\tau)$. The explicit coefficients of $p+q$ were computed for use in this check only. The maximum relative difference was noted and flagged if larger than roundoff error level. If no flag was noted then the zeros of $p+q$ were assumed to be reliably computed and their condition numbers were calculated. Of special interest was the condition number of the multiple zero ζ which should have been much smaller than the condition numbers of the ill conditioned zeros it replaced.

When computing q and $\|q\|$ in cases where we expect the last Lagrange multiplier to be zero, we usually forced it to be zero while solving the linear equations for q. We could, however, solve a system of linear equations of dimension one larger. Then, because of rounding error, we expect the last Lagrange multiplier to be small but not zero. So as a check we re-computed q and $\|q\|$ using the non-zero multiplier. The two values of $\|q\|$ are compared and flagged if they differ by more than a few units in the last place of precision.

Finally a number of random small perturbations of ζ were made and the distance to the nearest polynomial with the perturbed ζ as a multiple zero was computed. Since the original ζ was alleged to be a minimal point, a message was printed if any of the nearby polynomials were significantly closer to p.

All the experimental results to be presented in this chapter and the next satisfied these checks unless otherwise stated. Thus there is a basis for confidence that the various complicated equations that were solved for one or more ζ's were in fact formulated and solved correctly.

5. Setting Up a Problem with a Known Solution

While developing computer codes it is sometimes desirable to solve a problem whose answer is known. Although it is not known how, for instance, to set up a polynomial such that the globally nearest polynomial with an m-tuple zero has the m-tuple zero we specified in advance, it is a simple matter to set up such a polynomial so that a locally nearest polynomial has that specified m-tuple zero.

One's first thought might be to start with a trivial problem whose solution is known and apply a random perturbation. This is done for some problems described in the next chapter. For instance, a small random perturbation may be applied to the coefficients of a polynomial with a double zero to obtain a nearby polynomial with two ill conditioned zeros. Then the computer codes find that the nearest polynomial with a double zero has a double zero near the one we started with. Figure VIII.1 shows why the double zero is not the same; a perturbation in a random direction is not generally "orthogonal" to the surface. The change in the multiple zero is usually commensurate with the size of the perturbation when the multiple zero is well conditioned.

It is possible to set up a perturbation so we return to a specific multiple zero, however. Recall the equation, (III.6.2), to be solved for the polynomial nearest p with an m-tuple zero ζ:

$$f(\zeta,p) = \det(\tilde{A}p \;\vdots\; AW^{-1}A^*Z) = 0 \ .$$

Z is a constant truncator matrix, W depends only on the norm, and A and $\tilde{A}$ depend only on ζ.

Normally p is given and we seek ζ by solving a highly nonlinear equation. But now we wish to find p given ζ. From the

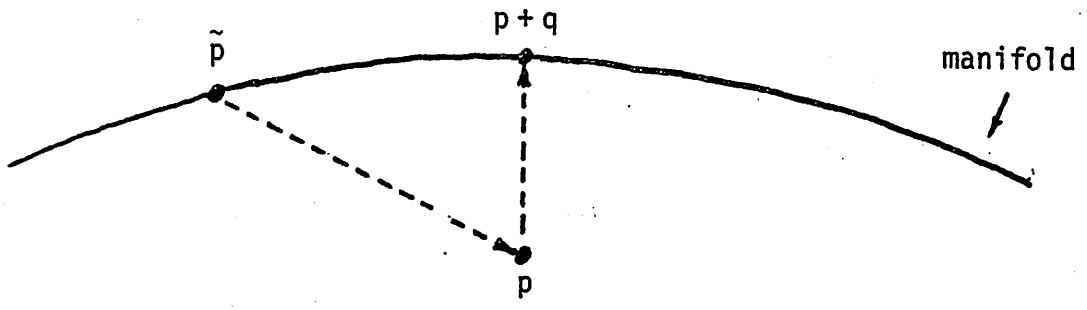

Figure VIII.1. $\tilde{p}$ has a multiple zero. A random perturbation to $\tilde{p}$ produces p. p + q is the polynomial with a multiple zero closest to p.

properties of determinants it is apparent that $f(\zeta,p)$ is a linear functional of the vector p, so $f(\zeta,p) = u_\zeta^* p$ for some u_ζ^* which depends on ζ but not p. Then to find such a p it is only necessary to obtain one of the members of the $(n-1)$ dimensional subspace of solutions of $u_\zeta^* p = 0$.

As an example, suppose we wish to start with a polynomial $\tilde{p}$ with a double zero at α, so $f(\alpha,\tilde{p}) = 0$. We then want to find a $\tilde{q}$ such that $\tilde{p}+\tilde{q}$ has a locally nearest polynomial with a double zero at α. Presumably that nearest polynomial would be $\tilde{p}$ if $\tilde{q}$ is not too large.

We find then that $f(\alpha,\tilde{q}) = 0$ is the requirement on $\tilde{q}$. We can find such a $\tilde{q}$ by letting $\tilde{q}_0$ be a polynomial with random coefficients and $\tilde{q}_1$ be the constant polynomial whose value is 1. Then

$$\tilde{q} = \tilde{q}_0 - \frac{f(\alpha,\tilde{q}_0)}{f(\alpha,\tilde{q}_1)}\,\tilde{q}_1$$

is the polynomial we seek. It may be verified that $f(\alpha,\tilde{q}_1) \neq 0$ for $m = 2$ or 3.

Then we may apply the computer codes to $\tilde{p}+\tilde{q}$ to verify that they do find a locally nearest polynomial with an m-tuple zero α.

We could impose an even more stringent requirement: that the closest polynomial to $\tilde{p}+\tilde{q}$ with a multiple zero be $\tilde{p}$ itself. This is just as easy to arrange. Recall the notation from chapter III for finding the polynomial $p+q$ with a multiple zero ζ nearest a polynomial p. For our present purpose $p = \tilde{p}+\tilde{q}$ and $p+q = \tilde{p}$ so $q = -\tilde{q}$. But

$$q = W^{-1}\hat{A}^*\hat{\ell}$$

for some $m-1$ dimensional vector $\hat{\ell}$ of Lagrange multipliers. So our recipe is: choose any random $m-1$ dimensional vector $\hat{u}$ and let

$$\tilde{p} + \tilde{q} = \tilde{p} - W^{-1}\hat{A}^*\hat{u}$$

be the perturbed polynomial. Then we may verify that the equation for ζ,

$$\det(\tilde{A}p \mid AW^{-1}A^*Z) = 0 \ ,$$

is trivially solved when $\zeta = \alpha$, for then

$$\begin{aligned} \tilde{A}p &= - AW^{-1}\hat{A}^*\hat{u} \\ &= - AW^{-1}Au \end{aligned}$$

where

$$u = \begin{pmatrix} \hat{u} \\ 0 \end{pmatrix} \ .$$

The matrix whose determinant we seek is just

$$AW^{-1}A^*(u \mid Z)$$

and the bottom row of the rightmost factor vanishes as does the determinant.

When solving for Lagrange multipliers $\hat{\ell}$,

$$\hat{A}W^{-1}\hat{A}^*\hat{\ell} = -\hat{\tilde{A}}p = \hat{A}W^{-1}\hat{A}^*\hat{u} \ ,$$

and since the rows of $\hat{A}W^{-1}\hat{A}^*$ are linearly independent, $\hat{\ell} = \hat{u}$ as we hoped, and $q = -\tilde{q}$. Thus

$$\{\tilde{q} \mid \tilde{q} = -W^{-1}\hat{A}^*\hat{u}\}$$

is indeed the subspace of perturbations of $\tilde{p}$ for which $\tilde{p}$ is a locally nearest polynomial with an m-tuple zero.

CHAPTER IX

NONPATHOLOGICAL EXPERIMENTAL RESULTS

1. Introduction

We turn now to presentation of some results of calculations performed on specific polynomials. The results in this chapter generally tend to vindicate the theory.

Calculations were usually based on the methods described in the previous chapter. The norms used were weighted least squares norms intended to minimize relative changes in the coefficients of the starting polynomial. Thus if the monic starting polynomial of degree n were

$$p(\tau) = \prod_{i=1}^{n} (\tau-\alpha_i) = \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j}$$

then polynomials

$$q(\tau) = \sum_{j=1}^{n} q_j \tau^{n-j}$$

would be sought such that $p+q$ had the desired multiplicity structure and

$$\|q\|_W^2 \equiv \sum_{j=1}^{n} w_j |q_j|^2$$

was minimized. Usually $w_j = 1/|p_j|^2$ but sometimes $w_j = 1/|\hat{p}_j|^2$ was used instead, where

$$\hat{p}(\tau) \equiv \prod_{i=1}^{n} (\tau - |\alpha_i|) \ .$$

The latter norm is applicable when some of the p_j's vanish.

The choice of norm also affects the condition numbers. Generally condition numbers for relative changes in the zeros are used.

In the first cases the "right answer" is obvious and the codes do indeed recover that answer.

2. n-tuple Zeros

Equations for finding the nearest polynomial with an n-tuple zero are given in section III.2. The present example was created by randomly perturbing a polynomial whose quintuple zero 1 has condition number .135. A perturbation of norm .749E-12 was applied in a random direction to create p whose zeros are

$$.99557908 \pm .32081885E\text{-}2\ i$$
$$1.00168511 \pm .52020041E\text{-}2\ i$$
$$1.00547160\ .$$

The condition numbers of these zeros vary from .353E+10 to .357E+10. The equations for finding the nearest polynomial with an n-tuple zero were solved by Newton's method, starting from the arithmetic mean of the five zeros of $p(\tau)$. The result was that the nearest polynomial with an n-tuple zero had the n-tuple zero 1.000000000000007 with condition number .135.

Corresponding results were obtained for similar polynomials of degrees 8 and 20. Although the n-tuple zero is easy to find, the nearest polynomial with a real double or triple zero is sometimes difficult to locate, especially if n is odd. There are usually numerous nearby polynomials with a complex double zero, and for some of these may be found a nearby real polynomial with a complex conjugate pair of double zeros.

3. Returning to a Double Zero

The next polynomial has six zeros -2, -1, 1, 1, 2, and 3. The worst conditioned of these is 3, with condition number 43.4. The double zero at 1 has condition number 5.04.

A random perturbation of norm .438E-8 was applied, creating a polynomial p:

Zero	Condition number
-2.00000000	2.89
-1.00000000	2.91
.99999998 ± .10462513E-3 i	.557E+5
2.00000011	19.5
2.99999992	43.4

The methods of chapter III were applied to find the nearest polynomial with a double zero, and a polynomial $p+q$ was soon found whose double zero at .99999998 has condition number 5.04. $\|q\|$ = .94E-9 and the other zeros were not changed by more than .0000007.

Other locally minimal polynomials with double zeros were also found. For instance the next closest one has a double zero at 2.5397 with condition number 3.85, and the worst conditioned zeros of $p+q$ are .952 ± .158i, with condition numbers 28.6. But $\|q\|$ = .385E-2, so this perturbation is over a million times larger than the previous one. By taking such a large step we manage to decrease the worst condition number only by a factor of 2, and this perturbation seems much less natural than the previous one.

Similarly when we seek the nearest polynomial with a triple zero, we find we must let $\|q\| = .017$ in order to reach the polynomial with

a triple zero at 1.20. The worst conditioned zero of that polynomial has condition number 8.64.

Thus we find that by forcing a large enough perturbation on p we can make its zeros as well conditioned as we want. However in this case we find that there is an "obvious" perturbation in which a comparatively small change in p results in a comparatively large improvement in the worst condition of p's zeros.

4. Returning to a Triple Zero

We start with the polynomial with simple zeros -2, -1, and 3, and triple zero 1. The condition of the triple zero is .797 and the worst zero is 3, with condition number 5.52.

Apply a random perturbation of norm .839E-10 to find p, a polynomial whose zeros and condition numbers are

-2.00000000	1.21
-1.00000000	.615
.99980426 ± .33876727E-3 i	.357E+7
1.00039148	.357E+7
2.99999999	5.52

When we search for nearby polynomials with double zeros, we find for instance one with a double zero .99999525 at distance .365E-10. The condition of that double zero is somewhat improved to .714E+5 but the condition of the third zero near 1 becomes .807E+10. Even though we can reach a double zero in a small step, the results are not interesting.

When we search for a nearby triple zero, however, we find that a perturbation of norm .495E-10 gets us to a polynomial with a triple zero 1.000000000014 with condition number .797. The worst zero has condition number 5.52. Comparing to the perturbation to a double zero, we find that a not much larger perturbation to a higher multiplicity structure yields a substantial improvement in condition.

Computer codes for quadruple zeros are not available but it seems doubtful that this p could be perturbed to a quadruple zero by a further small perturbation.

5. Returning to Two Double Zeros

The polynomial with simple zeros -2, 0, and 2, and double zeros -1 and +1 was perturbed by a random perturbation of norm .332E-7 to produce a polynomial whose ill conditioned zeros were ± .9999999991 ± .562407E-4 i with condition numbers .196E+4.

A polynomial at distance .143E-7 had a double zero but two remaining ill conditioned zeros. There was a polynomial with a triple zero at distance .629 with all zeros well conditioned. But the satisfactory polynomial had two double zeros at ± .999999997. All zeros were well conditioned but the perturbation q was only .219E-7 in norm.

6. Returning to a Complex Conjugate Pair of Double Zeros

Consider the eighth degree polynomial whose simple zeros are -3, -2, -1, and 4, and which also has double zeros at $2 \pm i$. The worst zero is 4, with condition number 55.0; the condition of the complex zeros is 7.98.

A random perturbation of norm .168E-8 produces a polynomial p whose zeros and condition numbers are

-3.00000000	9.98
-2.00000000	14.9
- .99999999	6.12
1.99982354 ± 1.00012355 i	.126E+6
2.00017652 ± .99987637 i	.126E+6
3.99999984	55.0

When we apply the methods of chapter IV we discover that there is a real polynomial $p+q$ with double zeros at $2.000000012 \pm .9999999946\,i$ with condition numbers 7.98. $\|q\|$ is .459E-9 and the worst zero is 3.9999993 with condition number 55.0.

Thus in the case of a complex conjugate double zero we can also find the answer when it is obvious. In this case no real double or triple zeros were found closer than .001. Of course there is no theoretical basis for asserting that they do not exist -- but if they are, they must be rather well hidden!

7. A Polynomial with Several Pairs of Complex Conjugate Zeros

Wilkinson presents a real polynomial [34, p. 63] all of whose 16 zeros are complex, most being rather ill conditioned. Condition numbers range from .878 to .107E+11.

No real ζ's were found other than 0, but 7 complex ζ's corresponding to complex perturbations were found. All of these complex ζ's lead to nearby real polynomials with complex conjugate pairs of double zeros. The closest of these is at a distance of .247E-13 and the worst conditioned zero of the perturbed polynomial has a condition number of .551E+10. So from the point of view of "explanation," clearly some higher multiplicity configuration is required. The value of this example is rather that it shows that the codes are capable of finding a number of complex conjugate pairs of double zeros when the problem is of a nature that several such solutions might reasonably be expected.

In the table below we list the unperturbed zeros α and their condition numbers on the left and, on the right, $\|q\|$, ζ, the condition of ζ, and the worst condition number of the perturbed polynomial.

Real	Imag	Cond(α)	$\|q\|$	ζ		Cond(ζ)	worst
-.305E-5	.312	.565E+10	.247E-13	-.884E-5	.312	.110E+8	.551E+10
-.148E-4	.312	.107E+11	.545E-13	-.354E-4	.311	.199E+8	.277E+10
-.471E-4	.311	.646E+10	.329E-12	-.116E-3	.309	.108E+8	.137E+10
-.143E-3	.309	.154E+10	.644E-11	-.417E-3	.306	.196E+7	.168E+9
-.491E-3	.304	.127E+9	.656E-9	-.201E-2	.295	.852E+5	.382E+7
-.232E-2	.293	.233E+7	.111E-5	-.166E-1	.260	.381E+3	.710E+4
-.187E-1	.253	.297E+4	.134E-1	-.121	.162	.483	.222E+2
-.132	.136	.878	.141E+1	0	0	.036	.227E+2

8. An Uninteresting Polynomial

In contrast to the previous examples, we consider now a polynomial all of whose zeros are well conditioned, just to see how the manifold of double zeros appears from a distance.

Let p be a cubic polynomial with zeros 1, 2, and 3, and condition numbers .87, 4.6, and 4.8. For this example we use the uniform norm for which all weights are 1. After a lengthy search we find the following interesting points:

ζ	$\|q\|$	Worst condition
Double at 2.49244540	.0551	.72
Double at 1.32286845	.152	2.7
Double at 0.0	12.53	1.0
Double at -3.20829919	12.57	.15
Double at -1.13700604	13.93	10.4
Triple at 1.87492441	57.18	.99E-2

Of these points, 0.0 turned out not to be a stationary point, and -1.13... turned out to be a maximum on the real axis, and a saddle point in the complex plane. The point 1.87... represents a minimum among perturbations to a real zero but a maximum among real perturbations to a double zero. The other three points are local minima in the complex plane.

This example supports the conclusion in chapter II that absence of ill condition implies distance from the manifolds of polynomials with multiple zeros.

9. Zeros in a Circle

The next example is a polynomial mentioned by Wilkinson [34]. Its zeros lie around the unit circle and are the twenty 20^{th} roots of unity. In the uniform norm the zeros are all very well conditioned; the real zeros have condition numbers .224 and the complex zeros have slightly smaller condition numbers, since only real perturbations are considered. Our codes were unable to find any solutions for double zeros other than zero or for complex conjugate pairs except by great labor, which produced unsatisfactory results. It turns out that

$$p(\tau) \equiv \tau^n - \beta \ ,$$

β real and positive, has non-zero solutions ζ constrained as follows for double zeros:

$$\beta/n < |\zeta|^n < (n-1)\beta \ ,$$

$$\arg \zeta = (2k+1)\pi/n \ , \quad k = 0,1,\ldots,n-1 \ .$$

Thus $\arg(\zeta^n) = \pi$ and if n is even there are no real solutions ζ.

10. Summary

The results presented in this chapter and other similar results lead to the following conclusions:

1. When there is an "obvious" nearby polynomial of a certain multiplicity structure, the computer codes find it. If insufficient multiplicity is requested, the codes find a polynomial that is close but has some zeros still very ill conditioned. When too much multiplicity is requested, the codes find a polynomial that is relatively far away although all its zeros are well conditioned. When the proper multiplicity is specified, the codes find a polynomial which is relatively close and has all zeros well conditioned.

2. When there is no obvious reason why a nearby polynomial would have substantially better conditioned zeros, the codes do not find any such polynomials.

3. The polynomials that the codes find are indeed critical points for $\|q\|$ and are usually minima. In other words, the answers are correct, but the codes may not be able to find all the answers.

With conclusions like these, based on simple cases, we have some basis for confidence in examining a more difficult polynomial in the next chapter.

CHAPTER X

WHAT'S WRONG WITH WILKINSON'S POLYNOMIAL?

1. Wilkinson's Polynomial

In [34] J. Wilkinson describes the astonishing ill condition of a polynomial whose zeros are the integers from 1 through 20. He observed that by changing one of the coefficients by less than one part in 1.0E+15 it was possible to create a polynomial some of whose zeros were complex conjugate pairs.

Our results in chapter II lead us to conclude that this badly behaved polynomial must be near the manifold of polynomials with double zeros, at least, and perhaps near manifolds corresponding to higher multiplicity configurations as well. Since this polynomial is precisely defined, we are not interested in "ameliorating" its ill condition but rather "explaining" that ill condition if possible. The results mentioned in the previous chapter show that ill condition is ideally explained by displaying a small perturbation to a nearby manifold of polynomials with some appropriate multiplicity configuration. We shall see that the experimental results presently available do not support any such simple explanation for Wilkinson's polynomial; rather they suggest that it is near a place where the manifolds of polynomials with multiple zeros are especially contorted.

After examining the well known Wilkinson polynomial we will look briefly at its translation to the origin and at another Wilkinson polynomial which is in some ways the opposite of the first.

2. Coefficients and Condition Numbers for Wilkinson's Polynomial

Two unusual things about Wilkinson's polynomial are the ranges in magnitude among the coefficients and among the condition numbers of the zeros.

The zeros are the integers from 1 through 20. Therefore the coefficients are exactly computable, but as a practical matter most have so many significant figures that they must be rounded to fit in 48 bits of a CDC computer word. Consequently the polynomial should be considered to be defined by its zeros, and the coefficients are only used to compute the weights in the norm on perturbations:

$$p(\tau) \equiv \prod_{i=1}^{n} (\tau - i) \equiv \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j}$$

$$\|q\|^2 = \sum w_j |q_j|^2$$

$$w_j = 1/|p_j|^2$$

This "relative" norm measures relative changes in the coefficients of p; we will also present results for the "uniform" norm in which all the weights are 1 and which measures absolute changes in the coefficients of p.

Some differences between these norms might be expected due to the large variation in those coefficients. In magnitude they range from 210 to 1E19; they are listed in Table X.2. Thus the corresponding weights for the relative norm range from 1E4 to 1E38.

The zeros are given in Table X.1 with their condition numbers. The first condition number is with respect to the uniform norm on the polynomial. The second condition number is with respect to the relative norm on the polynomial. All condition numbers are for absolute

Table X.1

Zeros of Wilkinson's Polynomial and Their Condition Numbers

Zero	Uniform Norm	Relative Norm
1	.368E-16	.187E+3
2	.946E-10	.355E+5
3	.173E-5	.234E+7
4	.226E-2	.778E+8
5	.620	.153E+10
6	.591E+2	.198E+11
7	.257E+4	.177E+12
8	.602E+5	.115E+13
9	.845E+6	.553E+13
10	.763E+7	.203E+14
11	.466E+8	.572E+14
12	.199E+9	.125E+15
13	.607E+9	.212E+15
14	.134E+10	.278E+15
15	.212E+10	.279E+15
16	.241E+10	.210E+15
17	.191E+10	.114E+15
18	.997E+9	.428E+14
19	.309E+9	.979E+13
20	.432E+8	.103E+13

changes in the zeros. The condition number for relative changes in α, say, may be obtained by dividing the listed condition number by $|\alpha|$.

The most striking facts about the condition numbers are

1) the magnitude of the ill condition of the worst,

2) the large group of zeros that are nearly as badly conditioned as the worst, and

3) the lack of any obvious partitioning into a set of well conditioned zeros and a set of ill conditioned ones.

The last fact distinguishes this polynomial from those of the previous chapter. There is no obviously best multiplicity configuration that we should look for. So we will try as many as we can, starting from the simplest.

Before giving the results, it is instructive to attempt to graph this polynomial. It turns out to be impossible to perceive all its features on one graph, so we present several successive magnifications of interesting parts. Figures X.1-X.4 were produced on a Tektronix 4051 Graphics System.

It is interesting to note that the symmetry of the polynomial about 10.5 is not reflected in the condition numbers, which reach their maximum near 15, depending on the norm. That is because the formula in chapter II for the condition number of a zero α has a numerator which is a monotonic increasing function of $|\alpha|$ and a denominator that depends only on the absolute spacing between α and the other zeros. The numerator is a rather rapidly increasing function of $|\alpha|$; for a simple zero, it is

$$\sum|\alpha^2|^{n-j}/w_j \ .$$

Intuitively it is hard to understand why the larger zeros should be so much more ill conditioned than the smaller ones. Indeed, by translating the entire polynomial by -10.5 so that it is symmetric about the origin, one can eliminate that part of the anomaly. Wilkinson did so and found substantial overall improvement in the condition of the zeros. Of course, if that translation were regarded as a perturbation, its norm would exceed 1 in the relative norm and 1E19 in the uniform norm, and we know that remarkable improvements in condition often accompany large perturbations.

We wish, however, to study Wilkinson's polynomial as an untranslated object. The next section gives our results. Some results for the translated polynomial are in a later section.

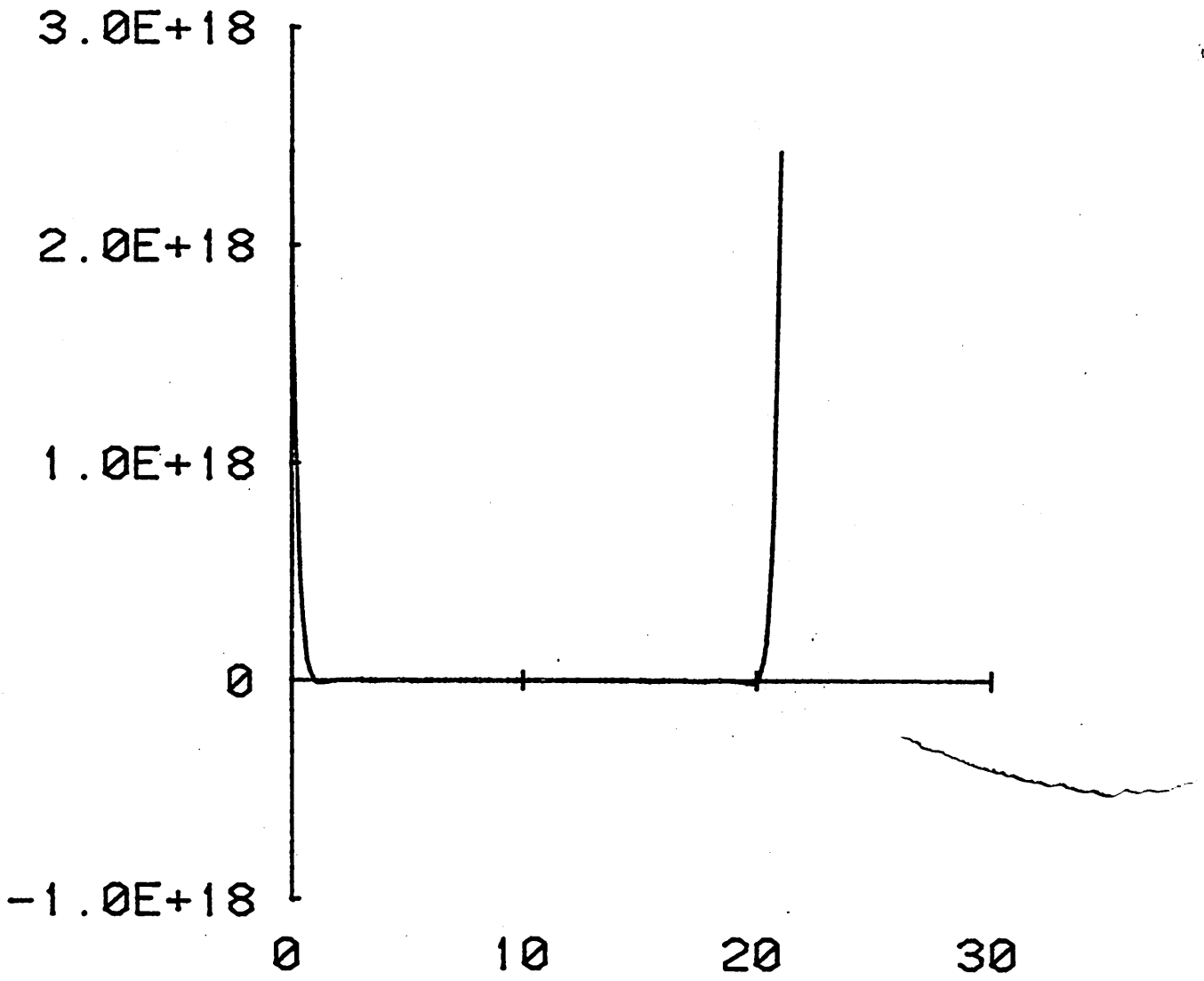

Figure X.1. Wilkinson's polynomial on [0,21].

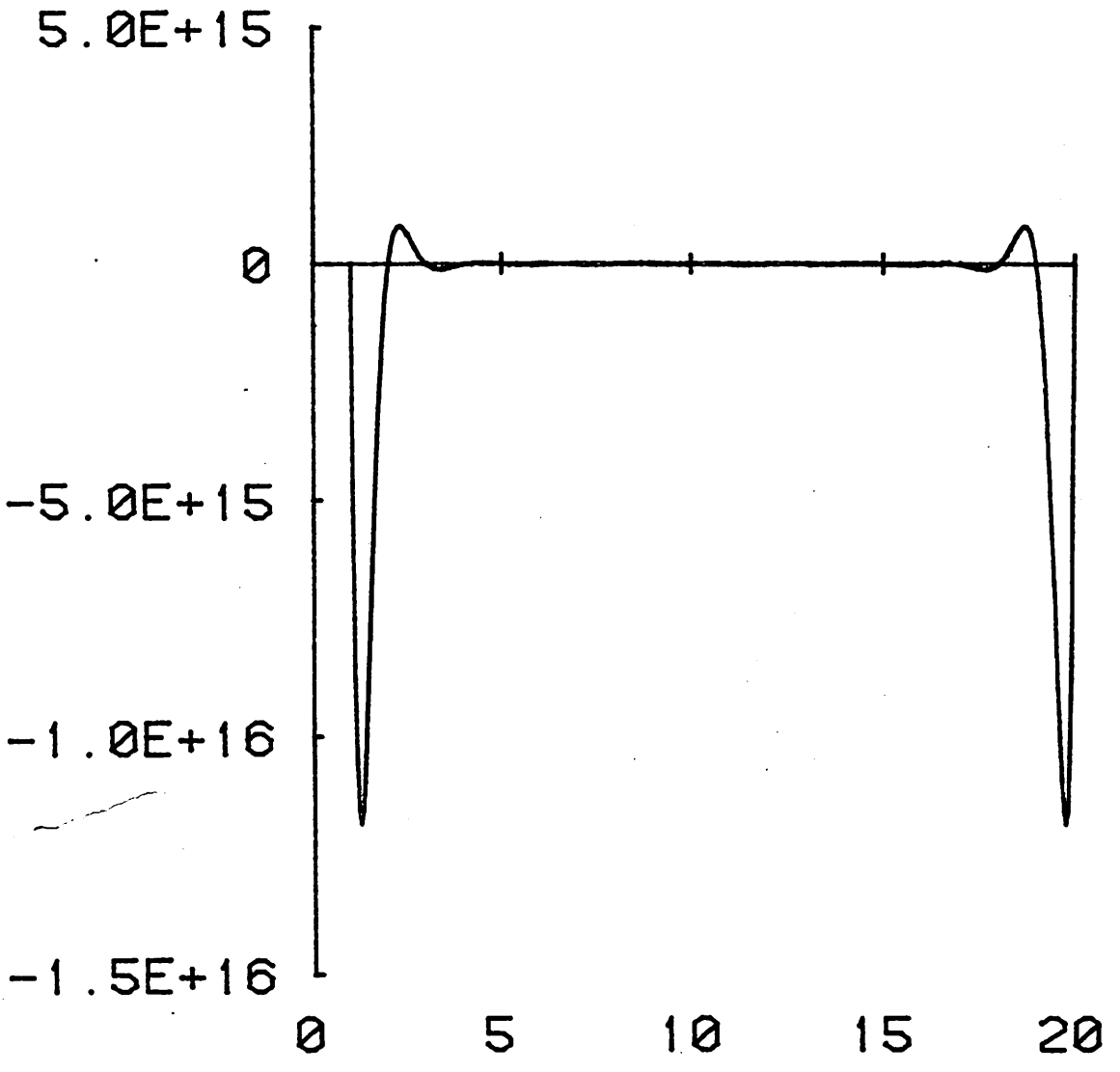

Figure X.2 Wilkinson's polynomial on [1,20].

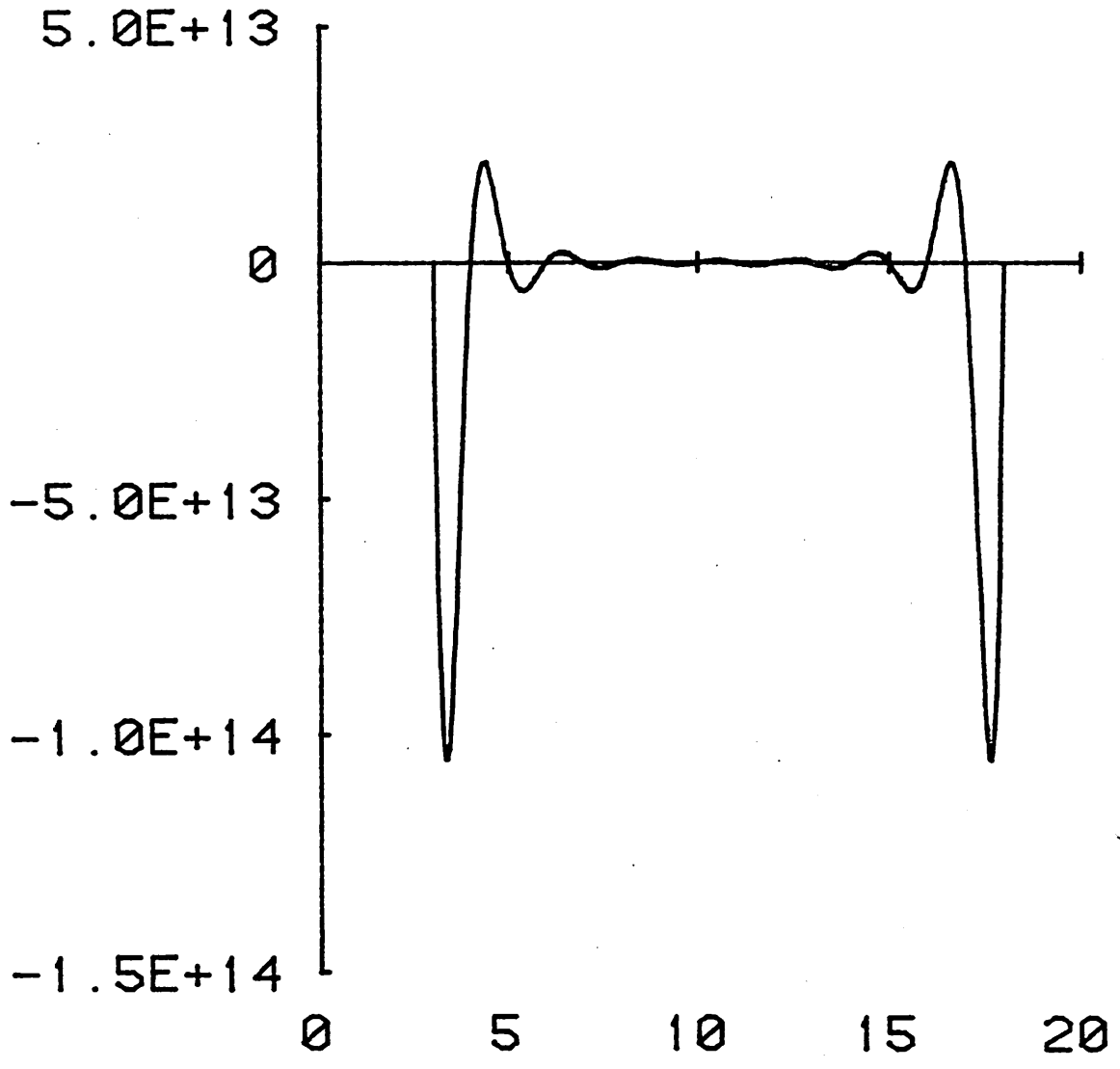

Figure X.3. Wilkinson's polynomial on [3,18].

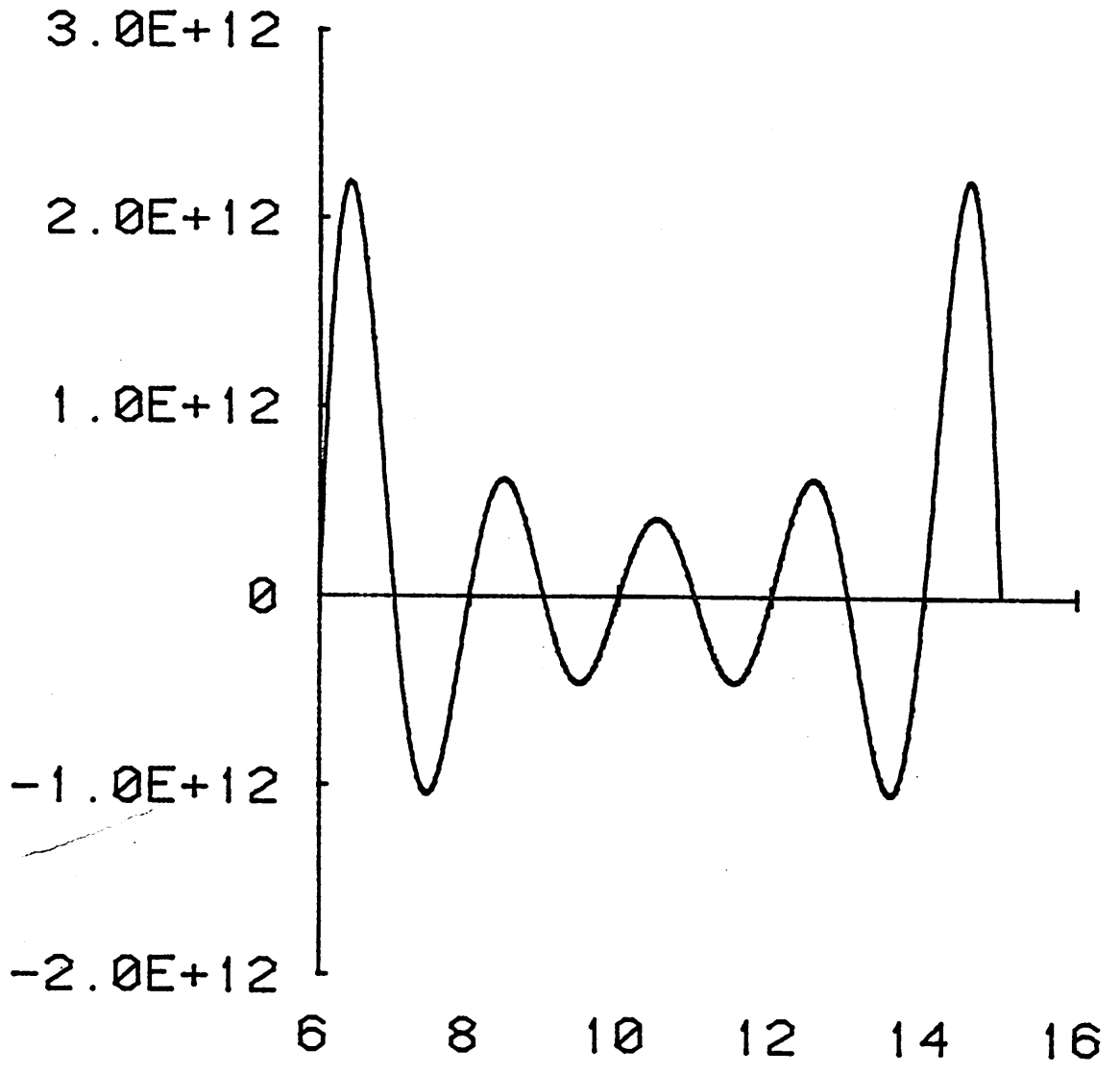

Figure X.4. Wilkinson's polynomial on [6,15].

3. The Nearest Polynomial with a Double Zero

There are many polynomials with a double zero that are close to Wilkinson's polynomial. In the next section we will list some of them. In the present section we will just present the facts about the closest known such polynomials in each norm.

In the relative norm the nearest polynomial on the manifold has a double zero at 14.499... . The distance $\|q\|$ is .11054E-14. The double zero and some of the nearby simple zeros are listed along with their condition numbers and their condition numbers prior to perturbation:

Unperturbed zero	and condition	Perturbed zero	and condition
12	.125E+15	12.15289	.174E+15
13	.212E+15	12.77240	.225E+15
14	.278E+15	14.49963	.963E+13
15	.279E+15	14.49963	.963E+13
16	.210E+15	16.22347	.215E+15
17	.114E+15	16.85795	.159E+15

The coefficients of q are in Table X.2.

The corresponding distance in the uniform norm is .13481E-9.

Unperturbed zero	and condition	Perturbed zero	and condition
13	.607E+9	13.09030	.753E+9
14	.134E+10	13.83087	.123E+10
15	.212E+10	15.48653	.325E+7
16	.241E+10	15.48653	.325E+7
17	.191E+10	17.25351	.205E+10
18	.997E+9	17.83934	.152E+10

In both cases we find that moving to the manifold of double zeros improved the condition of the coalescing zeros appreciably, and thereby improved the overall condition of the polynomial. But some of the nearby zeros actually worsened slightly in condition. Evidently moving to an even higher manifold is in order.

Table X.2

Coefficients of Wilkinson's Polynomial and of the Perturbations to the Nearest Polynomial with a Double Zero

j	p_j	q_j, uniform norm	q_j, relative norm
1	-210	.13452E-9	-.29637E-15
2	20615	.86866E-11	-.19697E-12
3	-1256850	.56091E-12	-.50496E-10
4	.53327 96400 E+8	.36219E-13	-.62696E-8
5	-.16722 80820 E+10	.23388E-14	-.42520E-6
6	.40171 77163 E+11	.15102E-15	-.16922E-4
7	-.75611 11845 E+12	.97516E-17	-.41346E-3
8	.11310 27700 E+14	.62969E-18	-.63804E-2
9	-.13558 51829 E+15	.40660E-19	-.63237E-1
10	.13075 35010 E+16	.26255E-20	-.40560
11	-.10142 29987 E+17	.16954E-21	-1.68308
12	.63030 81210 E+17	.10947E-22	-4.48311
13	-.31133 36432 E+18	.70689E-24	-7.54344
14	.12066 47804 E+19	.45646E-25	-7.81487
15	-.35999 79518 E+19	.29474E-26	-4.79737
16	.80378 11823 E+19	.19032E-27	-1.64939
17	-.12870 93125 E+20	.12290E-28	-.29168
18	.13803 75975 E+20	.79357E-30	-.23138E-1
19	-.87529 48037 E+19	.51242E-31	-.64163E-3
20	.24329 02008 E+19	.33088E-32	-.34188E-5

4. Interesting Polynomials near Wilkinson's

Tables X.3 and X.4 list a number of interesting polynomials near Wilkinson's which have one or more multiple zeros. The first columns list the distance to the polynomial from Wilkinson's, $\|q\|$, and the multiple zeros ζ. All multiple zeros are double except those marked (3) which are triple. In the last columns are listed the worst condition number of a multiple zero ζ and the worst condition number among the simple zeros.

Table X.3 is based on relative changes in the coefficients of Wilkinson's polynomial. Table X.4 is based on the uniform norm in which all the weights are 1. Some of the entries are incomplete; to conserve paper some of the computer codes involved did not print all details for some of the less interesting polynomials.

All the polynomials listed represent solutions of equations presented in chapters III-V. Most of the solutions are local minima. The likely candidates for global minima in each category are indicated by *.

There are apparently a very large number of solutions for the cases of 2, 3, or 4 double zeros. To keep computing expenses in bounds it was necessary to discontinue the computation after a certain arbitrary number, usually 20, of these solutions had been found. Even these are not all listed in the tables; some were omitted whose norms are larger than those listed.

Table X.3

Interesting Polynomials Near Wilkinson's, Relative Norm

	$\|q\|$	ζ's	Worst condition numbers Multiple zero	Simple zero
	unperturbed	polynomial		.279E+15
* 1.	.110E-14	14.4996	.963E+13	.225E+15
2.	.127E-14	15.5295	.771E+13	.122E+16
3.	.127E-14	13.472	.895E+13	.124E+16
* 4.	.128E-14	13.471, 15.531	.895E+13	.113E+15
5.	.192E-14	12.446	.631E+13	.349E+15
6.	.201E-14	16.562	.444E+13	.573E+15
7.	.202E-14	12.442, 16.563	.629E+13	.196E+15
8.	.376E-14	11.420	.340E+13	.111E+15
9.	.392E-14	12.467, 14.535	.963E+13	.188E+15
10.	.454E-14	17.600	.173E+13	.449E+15
11.	.454E-14	11.413, 17.600	.337E+13	.825E+14
12.	.485E-14	14.454, 16.543	.977E+13	.185E+15
13.	.615E-14	11.436, 15.573	.751E+13	.156E+15
14.	.889E-14	13.397, 17.587	.886E+13	.120E+15
15.	.956E-14	10.396	.141E+13	.459E+14
16.	.101E-13	12.509, 15.451	.841E+13	.218E+15
17.	.104E-13	11.454, 13.565	.965E+13	.301E+15
18.	.110E-13	11.454, 16.516	.479E+13	.711E+15
19.	.112E-13	13.573, 16.513	.980E+13	.120E+16
*20.	.112E-13	11.451, 13.572, 16.513	.980E+13	.224E+15
21.	.131E-13	10.407, 16.612	.417E+13	.816E+14
22.	.133E-13	12.527, 17.579	.706E+13	.200E+15
23.	.144E-13	11.468, 14.361	.108E+14	.201E+15
24.	.162E-13	18.646	.410E+12	.534E+14
25.	.163E-13	10.381, 18.645	.137E+13	.274E+14
26.	.176E-13	15.383, 17.570	.944E+13	.240E+15
27.	.182E-13	10.417, 14.668	.106E+14	.214E+15
28.	.197E-13	10.419, 15.378	.968E+13	.307E+15
29.	.197E-13	10.417, 17.571	.186E+13	.874E+15
30.	.197E-13	10.418, 15.373, 17.572	.978E+13	.250E+15

Table X.3 (continued)

	$\|q\|$	ζ's	Worst condition numbers Multiple zero	Simple zero
31.	.210E-13	14.708, 17.564	.114E+14	.360E+15
32.	.210E-13	10.412, 14.708, 17.564	.114E+14	.250E+15
33.	.261E-13	12.475, 15.407, 17.565	.881E+13	.231E+15
34.	.264E-13	12.300, 18.638	.597E+13	.619E+14
35.	.272E-13	11.487, 14.417, 16.559	.104E+14	.178E+15
36.	.298E-13	12.496, 14.492, 16.525	.926E+13	.532E+14
37.	.317E-13	9.372	.447E+12	.164E+14
*38.	.341E-13	13.978(3)	.482E+12	.431E+14
39.	.343E-13	11.474, 13.514, 15.494	.876E+13	.114E+15
40.	.367E-13	15.038(3)	.413E+12	.466E+14
41.	.423E-13	12.921(3)	.414E+12	.344E+14
42.	.547E-13	16.105(3)	.253E+12	.199E+15
43.	.696E-13	11.868(3)	.268E+12	.195E+14
*44.	.110E-12	11.458, 13.531, 15.466, 17.549	.886E+13	.170E+14
45.	.113E-12	19.710	.440E+11	.555E+13
46.	.118E-12	17.181(3)	.104E+12	.136E+14
47.	.130E-12	10.448, 12.557, 14.396, 18.617	.881E+13	.567E+14
48.	.136E-12	10.464, 12.526, 14.472, 16.543	.934E+13	.358E+14
49.	.138E-12	8.349	.107E+12	.465E+13
50.	.145E-12	9.417, 13.584, 15.444, 17.560	.857E+13	.405E+14
51.	.150E-12	10.816(3)	.131E+12	.818E+13
52.	.175E-12	10.462, 12.544, 15.440, 17.554	.822E+13	.225E+15
53.	.299E-12	10.477, 12.501, 14.542, 17.519	.102E+14	.164E+15
54.	.321E-12	9.443, 12.309, 14.708, 15.777	.123E+14	.270E+15
55.	.409E-12	18.273(3)	.257E+11	.438E+13
56.	.426E-12	9.766(3)	.485E+11	.268E+13
57.	.808E-12	7.325	.191E+11	.115E+13
58.	.160E-11	8.717(3)	.135E+11	.652E+12

Table X.3 (continued)

	$\|q\|$	ζ's	Worst condition numbers Multiple zero	Simple zero
59.	.285E-11	19.401(3)	.283E+10	.719E+12
60.	.652E-11	6.302	.247E+10	.230E+12
61.	.808E-11	7.669(3)	.281E+9	.116E+12
62.	.561E-10	6.620(3)	.424E+9	.200E+11
63.	.751E-10	5.277	.224E+9	.384E+11
64.	.555E-9	5.570(3)	.450E+8	.362E+10
65.	.131E-8	4.252	.135E+8	.633E+10
66.	.822E-8	4.519(3)	.320E+7	.609E+9
67.	.381E-7	3.226	.493E+6	.975E+9
68.	.198E-6	3.464(3)	.142E+6	.982E+8
69.	.213E-5	2.196	.950E+4	.162E+9
70.	.888E-5	2.404(3)	.345E+4	.157E+8
71.	.320E-3	1.160	.720E+2	.389E+8
72.	.976E-3	1.331(3)	.366E+2	.426E+7
73.	1.414	0.0	.615E+1	.165E+11
74.	1.732	0.0(3)	.286E+1	.120E+10
75.	2.19614	-117.314		
76.	2.73772	- 9.579		

Table X.4

Interesting Polynomials Near Wilkinson's, Uniform Norm

	$\|q\|$	ζ's	Worst condition numbers Multiple zero	Simple zero
	Unperturbed	polynomial		.241E+10
* 1.	.135E-9	15.487	.325E+7	.152E+10
2.	.142E-9	16.524	.271E+7	.295E+10
3.	.183E-9	14.452	.268E+7	.281E+10
4.	.223E-9	17.567	.148E+7	.186E+10
5.	.350E-9	13.419	.156E+7	.120E+10
6.	.570E-9	18.619	.471E+6	.624E+9
7.	.936E-9	12.388	.645E+6	.417E+9
8.	.294E-8	19.691	.664E+5	.126E+9
9.	.352E-8	11.358	.189E+6	.106E+9
*10.	.477E-8	14.465, 16.537	.271E+7	.856E+9
11.	.723E-8	13.431, 17.578	.159E+7	.137E+10
12.	.114E-7	15.449, 17.550	.327E+7	.117E+10
13.	.135E-7	12.397, 18.625		
14.	.159E-7	11.361, 19.692	.190E+6	.121E+9
15.	.190E-7	10.329	.384E+5	.219E+8
16.	.215E-7	13.454, 15.557	.333E+7	.165E+10
17.	.247E-7	14.387, 18.607	.261E+7	.911E+9
18.	.338E-7	14.547, 17.515	.299E+7	.186E+10
19.	.359E-7	13.478, 16.422	.306E+7	.177E+10
20.	.389E-7	12.414, 16.628	.262E+7	.955E+9
21.	.446E-7	13.493, 18.597	.177E+7	.159E+10
22.	.489E-7	12.421, 17.490	.172E+7	.167E+10
23.	.602E-7	16.341, 18.589		
24.	.633E-7	12.431, 14.643		
*25.	.106E-6	16.025(3)	.835E+4	.440E+9
26.	.107E-6	14.948(3)	.944E+4	.342E+9
27.	.151E-6	9.300	.529E+4	.319E+7
28.	.156E-6	13.877(3)	.724E+4	.212E+9
29.	.159E-6	17.112(3)	.477E+4	.214E+9
30.	.326E-6	12.811(3)	.385E+4	.901E+8

Table X.4 (continued)

	$\|q\|$	ζ's	Worst condition numbers Multiple zero	Simple zero
31.	.398E-6	18.216(3)	.158E+4	.445E+8
32.	.972E-6	11.747(3)	.143E+4	.262E+8
*33.	.120E-5	13.451, 16.372, 18.585	.326E+7	.198E+10
34.	.127E-5	12.442, 14.592, 17.539	.312E+7	.197E+10
35.	.186E-5	8.271		
36.	.192E-5	13.485, 15.496, 17.528	.309E+7	.245E+9
37.	.206E-5	19.358(3)	.226E+3	.251E+8
38.	.208E-5	11.396, 15.820, 18.596	.520E+7	.418E+10
39.	.221E-5	13.416, 15.638, 18.574	.385E+7	.193E+10
40.	.221E-5	11.397, 16.273, 18.598	.481E+7	.376E+10
41.	.239E-5	12.461, 15.384, 17.579	.364E+7	.128E+10
42.	.301E-5	12.457, 14.533, 16.469	.270E+7	.107E+10
43.	.337E-5	12.460, 14.522, 18.581		
44.	.379E-5	12.468, 16.428, 18.575		
45.	.417E-5	10.685(3)	.368E+3	.513E+7
46.	.440E-5	14.563, 16.439, 18.573		
47.	.496E-5	12.492, 15.530, 18.558		
48.	.502E-5	11.411, 15.564, 17.492		
49.	.531E-5	11.420, 14.116, 18.648		
50.	.746E-5	11.428, 14.255, 16.742		
51.	.768E-5	11.429, 14.269, 17.335		
52.	.896E-5	11.431, 13.597, 15.311		
53.	.948E-5	13.605, 15.299, 19.668		
54.	.265E-4	9.625(3)	.637E+2	.657E+6
55.	.309E-4	16.019, 17.181, 19.646		
56.	.378E-4	7.242	.255E+2	.252E+5
*57.	.520E-4	12.447, 14.547, 16.447, 18.570	.274E+7	.382E+8
58.	.252E-3	11.446, 13.567, 16.444, 18.564		
59.	.259E-3	8.565(3)	.718E+1	.526E+5
60.	.327E-3	11.455, 13.528, 15.512, 18.537		

Table X.4 (continued)

	$\|q\|$	ζ's	Worst condition numbers Multiple zero	Simple zero
61.	.340E-3	11.468, 14.434, 16.495, 18.556		
62.	.728E-3	12.580, 14.355, 17.039, 19.645		
63.	.141E-2	6.213	.757	.132E+4
64.	.416E-2	7.504(3)	.495	.361E+4
65.	.113	5.183	.106E-1	.397E+2
66.	.120	6.444(3)	.191E-1	.198E+3
67.	7.24	5.382(3)	.355E-3	.970E+1
68.	25.3	4.152	.539E-4	.193E+1
69.	1161.4	4.318(3)	.254E-5	.474E+0
70.	.256E+5	3.120		
71.	.770E+6	3.251(3)		
72.	.331E+9	2.085		
73.	.525E+10	2.181(3)		
74.	.679E+15	1.062		
75.	.144E+16	1.140(3)		

5. Discussion of Results

Apparently Wilkinson's polynomial lies near a thicket of intersecting branches of the manifold of polynomials with a double zero; see Figure X.5. Although there is a unique point on this manifold closest to p, there are other locally closest points in different directions that are not much further away. In turn the self-intersections of the manifold, which form the manifold of polynomials with two double zeros, may be found not much further from p than the first manifold. And by steps that are increasingly larger, but not overwhelmingly so, it is possible to obtain 3 or 4 double zeros or a triple zero.

Perhaps the polynomials whose zeros are the integers from 1 to n form a family akin to the finite segments of the Hilbert matrix [11]. These ill conditioned matrices have the property that there is no obvious perturbation to a matrix of lower rank that results in a perturbed matrix of satisfactory condition. For large n, rather, there is a sequence of possible perturbations to nearest matrices of rank n-1, n-2, etc. Each perturbation in this sequence has the property that it is neither much larger than the previous perturbation nor much smaller than the next. Furthermore the corresponding sequence of nearest matrices of rank n-1, n-2, etc. has the property that each matrix is somewhat better conditioned than the previous but somewhat less well conditioned than the next one. Thus the ill condition of a Hilbert segment can not be satisfactorily "explained" as due to a small perturbation of a well conditioned matrix of lower rank.

If an analogy with the Hilbert segments is appropriate, then Wilkinson's polynomial can not be satisfactorily "explained" by means

of the numerical methods described in previous chapters. A satisfactory explanation would entail an understanding and description of the geometry of the manifolds of polynomials with multiple zeros and their intersections.

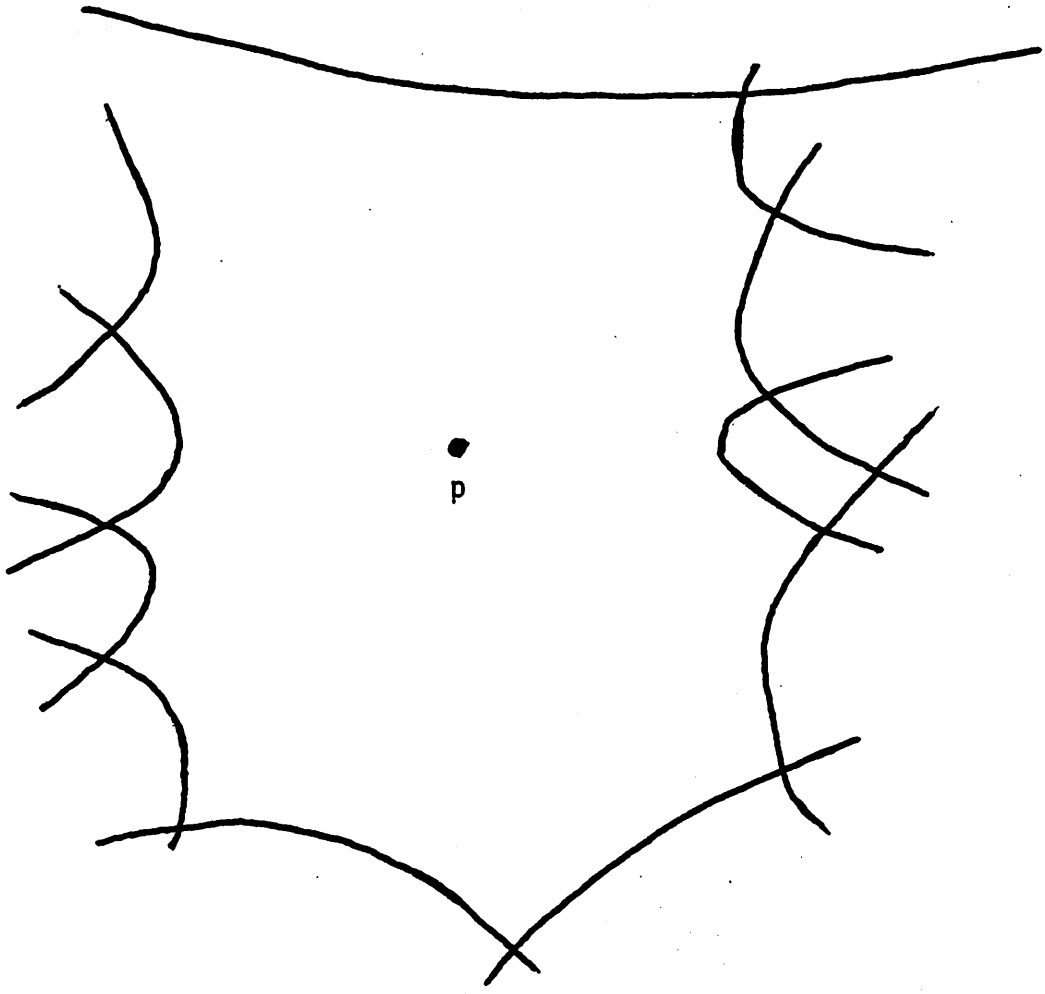

Figure X.5. A mental picture of the manifold thicket surrounding Wilkinson's polynomial.

6. Numerical Results for Translation

Here we summarize some results for translating Wilkinson's polynomial. The zeros of the translated polynomial are $\pm 0.5, \pm 1.5, \ldots, \pm 9.5$. In the uniform norm the worst conditioned of these are ± 8.5 with condition numbers of 72, which are well enough conditioned for most purposes. In contrast the condition numbers for ± 0.5 are .877E-12.

The nearest polynomial with a double zero had $\zeta = \pm 7.979$ and $\|q\| = .437E\text{-}2$. Only the condition of the coalescing zeros was improved significantly, to .402.

Thus in this norm the effects of translation go much farther toward "amelioration" of ill condition than do any of the movements to manifolds of multiple zeros.

When the translation to an even polynomial is carried out, some of the coefficients in the translated polynomial vanish. Thus, in the norm that measures relative changes in coefficients, some of the weights become infinite. Some of the computer codes do not handle this case properly so only partial results are available.

The worst zeros now are ± 7.5 with condition numbers .127E+5. The nearest polynomial with a double zero appears to be a polynomial with two double zeros at $\zeta = \pm 6.979$. Two double zeros are to be expected since the infinite weights constrain the perturbation to be even. In contrast, only one double zero was obtained for the uniform weights.

Numerical difficulties prevented accurate determination of $\|q\|$. The difficulties arose from the fact that the code expected only one double zero so that the other one was poorly determined as two single zeros. The codes for two double zeros found $\zeta = \pm 8.201$ with

$\|q\| = .247E-4$ but they seem to have missed the polynomial with $\zeta = \pm 6.979$.

7. Zeros in Geometrical Progression

In [34] Wilkinson also discussed the polynomial of degree 20 whose zeros are in the geometrical series $2^{-1},2^{-2},\ldots,2^{-20}$. From one point of view these zeros are all remarkably well conditioned despite their apparent crowding near zero. Thus just as the first polynomial was ill conditioned yet free from clustering in its zeros, this second polynomial seems well conditioned despite what seems to be extreme clustering.

For this polynomial, however, all depends on the point of view. Whereas the first polynomial was ill conditioned whether uniform or relative perturbations were considered, the second is only well conditioned when relative perturbations are at stake.

When relative changes both in the coefficients and the zeros are considered, the worst zero is 2^{-11} and its condition number is 65.7; the other condition numbers are remarkably similar, the best being 8.43. In contrast, when absolute changes in the coefficients and zeros are at issue, the worst is 2^{-19} with a condition number of .109E+59; the best is 2^{-1} with condition number .210E+7. In the uniform norm, then, this polynomial is far worse conditioned than the better known one with a linear distribution of zeros.

It should be realized that the coefficients of p range from 1 to 1E63 in magnitude. With such a wide range of magnitudes of both coefficients and zeros, numerical problems made it difficult to obtain meaningful results. What results were obtained frequently failed some of the tests described in chapter VIII. Since floating point underflow is not detected by the CDC 6400 and overflow is known to have occurred, we will not discuss these possibly contaminated results.

CHAPTER XI

CONCLUDING REMARKS

We have given methods for finding nearby polynomials with various configurations of multiple zeros. We have exhibited examples to show that these methods provide the answer we would expect when the correct answer is obvious.

For a polynomial like Wilkinson's, however, there is no obvious answer, and these methods do not provide satisfactory explanations of the ill condition of such polynomials. Rather the numerical results provide evidence of an inherently complicated structure of the manifold of polynomials with multiple zeros.

Finally there are intermediate polynomials for which the "correct" answer is no longer so obvious but which do not seem to present so confusing a picture as Wilkinson's polynomial. For such intermediate cases our methods sometimes provide results that seem satisfactory and sometimes do not. But it is not yet clear whether "unsatisfactory" results are due to defects in algorithms or inappropriate expectations about the existence of satisfactory nearby polynomials.

In each of these areas there is ample scope for further research. For the "obvious" cases we would like to be able to specify starting points for iterative methods which could be guaranteed to converge quickly to the global minimum.

For the intermediate cases we would like to know simple criteria for deciding when, for instance, nearby polynomials with complex conjugate pairs of double zeros may exist. More generally we would like to know when a solution ζ of the equations we wish to solve does <u>not</u> exist in a particular region, so that we need not waste time looking there.

Sketchy information on where to look for ζ is known for the case of one double zero, but for other configurations the only known facts are that the dimensionality of the problem is less than might have been thought, because certain Lagrange multipliers vanish in the complex case. We would like to have a simple criterion in the real case, that will tell us when we may rely on that theorem about Lagrange multipliers, when we must check real configurations of higher multiplicity, and when we must check for complex conjugate multiple zeros.

The new expansion technique discussed in chapter VII provides some interesting questions. In how large a region can realistic bounds be computed easily? It would be desirable to have a symbolic algebra program to provide these tedious bounds automatically. Do these bounds have any significant advantages over Smith's [42]?

A task of a different sort is to render the existing mass of algorithmic ideas and devices into mathematical software. The computer codes with which the research reported here was conducted were constantly changing and required considerable experience to direct the search and interpret the results. They were dependent on the local computing environment in many ways and most likely contain some errors, which would probably not affect the results presented in previous chapters.

In contrast, respectable mathematical software is carefully specified, written, documented, and tested. Then it is independently examined and tested again. The experienced computer programmer now recognizes, moreover, that the production of quality mathematical software from its raw materials entails as much effort as providing those

raw materials. Consequently that production will be deferred to another occasion in this case.

The final, and perhaps most difficult, challenge is to unravel the nature of the manifold of polynomials with multiple zeros, particularly in the vicinity of polynomials like Wilkinson's. Although numerical investigations may sometimes be helpful, probably the principal factor for success will be the investigator's competence in algebraic geometry.

Turning now to a more general point of view, we should recall that one reason for studying polynomials is that they are simpler than the often infinite dimensional eigenvalue problems they frequently replace. Thus the more general problem might be stated as follows: given a linear operator, some of whose eigenvalues are ill conditioned, what is the nearest linear operator whose eigenvalues, some of them multiple, are all well conditioned?

Ruhe [27], Wilkinson [36], and Kahan [16] have all given bounds for the distance to the nearest matrix with a multiple eigenvalue. Kahan [17] and Golub and Wilkinson [39] have also surveyed the known theory. But there are no known computational techniques which are even as reliable as those discussed previously for zeros of polynomials. The closest related work is that of Kågström and Ruhe [15] on finding the Jordan canonical form of a matrix. Otherwise the many refractory aspects of the problem remain untouched for future investigators.

APPENDICES

1. Using the Zeros of a Polynomial to Compute Its Coefficients

Our object is to display the well known algorithm for computing the coefficients of a monic polynomial from its zeros. If we are to determine the p_j in

$$\prod_{j=1}^{n}(\tau-\zeta_j) = \tau^n + \sum_{j=1}^{n} p_j \tau^{n-j}$$

and we expand directly we find

$$p_j = \sum_{\begin{pmatrix}\text{over all } \binom{n}{j} \text{ combinations}\\ \text{of the n } (-\zeta_j)\text{'s taken j}\\ \text{at a time}\end{pmatrix}} [\Pi \text{ (of the } (-\zeta_j)\text{'s in each combination)}]$$

We can avoid this n! calculation by building up the coefficients recursively. If we have a polynomial

$$p^k(\tau) = \sum_{j=0}^{k} p_j^k \tau^{k-j} = \prod_{j=1}^{k}(\tau-\zeta_j) , \qquad p_0^k = 1 ,$$

we can form the polynomial of degree $k+1$ by multiplication by $(\tau-\zeta_{k+1})$:

$$\begin{aligned} p^{k+1}(\tau) &= \Big(\sum_{j=0}^{k} p_j^k \tau^{k-j}\Big)(\tau-\zeta_{k+1}) \\ &= \sum_{j=0}^{k} p_j^k \tau^{k+1-j} - \sum_{j=0}^{k} p_j^k(-\zeta_{k+1})\tau^{k-j} \\ &= \sum_{j=0}^{k+1} p_j^{(k+1)} \tau^{k+1-j} \end{aligned}$$

where

$$p_j^{k+1} = \begin{cases} -\zeta_{k+1} p_k^k & , \quad j = k+1 , \\ p_j^k - \zeta_{k+1} p_{j-1}^k & , \quad j = k,k-1,\ldots,2,1 , \\ 1 & , \quad j = 0 . \end{cases}$$

We list the coefficients in the order that they could be successively computed and overlaid in storage.

In the case of real polynomials, we wish to avoid complex arithmetic by considering complex zeros and their conjugates together. Then

$$p^{k+2}(\tau) = p^k(\tau)\cdot(\tau^2 - 2(\mathrm{Re}\ \zeta_{k+1})\tau + |\zeta_{k+1}|^2)$$

so

$$p_j^{k+2} = \begin{cases} |\zeta_{k+1}|^2 p_k^k & , \quad j = k+2 , \\ -2(\mathrm{Re}\ \zeta_{k+1}) p_k^k + |\zeta_{k+1}|^2 p_{k-1}^k & , \quad j = k+1 , \\ p_j^k - 2(\mathrm{Re}\ \zeta_{k+1}) p_{j-1}^k + |\zeta_{k+1}|^2 p_{j-2}^k & , \quad j = k,\ldots,3,2 , \\ p_1^k - 2(\mathrm{Re}\ \zeta_{k+1}) & , \quad j = 1 , \\ 1 & , \quad j = 0 . \end{cases}$$

It may happen that we are only interested in the last few coefficients or the first few. The formulas above may be used for the first few coefficients corresponding to high powers of τ.

To find formulas for the last few coefficients, corresponding to low powers of τ, we redefine the p_j^k as follows:

$$p^k(\tau) = \sum_{j=0}^{k} p_j^k \tau^j .$$

Then

$$p_j^{k+1} = \begin{cases} 1 & , \quad j = k+1 , \\ p_{j-1}^k - \zeta_{k+1} p_j^k & , \quad j = k,\ldots,2,1 , \\ -\zeta_{k+1} p_0^k & , \quad j = 0 , \end{cases}$$

and in the case $\zeta_{k+2} = \overline{\zeta_{k+1}}$,

$$p_j^{k+2} = \begin{cases} 1 & , \quad j = k+2 , \\ p_{k-1}^k - 2 \operatorname{Re} \zeta_{k+1} & , \quad j = k+1 , \\ p_{j-2}^k - 2 \operatorname{Re} \zeta_{k+1} p_{j-1}^k + |\zeta_{k+1}|^2 p_j^k & , \quad j = k,\ldots,2 , \\ -2 \operatorname{Re} \zeta_{k+1} p_0^k + |\zeta_{k+1}|^2 p_1^k & , \quad j = 1 , \\ |\zeta_{k+1}|^2 p_0^k & , \quad j = 0 . \end{cases}$$

2. Simultaneous Evaluation of a Polynomial and Some of its Derivatives

Ways of efficiently evaluating a polynomial and its derivatives simultaneously from the coefficients have been studied by Shaw and Traub [29] among others.

Rice [26] has argued that, given the zeros ζ_j of a polynomial, computing the product

$$p(\tau) = \prod_{j=1}^{n} (\tau-\zeta_j)$$

is usually the method of evaluation that minimizes the uncertainty in $p(\tau)$. When the polynomial is evaluated in this way the relative error in the final result, due to rounding errors, is always small on a properly designed machine. In contrast the relative error of the evaluation from the coefficients is usually large when τ is near one of the ζ_j.

Furthermore if the zeros are the primary data, rather than the coefficients, the attempt to compute the coefficients from the zeros will, in the presence of rounding errors, produce wrong coefficients which will be the coefficients of another polynomial with different zeros. If the new zeros are ill conditioned they may be rather far removed from the zeros we started with.

Therefore we prefer to evaluate polynomials and their derivatives directly from the zeros if they are the primary data. Typical expressions for the polynomial and two derivatives follow:

Method N:

$$p(\tau) = \prod_{j=1}^{n} (\tau-\zeta_j)$$

$$\frac{p'(\tau)}{p(\tau)} = \sum_{j=1}^{n} \frac{1}{\tau-\zeta_j}$$

$$\frac{p''(\tau)}{p(\tau)} = \left(\sum_{j=1}^{n} \frac{1}{\tau-\zeta_j}\right)^2 - \sum_{j=1}^{n} \frac{1}{(\tau-\zeta_j)^2}$$

Similar expressions for higher derivatives may be found by means of Newton's identities which are described in Householder [12]. These expressions have the defect, however, that in the presence of rounding errors, they tend to have high relative errors which are revealed by cancellation at the end. Thus if $\tau \doteq \zeta_j$ in the expression for p''/p, the two subexpressions will tend to cancel with subsequent severe loss of significant figures. By algebraic manipulation we may be able to find forms for these expressions in which cancellation is not preordained. For instance

$$\frac{p''}{p} = 2 \sum_{j=1}^{n-1} \frac{1}{\tau-\zeta_j}\left(\sum_{k=j+1}^{n} \frac{1}{\tau-\zeta_k}\right)$$

but this expression is not applicable when $\tau = \zeta_j$ exactly.

Therefore it is helpful to use different methods for computing a polynomial and its derivatives from its zeros. These methods are based on the observation that if

$$p(\tau) = \prod_{j=1}^{n} (\tau-\zeta_j) = \sum_{j=0}^{n} p_j \tau^{n-j}, \qquad p_0 = 1,$$

then $p(0) = p_n$, $p'(0) = p_{n-1}$, and in general, $p^k(0) = k!p_{n-k}$. Therefore we can evaluate the polynomial and m derivatives at 0 by computing the last $m+1$ coefficients of p from its zeros ζ_j.

Moreover we can evaluate p and its derivatives at α by computing the coefficients of the polynomial whose zeros are $\zeta_j - \alpha$:

Method A:

$$p^{(k)}(\alpha) = k!\{n-k \text{ coefficient of polynomial whose zeros are } \zeta_j-\alpha\}$$

Another method is based on the observation that

$$q(\tau) = \sum_{j=0}^{n} p_{n-j}\tau^{n-j} = \tau^n p(\tfrac{1}{\tau}) = \tau^n \prod_{j=1}^{n} (\tfrac{1}{\tau} - \zeta_j) = \prod_{j=1}^{n} (-\zeta_j) \prod_{j=1}^{n} (\tau - \tfrac{1}{\zeta_j}) .$$

Then

$$\begin{aligned} q_k &= p_{n-k} \\ &= \prod_{j=1}^{n} (-\zeta_j) \cdot \{k^{\text{th}} \text{ coefficient of polynomial whose zeros are } \tfrac{1}{\zeta_j}\} \\ &= p^{(k)}(0)/k! \quad . \end{aligned}$$

So continuing as before,

Method B:

$$p^{(k)}(\alpha) = k!p(\alpha)\{k \text{ coefficient of polynomial whose zeros are } \tfrac{1}{\zeta_j-\alpha}\}$$

Like Newton's identities, however, this method is undefined if $\alpha = \zeta_j$.

We might conduct operation counts to help choose from among these methods. They all require $mn + O(m^2) + O(n)$ operations to evaluate a polynomial and m derivatives. Therefore we choose Method A since it is applicable even when $\tau = \zeta_j$.

3. Partial Derivatives of a Deflated Function of a Complex Variable

When minimizing norms of functions of complex variables we are often required to find zeros of non-analytic functions of a complex variable. There seems to be little general theory for such functions other than that of two real analytic functions of two real variables. Consequently when finding zeros of such functions by Newton's method we solve systems of two equations.

Having found one solution we may wish to deflate it out in order to find other solutions. Fortunately there is a way of deflating such functions that makes sense. In contrast, there is no completely satisfactory way of deflating solutions of systems of n real equations in n variables for $n \geq 2$.

$f(\tau)$ will be the function to be deflated; it is not analytic. Let $\zeta_1,\ldots,\zeta_k$ be the zeros to be removed; we will divide $f(\tau)$ by the polynomial

$$p(\tau) = \prod_{j=1}^{k} (\tau-\zeta_k) .$$

The deflated function $g(\tau) = f(\tau)/p(\tau)$ is not analytic, but the analyticity of p will simplify the expressions for the partial derivatives of $\operatorname{Re} g$ and $\operatorname{Im} g$.

Let $(\dot{})$ represent a differential operator, either $\frac{\partial}{\partial \operatorname{Re}\tau}$ or $\frac{\partial}{\partial \operatorname{Im}\tau}$. Then

$$(\operatorname{Re}\dot{g}) = (\operatorname{Re}\dot{\frac{f}{p}}) = [\operatorname{Re} f \operatorname{Re}(\dot{\frac{1}{p}}) - \operatorname{Im} f \operatorname{Im}(\dot{\frac{1}{p}})]$$

$$= [(\operatorname{Re}\dot{f})\operatorname{Re}(\frac{1}{p}) + \operatorname{Re} f\,(\operatorname{Re}\dot{\frac{1}{p}}) - \operatorname{Im} f\,(\operatorname{Im}\dot{(\frac{1}{p})}) - (\operatorname{Im}\dot{f})\operatorname{Im}(\frac{1}{p})]$$

and

$$\frac{\partial \operatorname{Re} g}{\partial \operatorname{Re} \tau} = \operatorname{Re}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Re} f}{\partial \operatorname{Re} \tau} - \operatorname{Im}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Im} f}{\partial \operatorname{Re} \tau} - \operatorname{Re}\left(\frac{fp'}{p^2}\right) ,$$

where p' represents the complex derivative of p, $\frac{\partial p(\tau)}{\partial \tau}$. Similarly

$$\frac{\partial \operatorname{Re} g}{\partial \operatorname{Im} \tau} = \operatorname{Re}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Re} f}{\partial \operatorname{Im} \tau} - \operatorname{Im}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Im} f}{\partial \operatorname{Im} \tau} + \operatorname{Im}\left(\frac{fp'}{p^2}\right) ,$$

and

$$\frac{\partial \operatorname{Im} g}{\partial \operatorname{Re} \tau} = \operatorname{Re}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Im} f}{\partial \operatorname{Re} \tau} + \operatorname{Im}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Re} f}{\partial \operatorname{Re} \tau} - \operatorname{Im}\left(\frac{fp'}{p^2}\right) ,$$

and

$$\frac{\partial \operatorname{Im} g}{\partial \operatorname{Im} \tau} = \operatorname{Re}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Im} f}{\partial \operatorname{Im} \tau} + \operatorname{Im}\left(\frac{1}{p}\right)\frac{\partial \operatorname{Re} f}{\partial \operatorname{Im} \tau} - \operatorname{Re}\left(\frac{fp'}{p^2}\right) .$$

These partial derivatives are now in a form suitable for use in Newton's method applied to a system of two real equations in two unknowns.

4. Computing the Divided Differences Required in the Equations to be Solved for Complex Conjugate Double Zeros in Chapter IV

Below will be found the recurrences required to compute

$$\Delta_p, \quad \Delta_{p'}, \quad \text{and} \quad \Delta_k ,$$

the divided differences of section IV.3. We will also obtain derivatives with respect to $\operatorname{Re}\zeta$ and $\operatorname{Im}\zeta$ for use with Newton's method.

$$\Delta_k \equiv (\operatorname{Im}\zeta^k)/(\operatorname{Im}\zeta)$$

so $\Delta_0 = 0$, $\Delta_1 = 1$, and

$$\Delta_k = \operatorname{Re}\zeta\,\Delta_{k-1} + \operatorname{Re}(\zeta^{k-1}) .$$

If we write

$$\Delta_k^r = \frac{\partial \Delta_k}{\partial \operatorname{Re}\zeta}$$

and

$$\Delta_k^m = \frac{\partial \Delta_k}{\partial \operatorname{Im}\zeta}$$

we find:

$$\begin{aligned}
&\Delta_0 = 0 , \\
&\Delta_k = \operatorname{Re}\zeta\,\Delta_{k-1} + \operatorname{Re}(\zeta^{k-1}) , \\
&\Delta_0^m = \Delta_1^m = \Delta_2^m = 0 , \\
&\Delta_k^m = \operatorname{Re}\zeta\,\Delta_{k-1}^m - (k-1)\operatorname{Im}(\zeta^{k-2}) , \\
&\Delta_0^r = \Delta_1^r = 0 , \\
&\Delta_k^r = \Delta_{k-1} + \operatorname{Re}\zeta\,\Delta_{k-1}^r + (k-1)\operatorname{Re}(\zeta^{k-2}) .
\end{aligned}$$

In order to compute Δ_p and $\Delta_{\zeta p'}$ it is necessary to start by recalling the formulas for updating p and p'. If the zeros of p are ζ_i, $1 \le i \le n$, then we could define

$$p_k = \prod_{i=1}^{k} (\zeta-\zeta_i) \ .$$

Then we may imagine updating p_k by one real zero ζ_+ or by two complex conjugate zeros ζ_+ and $\bar{\zeta}_+$. Then

$$\begin{aligned}
p_0 = 1 \ , \quad & p_{k+1} = (\zeta-\zeta_+)p_k \ , \\
& p_{k+2} = \{\zeta^2 - 2(\mathrm{Re}\ \zeta_+)\zeta + |\zeta_+|^2\}p_k \ , \\
p_0' = 0 \ , \quad & p_{k+1}' = (\zeta-\zeta_+)p_k' + p_k \ , \\
& p_{k+2}' = \{\zeta^2 - 2(\mathrm{Re}\ \zeta_+)\zeta + |\zeta_+|^2\}p_k' + 2(\zeta - \mathrm{Re}\ \zeta_+)p_k \ , \\
p_0'' = 0 \ , \quad & p_{k+1}'' = (\zeta-\zeta_+)p_k'' + 2p_k' \ , \\
& p_{k+2}'' = \{\zeta^2 - 2(\mathrm{Re}\ \zeta_+)\zeta + |\zeta_+|^2\}p_{k+2}'' \\
& \qquad + 4(\zeta - \mathrm{Re}\ \zeta_+)p_k' + 2p_k \ .
\end{aligned}$$

The formulas for computing Δ_p and its derivatives are as follows:

$$\begin{aligned}
\frac{\mathrm{Im}\ p_0}{\mathrm{Im}\ \zeta} &= 0 \ , \\
\frac{\mathrm{Im}\ p_{k+1}}{\mathrm{Im}\ \zeta} &= \mathrm{Re}\ p_k + \mathrm{Re}(\zeta-\zeta_+)\left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) \ , \\
\frac{\mathrm{Im}\ p_{k+2}}{\mathrm{Im}\ \zeta} &= \mathrm{Re}\{\zeta^2 - 2(\mathrm{Re}\ \zeta_+)\zeta + |\zeta_+|^2\}\cdot\left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) + 2\,\mathrm{Re}\,(\zeta-\zeta_+)\mathrm{Re}\ p_k \ , \\
\frac{\partial}{\partial\,\mathrm{Re}\,\zeta}\left(\frac{\mathrm{Im}\ p_{k+1}}{\mathrm{Im}\ \zeta}\right) &= \mathrm{Re}\ p_k' + \left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) + \mathrm{Re}(\zeta-\zeta_+)\frac{\partial}{\partial\,\mathrm{Re}\,\zeta}\left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) \ , \\
\frac{\partial}{\partial\,\mathrm{Re}\,\zeta}\left(\frac{\mathrm{Im}\ p_{k+2}}{\mathrm{Im}\ \zeta}\right) &= 2\ \mathrm{Re}(\zeta-\zeta_+)\left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) + 2\,\mathrm{Re}\,p_k + 2\ \mathrm{Re}(\zeta-\zeta_+)\mathrm{Re}\ p_k' \\
&\quad + \mathrm{Re}\{\zeta^2 - 2(\mathrm{Re}\ \zeta_+)\zeta + |\zeta_+|^2\}\frac{\partial}{\partial\,\mathrm{Re}\,\zeta}\left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) \ , \\
\frac{\partial}{\partial\,\mathrm{Im}\,\zeta}\left(\frac{\mathrm{Im}\ p_{k+1}}{\mathrm{Im}\ \zeta}\right) &= -\mathrm{Im}\ p_k' + \mathrm{Re}(\zeta-\zeta_+)\frac{\partial}{\partial\,\mathrm{Im}\,\zeta}\left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) \ , \\
\frac{\partial}{\partial\,\mathrm{Im}\,\zeta}\left(\frac{\mathrm{Im}\ p_{k+2}}{\mathrm{Im}\ \zeta}\right) &= -2\ \mathrm{Re}(\zeta-\zeta_+)\mathrm{Im}\ p_k' - 2\,\mathrm{Im}\,p_k \\
&\quad + \mathrm{Re}\{\zeta^2 - 2\,\mathrm{Re}(\zeta_+)\zeta + |\zeta_+|^2\}\frac{\partial}{\partial\,\mathrm{Im}\,\zeta}\left(\frac{\mathrm{Im}\ p_k}{\mathrm{Im}\ \zeta}\right) \ .
\end{aligned}$$

Note in passing that

$$\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta} = \frac{\operatorname{Im} p_k'}{\operatorname{Im}\zeta}\,.$$

We now state the corresponding formulas required for $\Delta_{\zeta p'}$:

$$\frac{\operatorname{Im}\zeta p'}{\operatorname{Im}\zeta} = \operatorname{Re} p' + \operatorname{Re}\zeta\,\frac{\operatorname{Im} p'}{\operatorname{Im}\zeta} = \operatorname{Re} p' + \operatorname{Re}\zeta\,\frac{\partial}{\partial \operatorname{Re}\zeta}\left(\frac{\operatorname{Im} p}{\operatorname{Im}\zeta}\right),$$

$$\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im}\zeta p'}{\operatorname{Im}\zeta} = \operatorname{Re} p'' + \frac{\partial}{\partial \operatorname{Re}\zeta}\left(\frac{\operatorname{Im} p}{\operatorname{Im}\zeta}\right) + \operatorname{Re}\zeta\,\frac{\partial^2}{(\partial \operatorname{Re}\zeta)^2}\left(\frac{\operatorname{Im} p}{\operatorname{Im}\zeta}\right),$$

$$\frac{\partial}{\partial \operatorname{Im}\zeta}\frac{\operatorname{Im}\zeta p'}{\operatorname{Im}\zeta} = -\operatorname{Im} p'' + \operatorname{Re}\zeta\,\frac{\partial^2}{\partial \operatorname{Re}\zeta\,\partial \operatorname{Im}\zeta}\left(\frac{\operatorname{Im} p}{\operatorname{Im}\zeta}\right),$$

$$\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_{k+1}}{\operatorname{Im}\zeta} = \operatorname{Re} p_k' + \frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta} + \operatorname{Re}(\zeta-\zeta_+)\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta},$$

$$\frac{\partial^2}{(\partial \operatorname{Re}\zeta)^2}\frac{\operatorname{Im} p_{k+1}}{\operatorname{Im}\zeta} = \operatorname{Re} p_k'' + 2\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta} + \operatorname{Re}(\zeta-\zeta_+)\frac{\partial^2}{(\partial \operatorname{Re}\zeta)^2}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta},$$

$$\frac{\partial^2}{\partial \operatorname{Im}\zeta\,\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_{k+1}}{\operatorname{Im}\zeta} = -\operatorname{Im} p_k'' + \frac{\partial}{\partial \operatorname{Im}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta} + \operatorname{Re}(\zeta-\zeta_+)\frac{\partial^2}{\partial \operatorname{Im}\zeta\,\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta},$$

$$\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_{k+2}}{\operatorname{Im}\zeta} = 2\operatorname{Re} p_k + 2\operatorname{Re}(\zeta-\zeta_+)\left(\operatorname{Re} p_k' + \frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta}\right) + \operatorname{Re}\{\zeta^2 - 2\operatorname{Re}(\zeta_+)\zeta + |\zeta_+|^2\}\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta},$$

$$\frac{\partial^2}{(\partial \operatorname{Re}\zeta)^2}\frac{\operatorname{Im} p_{k+2}}{\operatorname{Im}\zeta} = 4\operatorname{Re} p_k' + 2\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta} + 2\operatorname{Re}(\zeta-\zeta_+)\left(\operatorname{Re} p_k'' + 2\frac{\partial}{\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta}\right) + \operatorname{Re}\{\zeta^2 - 2(\operatorname{Re}\zeta_+)\zeta + |\zeta_+|^2\}\frac{\partial^2}{(\partial \operatorname{Re}\zeta)^2}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta},$$

$$\frac{\partial^2}{\partial \operatorname{Im}\zeta\,\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_{k+2}}{\operatorname{Im}\zeta} = -4\operatorname{Im} p_k' + 2\operatorname{Re}(\zeta-\zeta_+)\left(-\operatorname{Im} p_k'' + \frac{\partial}{\partial \operatorname{Im}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta}\right) + \operatorname{Re}\{\zeta^2 - 2(\operatorname{Re}\zeta_+)\zeta + |\zeta_+|^2\}\frac{\partial}{\partial \operatorname{Im}\zeta\,\partial \operatorname{Re}\zeta}\frac{\operatorname{Im} p_k}{\operatorname{Im}\zeta}.$$

5. Computing the Divided Differences Required in the Equations to be Solved for Two Real Double Zeros in Chapter V

The equations which follow provide recurrent methods for computing the divided differences required to solve (6.6) of chapter V.

Recall

$$\Delta_k \equiv (\zeta_1^k - \zeta_2^k)/(\zeta_1 - \zeta_2) .$$

Therefore $\Delta_0 = 0$, and $\Delta_1 = 1$. We may verify that

$$\Delta_k = \zeta_1 \Delta_{k-1} + \zeta_2^{k-1} = \zeta_2 \Delta_{k-1} + \zeta_1^{k-1} ,$$

$$\frac{\partial \Delta_k}{\partial \zeta_1} = \zeta_1 \frac{\partial \Delta_{k-1}}{\partial \zeta_1} + \Delta_{k-1} ,$$

$$\frac{\partial \Delta_k}{\partial \zeta_2} = \zeta_2 \frac{\partial \Delta_{k-1}}{\partial \zeta_2} + \Delta_{k-1} .$$

The equations for $\Delta_{p,k}$ are more complicated. Recall that

$$\Delta_{p,k} \equiv \frac{\zeta_2^k p(\zeta_1) - \zeta_1^k p(\zeta_2)}{\zeta_1 - \zeta_2} .$$

To compute $\Delta_{p,k}$ when p is given in factored form, it is necessary to fix k and develop $\Delta_{p,k}$ recursively by considering each factor of p in turn. To start, suppose $p \equiv 1$; then $\Delta_{p,k} = -\Delta_k$. Now suppose that $\Delta_{p,k}$ is known and p is to be multiplied by a linear factor $(\tau-\alpha)$. Then denoting the new divided difference by $\underset{+1}{\Delta}_{p,k}$,

$$\begin{aligned} \underset{+1}{\Delta}_{p,k} &\equiv (\zeta_2^k p(\zeta_1)(\zeta_1-\alpha) - \zeta_1^k p(\zeta_2)(\zeta_2-\alpha))/(\zeta_1-\zeta_2) \\ &= \zeta_1\zeta_2\Delta_{p,k-1} - \alpha\Delta_{p,k} . \end{aligned}$$

Furthermore

$$\frac{\partial \underset{+1}{\Delta} p,k}{\partial \zeta_1} = \zeta_2\{\zeta_1 \frac{\partial \Delta_{p,k-1}}{\partial \zeta_1} + \Delta_{p,k-1}\} - \alpha \frac{\partial \Delta_{p,k}}{\partial \zeta_1}$$

and

$$\frac{\partial \underset{+1}{\Delta} p,k}{\partial \zeta_2} = \zeta_1\{\zeta_2 \frac{\partial \Delta_{p,k-1}}{\partial \zeta_2} + \Delta_{p,k-1}\} - \alpha \frac{\partial \Delta_{p,k}}{\partial \zeta_2} .$$

If p were real and α were complex it would be desirable to update p by the real quadratic factor $(\tau-\alpha)(\tau-\bar{\alpha})$. Let $\underset{+2}{\Delta} p,k$ represent this updated divided difference:

$$\underset{+2}{\Delta} p,k \equiv \frac{\zeta_2^k(\zeta_1^2 - 2\,\mathrm{Re}\,\alpha\,\zeta_1 + |\alpha|^2) - \zeta_1^k(\zeta_2^2 - 2\,\mathrm{Re}\,\alpha\,\zeta_2 + |\alpha|^2)}{\zeta_1 - \zeta_2}$$

$$= \zeta_1^2\zeta_2^2\Delta_{p,k-2} - 2(\mathrm{Re}\,\alpha)\zeta_1\zeta_2\Delta_{p,k-1} + |\alpha|^2\zeta_{p,k}$$

$$\frac{\partial \underset{+2}{\Delta} p,k}{\partial \zeta_1} = \zeta_1\zeta_2^2\{2\Delta_{p,k-2} + \zeta_1 \frac{\partial \Delta_{p,k-2}}{\partial \zeta_1}\} + |\alpha|^2 \frac{\partial \Delta_{p,k}}{\partial \zeta_1}$$

$$- 2(\mathrm{Re}\,\alpha)\zeta_2\{\Delta_{p,k-1} + \zeta_1 \frac{\partial \Delta_{p,k-1}}{\partial \zeta_1}\} .$$

The corresponding equation for $\frac{\partial \underset{+2}{\Delta}}{\partial \zeta_2}$ may be found by interchanging ζ_1 and ζ_2.

Similar methods may be applied to

$$\Delta_{p',k} \equiv \frac{\zeta_2^k p'(\zeta_1) - \zeta_1^k p'(\zeta_2)}{\zeta_1 - \zeta_2} .$$

Note first that

$$p_{+1}(\tau) = (\tau-\alpha)p(\tau) ,$$
$$p'_{+1}(\tau) = (\tau-\alpha)p'(\tau) + p(\tau) ,$$
$$p''_{+1}(\tau) = (\tau-\alpha)p''(\tau) + 2p'(\tau) ,$$

and

$$p_{+2}(\tau) = (\tau-\alpha)(\tau-\bar{\alpha})p(\tau) ,$$
$$p'_{+2}(\tau) = 2(\tau - \text{Re } \alpha)p(\tau) + (\tau-\alpha)(\tau-\bar{\alpha})p'(\tau) ,$$
$$p''_{+2}(\tau) = 2p(\tau) + 4(\tau - \text{Re } \alpha)p'(\tau) + (\tau-\alpha)(\tau-\bar{\alpha})p''(\tau) .$$

Then

$$\Delta_{+1} p',k = \zeta_1\zeta_2\Delta_{p',k-1} - \alpha\Delta_{p',k} + \Delta_{p,k} ,$$

$$\frac{\partial \Delta_{+1} p',k}{\partial \zeta_1} = \zeta_2\{\zeta_1 \frac{\partial \Delta_{p',k-1}}{\partial \zeta_1} + \Delta_{p',k-1}\} - \alpha \frac{\partial \Delta_{p',k}}{\partial \zeta_1} + \frac{\partial \Delta_{p,k}}{\partial \zeta_1} ,$$

$$\Delta_{+2} p',k = (\zeta_1\zeta_2)^2\Delta_{p',k-2} - 2(\text{Re } \alpha)\zeta_1\zeta_2\Delta_{p',k-1} + |\alpha|^2\Delta_{p',k}$$
$$+ 2\zeta_1\zeta_2\Delta_{p,k-1} - 2(\text{Re } \alpha)\Delta_{p,k} ,$$

$$\frac{\partial \Delta_{+2} p',k}{\partial \zeta_1} = \zeta_1\zeta_2^2\{\zeta_1 \frac{\partial \Delta_{p',k-2}}{\partial \zeta_1} + 2\Delta_{p',k-2}\} + |\alpha|^2 \frac{\partial \Delta_{p',k}}{\partial \zeta_1}$$
$$- 2(\text{Re } \alpha)\zeta_1\{\zeta_2 \frac{\partial \Delta_{p',k-1}}{\partial \zeta_1} + \Delta_{p',k-1}\}$$
$$+ 2\zeta_2\{\zeta_1 \frac{\partial \Delta_{p,k-1}}{\partial \zeta_1} + \Delta_{p,k-1}\} - 2(\text{Re } \alpha)\frac{\partial \Delta_{p,k}}{\partial \zeta_1} .$$

These formulas may be used to calculate $\Delta_{p,k}$ and $\Delta_{p',k}$ except when $k = 0$ or $k = 1$. In those cases the formulas would require $\Delta_{p,-1}$ which is not defined.

To deal with that difficulty, different formulas for divided differences must be used for $k = 0$ and $k = 1$. These formulas will be based on the finite difference analog of Leibniz' rule:

$$\begin{aligned}\Delta(xy)(\theta_1,\theta_2) &\equiv \frac{xy(\theta_1) - xy(\theta_2)}{\theta_1 - \theta_2} \\ &= \left(\frac{x(\theta_1) + x(\theta_2)}{2}\right)\left(\frac{y(\theta_1) - y(\theta_2)}{\theta_1 - \theta_2}\right) \\ &\quad + \left(\frac{y(\theta_1) + y(\theta_2)}{2}\right)\left(\frac{x(\theta_1) - x(\theta_2)}{\theta_1 - \theta_2}\right) .\end{aligned}$$

Here x and y are functions of a single variable; the divided difference of the product xy is sought for the points (θ_1,θ_2). This and other divided difference identities may be found in the book by Milne-Thomson [23].

For our application, x will be $p(\tau)$ or $p'(\tau)$ and y will be the updating factor $(\tau-\alpha)$ or $(\tau-\alpha)(\tau-\bar{\alpha})$. We find that

$$\begin{aligned}\overset{\Delta}{{}_{+1}}p,0 &= \frac{(\zeta_1-\alpha) + (\zeta_2-\alpha)}{2}\Delta_{p,0} + \frac{p(\zeta_1) + p(\zeta_2)}{2} , \\ \frac{\partial \overset{\Delta}{{}_{+1}}p,0}{\partial\zeta_1} &= \frac{(\zeta_1-\alpha) + (\zeta_2-\alpha)}{2}\frac{\partial\Delta_{p,0}}{\partial\zeta_1} + \frac{1}{2}\Delta_{p,0} + \frac{1}{2}p'(\zeta_1) , \\ \overset{\Delta}{{}_{+2}}p,0 &= \left(\frac{(\zeta_1-\alpha)(\zeta_1-\bar{\alpha}) + (\zeta_2-\alpha)(\zeta_2-\bar{\alpha})}{2}\right)\Delta_{p,0} \\ &\quad + \frac{p(\zeta_1) + p(\zeta_2)}{2}((\zeta_1 - \mathrm{Re}\ \alpha)(\zeta_2 - \mathrm{Re}\ \alpha)) , \\ \frac{\partial \overset{\Delta}{{}_{+2}}p,0}{\partial\zeta_1} &= \left(\frac{(\zeta_1-\alpha)(\zeta_1-\bar{\alpha}) + (\zeta_2-\alpha)(\zeta_2-\bar{\alpha})}{2}\right)\frac{\partial\Delta_{p,0}}{\partial\zeta_1} + (\zeta_1 - \mathrm{Re}\ \alpha)\Delta_{p,0} \\ &\quad + \frac{p(\zeta_1) + p(\zeta_2)}{2} + \frac{1}{2}p'(\zeta_1)((\zeta_1 - \mathrm{Re}\ \alpha)(\zeta_2 - \mathrm{Re}\ \alpha)) ,\end{aligned}$$

$$\underset{+2}{\Delta} p,1 = \left(\frac{\zeta_1+\zeta_2}{2} - 2\,\mathrm{Re}\,\alpha\right)\zeta_1\zeta_2\underset{}{\Delta}_{p,0} + |\alpha|^2\underset{}{\Delta}_{p,1} + \zeta_1\zeta_2\left(\frac{p(\zeta_1)+p(\zeta_2)}{2}\right) ,$$

$$\frac{\partial \underset{+2}{\Delta} p,1}{\partial\zeta_1} = \left(\frac{\zeta_1+\zeta_2}{2} - 2\,\mathrm{Re}\,\alpha\right)\zeta_1\zeta_2\frac{\partial\underset{}{\Delta}_{p,0}}{\partial\zeta_1} + \zeta_2\left(\zeta_1 + \tfrac{1}{2}\zeta_2 - 2\,\mathrm{Re}\,\alpha\right)\underset{}{\Delta}_{p,0}$$

$$+ |\alpha|^2\frac{\partial\underset{}{\Delta}_{p,1}}{\partial\zeta_1} + \tfrac{1}{2}\zeta_2\left(\zeta_1 p'(\zeta_1) + p(\zeta_1)\right) .$$

Similarly

$$\underset{+1}{\Delta} p',0 = \left(\frac{(\zeta_1-\alpha)+(\zeta_2-\alpha)}{2}\right)\underset{}{\Delta}_{p',0} + \left(\frac{p'(\zeta_1)+p'(\zeta_2)}{2}\right) + \underset{}{\Delta}_{p,0} ,$$

$$\frac{\partial \underset{+1}{\Delta} p',0}{\partial\zeta_1} = \left(\frac{(\zeta_1-\alpha)+(\zeta_2-\alpha)}{2}\right)\frac{\partial\underset{}{\Delta}_{p',0}}{\partial\zeta_1} + \tfrac{1}{2}\underset{}{\Delta}_{p',0} + \tfrac{1}{2}p''(\zeta_1) + \frac{\partial\underset{}{\Delta}_{p,0}}{\partial\zeta_1} ,$$

$$\underset{+2}{\Delta} p',0 = \left(\frac{p'(\zeta_1)+p'(\zeta_2)}{2}\right)\{(\zeta_1 - \mathrm{Re}\,\alpha) + (\zeta_2 - \mathrm{Re}\,\alpha)\}$$

$$+ \left(\frac{(\zeta_1-\alpha)(\zeta_1-\bar{\alpha}) + (\zeta_2-\alpha)(\zeta_2-\bar{\alpha})}{2}\right)\underset{}{\Delta}_{p',0}$$

$$+ \left((\zeta_1 - \mathrm{Re}\,\alpha) + (\zeta_2 - \mathrm{Re}\,\alpha)\right)\underset{}{\Delta}_{p,0} + p(\zeta_1) + p(\zeta_2) ,$$

$$\frac{\partial \underset{+2}{\Delta} p',0}{\partial\zeta_1} = \left(\frac{(\zeta_1 - \mathrm{Re}\,\alpha) + (\zeta_2 - \mathrm{Re}\,\alpha)}{2}\right)p''(\zeta_1) + \frac{p'(\zeta_1)+p'(\zeta_2)}{2}$$

$$+ (\zeta_1 - \mathrm{Re}\,\alpha)\underset{}{\Delta}_{p',0} + \left(\frac{(\zeta_1-\alpha)(\zeta_1-\bar{\alpha}) + (\zeta_2-\alpha)(\zeta_2-\bar{\alpha})}{2}\right)\frac{\partial\underset{}{\Delta}_{p',0}}{\partial\zeta_1}$$

$$+ \left((\zeta_1 - \mathrm{Re}\,\alpha) + (\zeta_2 - \mathrm{Re}\,\alpha)\right)\frac{\partial\underset{}{\Delta}_{p,0}}{\partial\zeta_1} + \underset{}{\Delta}_{p,0} + p'(\zeta_1) .$$

Finally

$$\underset{+2}{\Delta} p',1 = \zeta_1\zeta_2\left(\frac{p'(\zeta_1)+p'(\zeta_2)}{2}\right) + \left(\frac{\zeta_1+\zeta_2}{2} - 2\,\mathrm{Re}\,\alpha\right)\zeta_1\zeta_2\underset{}{\Delta}_{p',0}$$

$$+ |\alpha|^2\underset{}{\Delta}_{p',1} + 2\zeta_1\zeta_2\underset{}{\Delta}_{p,0} - 2(\mathrm{Re}\,\alpha)\underset{}{\Delta}_{p,1} ,$$

$$\frac{\partial \Delta_{+2}^{p',1}}{\partial \zeta_1} = \frac{1}{2}\zeta_1\zeta_2 p''(\zeta_1) + \zeta_2\left(\frac{p'(\zeta_1)+p'(\zeta_2)}{2}\right) - 2(\mathrm{Re}\ \alpha)\frac{\partial \Delta_{p,1}}{\partial \zeta_1}$$
$$+ \zeta_2\{\frac{1}{2}\zeta_1 + (\frac{\zeta_1+\zeta_2}{2} - 2\,\mathrm{Re}\,\alpha)\}\Delta_{p',0}$$
$$+ (\frac{\zeta_1+\zeta_2}{2} - 2\,\mathrm{Re}\,\alpha)\zeta_1\zeta_2\frac{\partial \Delta_{p',0}}{\partial \zeta_1}$$
$$+ |\alpha|^2\frac{\partial \Delta_{p',1}}{\partial \zeta_1} + 2\zeta_2\{\zeta_1\frac{\partial \Delta_{p,0}}{\partial \zeta_1} + \Delta_{p,0}\}\ .$$

Taken together, the foregoing equations provide all the divided differences required in chapter V. To inhibit convergence to the remaining unwanted solutions it is still necessary to use the deflation techniques of section 5 of that chapter.

6. The Lagrange Multiplier Theorem

The following corollary of the Fredholm Alternative Theorem provides the basis for the use of Lagrange multipliers to find stationary points of functions subject to constraints. The vector ℓ^* is the vector of Lagrange multipliers.

Theorem. Let B map C^n to C^m. Then

$$(\text{for every } x \in C^n,\ Bx = 0 \Rightarrow y^*x = 0)$$

if and only if there exists an $\ell^* \in C^m$ such that

$$y^* = \ell^* B .$$

See Dunford and Schwartz [9, p. 609] for a statement of the Fredholm Alternative Theorem in an arbitrary Banach space, and for references to a proof.

REFERENCES

1. G. Bliss, Algebraic Functions, American Mathematical Society, Providence, 1933.

2. R. Brent, Algorithms for Minimization without Derivatives, Prentice Hall, 1973.

3. R. Buck, "Applications of Duality in Approximation Theory," in H. Garabedian, ed., Proceedings of the Symposium on Approximation of Functions, Elsevier, pp. 27-42, 1965.

4. B. Carnahan, H. Luther, and J. Wilkes, Applied Numerical Methods, Wiley, 1969.

5. G. Collins, "Subresultants and Reduced Polynomial Remainder Sequences," Journal of the ACM 14, pp. 128-142, 1967.

6. G. Dahlquist and A. Björck, Numerical Methods, Prentice Hall, 1974.

7. J. Daniel, "Correcting Approximations to Multiple Roots of Polynomials," Numerische Mathematik 9, pp. 99-102, 1966.

8. D. Dunaway, "Calculation of Zeros of a Real Polynomial through Factorization using Euclid's Algorithm," SIAM Journal on Numerical Analysis 11, pp. 1105-1120, 1974.

9. N. Dunford and J. Schwartz, Linear Operators Part I: General Theory, Interscience, 1958.

10. M. Eichler, Introduction to the Theory of Algebraic Numbers and Functions, Academic Press, 1966.

11. G. Forsythe and C. Moler, Computer Solution of Linear Algebraic Equations, Prentice Hall, 1967.

12. A. Householder, The Numerical Treatment of a Single Nonlinear Equation, McGraw Hill, 1970.

13. IMSL, IMSL Library 3 Reference Manual, International Mathematical and Statistical Libraries, Houston, 1975.

14. M. Jenkins, "Algorithm 493. Zeros of a Real Polynomial," ACM Transactions on Mathematical Software 1, pp. 178-189, 1975.

15. B. Kågström and A. Ruhe, An Algorithm for Numerical Computation of the Jordan Normal Form of a Complex Matrix, Department of Information Processing, University of Umeå, Sweden, 1974.

16. W. Kahan, Ruhe's Theorem on Ill-Conditioned Eigenvalues, Technical Report #5, Department of Computer Science, University of California, Berkeley, 1971.

17. W. Kahan, Conserving Confluence Curbs Ill Condition, Technical Report #6, Department of Computer Science, University of California, Berkeley, 1972.

18. W. Kahan, Implementation of Algorithms, Technical Report #20, Department of Computer Science, University of California, Berkeley, 1973.

19. W. Kahan, unpublished notes on polynomials, 1973.

20. T. Kibble, Classical Mechanics, McGraw Hill, London, p. 25, 1965.

21. M. Marden, The Geometry of Polynomials, American Mathematical Society, Providence, 1966.

22. A. Markushevich, Theory of Functions of a Complex Variable, Volume 2, Prentice Hall, pp. 105-112, 1965.

23. L. Milne-Thomson, The Calculus of Finite Differences, Macmillan, London, 1960.

24. R. Moore, Interval Analysis, Prentice Hall, 1966.

25. A. Ostrowski, Solutions of Equations and Systems of Equations, Academic Press, 1966.

26. J. Rice, "On the Conditioning of Polynomial and Rational Forms," Numerische Mathematik 7, pp. 426-435, 1965.

27. J. Rice, "A Theory of Condition," SIAM Journal on Numerical Analysis 3, p. 287, 1966.

28. A. Ruhe, "Properties of a Matrix with a Very Ill-Conditioned Eigenproblem," Numerische Mathematik 15, pp. 57-60, 1970.

29. M. Shaw and J. Traub, "Analysis of a Family of Algorithms for the Evaluation of a Polynomial and some of its Derivatives," Journal of the ACM 21, pp. 161-167, 1974.

30. B. Smith, "ZERPOL, A Zero Finding Algorithm for Polynomials Using Laguerre's Method," M.S. Thesis, Department of Computer Science, University of Toronto, 1967.

31. G. Stewart, "Error Analysis of the Algorithm for Shifting the Zeros of a Polynomial by Synthetic Division," Mathematics of Computation 25, pp. 135-139, 1971.

32. G. Stewart, "The Convergence of Multipoint Iterations to Multiple Zeros," SIAM Journal on Numerical Analysis 11, pp. 1105-1120, 1974.

33. J. Traub, Iterative Methods for Solution of Equations, Prentice Hall, 1964.

34. J. Wilkinson, Rounding Errors in Algebraic Processes, Prentice Hall, 1963.

35. J. Wilkinson and C. Reinsch, Linear Algebra, Springer Verlag, 1971.

36. J. Wilkinson, "Note on Matrices with a Very Ill-Conditioned Eigenproblem," Numerische Mathematik 19, pp. 66-68, 1972.

37. V. Zaguskin, Handbook of Numerical Methods for the Solution of Algebraic and Transcendental Equations, Pergamon Press, 1961.

38. Association for Computing Machinery, Proceedings of the Second Symposium on Symbolic and Algebraic Manipulation, ACM, 1971.

39. G. Golub and J. Wilkinson, "Ill Conditioned Eigensystems and the Computation of the Jordan Canonical Form," SIAM Review 18, pp. 578-619, 1976.

40. H. Kung and J. Traub, All Algebraic Functions can be Computed Fast, Department of Computer Science, Carnegie-Mellon University, 1976.

41. J. Pinkert, "An Exact Method for Finding the Roots of a Complex Polynomial," ACM Transactions on Mathematical Software 2, pp. 351-363, 1976.

42. B. Smith, "Error Bounds for Zeros of a Polynomial Based Upon Gerschgorin's Theorems," Journal of the ACM 17, pp. 661-674, 1970.

unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Electronics Research Laboratory University of California Berkeley, CA 94720	unclassified 2b. GROUP

3. REPORT TITLE

EXPLAINING AND AMELIORATING THE ILL CONDITION OF ZEROS OF POLYNOMIALS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

David Granville Hough

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 6, 1977	305	42

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-76-C-0013 b. PROJECT NO.	UCB-ERL-M77/30
c. d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

unlimited distribution

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Department of Navy Arlington, VA 22217

13. ABSTRACT

Physical systems can frequently be modeled by polynomial equations. Then interesting properties of the systems can be determined from the zeros of the polynomials. Standard codes compute those zeros from the coefficients in a stable fashion. But what should be done if the zeros are inherently hypersensitive to changes in the coefficients of their polynomials? Newly developed methods can be used to explain such an ill conditioned polynomial by exhibiting a nearby polynomial with one or more multiple zeros which are well conditioned. Furthermore these methods can be abused by uncritically replacing the ill conditioned polynomial with the well conditioned one nearby. When such a replacement is unwarranted, bounds can be obtained on the variation of the zeros corresponding to the uncertainty in the coefficients. One way to obtain such bounds is to exploit the nearby well conditioned polynomial to obtain a revision of the classical Puiseux fractional power series expansions of the zeros.

These notions have been investigated experimentally in a long series of computer calculations. In the course of these calculations the existing stock of numerical techniques has been augmented. A new way is now known for computing the condition numbers which measure the condition of zeros. The previously known equations to be solved for the nearest polynomial with a single multiple zero are now joined by equations for the nearest polynomial with a complex conjugate pair of double zeros and equations for the nearest polynomial with several distinct double zeros. All these equations have simplified forms because certain Lagrange multipliers vanish in the complex case. But some examples demonstrate that when only real perburbations are considered, the Lagrange multipliers do not always vanish. Finally, there is some theory about the location of the nearest polynomial with a double zero. (over)

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT

The numerical experiments show that Newton's method may be used successfully to solve the equations in the cases of greatest interest when the expected result is sufficiently simple. The techniques may also be applied to polynomials such as Wilkinson's famous example whose zeros are the integers from 1 to 20. But then the numerical results suggest that that ill conditioned polynomial can not be explained successfully as a small perturbation of a well conditioned polynomial. Instead Wilkinson's polynomial lies in a region of polynomial space whose geometry seems to be exceptionally complicated.

Bounds on uncertainties in zeros corresponding to uncertainties in coefficients are customarily computed with Taylor series. For ill conditioned simple zeros these Taylor series have radii of convergence that are much too small. The well conditioned multiple zeros of a nearby polynomial are not amenable to Taylor series expansions but may be expanded in a Puiseux fractional power series. These fractional power series, however, also have unsatisfactory regions of convergence. But by choosing a different starting point the convergence problem of the Puiseux series can be overcome to produce, in principle, series that converge rapidly throughout the region of interest. In practice those series are used to produce realistic bounds on the uncertainties in the zeros. Full exploitation of these techniques awaits adequate facilities for symbolic algebra.