38

ITERATION PROCEDURES FOR SIMULTANEOUS EQUATIONS

by

EDWARD JOSEPH CRAIG

S.B., Union College
(1948)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January, 1954

Signature of Author ______________________________
Department of Electrical Engineering, January 11, 1954

Certified by ______________________________
Thesis Supervisor

____
Chairman, Department Committee on Graduate Students

"All things were made through Him, and without Him was made nothing that has been made."

Gospel according to St. John: 1, 3.

ITERATION PROCEDURES FOR SIMULTANEOUS EQUATIONS

by

EDWARD JOSEPH CRAIG

Submitted to the Department of Electrical Engineering on January 11, 1954, in partial fulfillment of the requirements for the degree of Doctor of Science.

ABSTRACT

The main object of this thesis is to offer another attack on the solution of simultaneous equations. The bulk of the work is on linear equations.

Much work has been done on approximation or iteration procedures for the solution of such problems in the last twenty years. The work has been concentrated on linear equations. Until 1950 all these trial and error methods took infinitely many corrections to complete, but some converged so nicely that they have been used extensively. The real attraction, the author believes, lies in the process of guessing an answer. Engineers seem to be fascinated by procedures which will relieve them of the drudgery of mathematics.

As more interest was shown in this type of solution, several authors made great strides. To progress, some picture or theory had to be visualized. This picture amounts to the elevation of the tasteless task of solving linear sets to a geometric problem consisting of locating a point in space.

Once this is done, all procedures can be compared and evaluated. For certain sets of equations (called ill-conditioned) all the procedures which require infinitely many steps are very poor. This thesis discusses finite-step procedures.

It is the aim of this work to show that finite step procedures are possible,* and are the best one can obtain. Moreover, all finite step procedures are a variation on a general procedure which the author states and which is due to Lanczos. To demonstrate this, several such procedures have been devised by the author and are applicable to all types of equations: non-symmetric, non-hermitian complex, skew-symmetric, etc.

The author believes that the best methods have now been generalized, and that the "guess and try" method of solution has now been substantially solved. A complete solution would require a complete evaluation of

* Lest this be misunderstood, the author wishes to point out that finite step procedures were known at least six years ago, and perhaps even by the ancient Greeks.

roundoff error - i.e., the errors resulting when a limited number of digits are used in the computations. This aspect is discussed, but by no means solved.

Thesis Supervisor: William K. Linvill
Title: Associate Professor of Electrical Engineering

ACKNOWLEDGMENT

It would be unfair and incomplete if certain acknowledgments were not made at the start. If there is anything of value in this dissertation, it has been a direct result of the encouragement and advice of Professor William K. Linvill. To him goes my everlasting gratitude.

Thanks also go to Professor Ernst Guillemin and Professor Parry Moon for their cooperation and assistance, and to Messrs. Kenneth Ralston, Jack Porter, and Dean Arden of the Digital Computer Laboratory at M.I.T. who made possible the programming and execution of the experimental results.

Thanks also to Mrs. Marilyn Susskind who typed the thesis.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

All too frequently the solution to engineering problems is obtained implicitly in a set of simultaneous equations. The problem of extracting an explicit answer from these is generally a large one, and one that is usually so difficult that the engineer may not know how it is to be solved.

If the solution of the problem can be reduced to the solution of a set of simultaneous differential equations, there are methods available to the engineer. Most of these are approximate, unless the equations are linear with constant coefficients. If the equations are ordinary, that is, with no derivatives of the unknown present, then again the methods are approximate, unless the equations are linear.

It will be recalled that a linear equation is one in which the unknown or unknowns or any of their derivatives appear in the first degree only. If the system is not linear, one is forced to employ a "guess and correct" procedure to obtain an answer - or to make assumptions so that the set of equations is reduced to a different set, the solution to which is known.

Of late, the construction of large scale automotons to do the burdensome work of carrying out numerical solutions has given engineers a new lease on life. It has also given much impetus to logical "cut and try" procedures, which the author chooses to call iteration procedures.

While the task of solving a large set of simultaneous linear (not differential) equations is a straightforward matter, it is still an unpleasant task to perform by hand. Cramer's Rule, or elimination procedures such as that due to Gauss (or Gauss-Jordan reduction - see ref. 26, Ch. I), are simple in principle but involved in detail and susceptible to error.

Lest the reader become alarmed at this last remark, the author hastens to state that he has no intention of scrapping these procedures or of belittling them. He intends merely to exploit another procedure based originally on a lazier attack on the problem.

The latter procedure was fathered by the solution of non-linear simultaneous equations. The best approximation procedure evolved to date for these is called the Newton-Raphson procedure. In this procedure one guesses or tries to guess the answer, and then attempts to add corrections to this guess by using information obtained from the equations and his trial answer.

Naturally the idea occurred to someone that if a procedure of the iterative "cut and try" type were to work well on linear equations, it might work agreeably well on non-linear equations. In recent years, therefore, in an attempt to expand knowledge of methods for the solution of non-linear equations, much emphasis has been placed on iteration procedures for linear equations. It is hoped that a procedure which works very well with a linear set can be set up to work well with a non-linear set.

With these ideas in mind, the author believes that this thesis presents the best type of procedures for linear equations. The procedures are then extended to non-linear equations.

1.0 Iteration

Consider a specific set of simultaneous equations:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = y_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = y_2$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = y_3.$$

The "a's" and "y's" are known constants, and the "x's" are to be determined. The subscripts on the letters "a" refer to the equation number and variable respectively.

One begins by guessing three numbers to use as trials for x_1, x_2, and x_3. These trials are now checked by substitution in the equations. It is quite obvious that if this is done and the equations are satisfied exactly, then one must have the answer. (Of course, one assumes that the answer to the problem is unique. If more than one answer is possible, then it can be easily shown that an infinite number is possible. Unless specifically stated to the contrary, it will be assumed that in all problems considered there exists an unique answer.)

The probability that one will guess the correct answer is zero; hence after substitution of the trial numbers in the left-hand side and performing the indicated multiplications and additions, one obtains three numbers which in general will not be y_1, y_2, and y_3. Intuition tells us that if these numbers are near the values y, then one is close to the true answer. This is not, unhappily, always true in an absolute sense, but it is true in a relative way.

One is now forced to the rather obvious conclusion that the only information one has concerning the error in these trials is the difference between the computed left side of these equations and the right side. These differences are usually called <u>residuals</u> and are defined in this thesis as

$$r_{1i} = a_{11}x_{1i} + a_{12}x_{2i} + a_{13}x_{3i} - y_1$$

$$r_{2i} = a_{21}x_{1i} + a_{22}x_{2i} + a_{23}x_{3i} - y_2$$

$$r_{3i} = a_{31}x_{1i} + a_{32}x_{2i} + a_{33}x_{3i} - y_3.$$

Here the subscript "i" refers to the ith trial or approximation to the answer.

At this point one is led to the conclusion that the problem is solved if the residuals are zero, and one need only continually try guesses until this condition is obtained. This is surely a hit or miss way of doing the problem, so a short list of some of the methods previously devised will be given.

A. Alter x_{1i} sufficiently so that r_{1i} is reduced to zero. This changes the other two residuals, however, so then alter x_{2i} to reduce r_{2i} to zero, and so on. When this is done, each residual in turn is reduced to zero, but all the other residuals change at each step. Thus, after reducing the first residual to zero, the reduction of the second residual to zero causes the first residual to change, and it is never possible theoretically to make more than one residual zero at a time. The problem is similar to that of a mother putting six small children to bed. While she puts one child in bed and tucks him (or her) in, the other five get up again. Her only hope is to tire them out one at a time (or employ more drastic means). This type of iteration sometimes converges (i.e., she sometimes gets them all in bed at the same time) and sometimes diverges (i.e., the children get wilder at each step). The method is aptly termed a "relaxation" procedure, but the author is certain that this name did not arise in connection with the example cited, however well applicable! The condition for convergence in this procedure is that the coefficients a_{ij} and a_{ji} are equal. This is usually called a symmetry condition. There are naturally several variations on this scheme, but the exact answer cannot be obtained in a finite number of steps.

B. Instead of changing one trial component at a time, change all of them (three in the example above) such an amount that the sum of the squared residuals is minimized. Such a procedure is evidently assured of convergence. This was first done by subtracting from each trial x_i a piece of the residual r_i, or a weighted function of r_i. This thesis deals solely with iterative procedures in which the new trial guess is obtained from the old by subtracting (adding) corrections. It must be clear that this is surely not the only way this can be done, but it is believed to be the simplest. Thus, if p_{1i}, p_{2i}, p_{3i} are three numbers,

$$x_{1(i+1)} = x_{1i} - p_{1i}$$

$$x_{2(i+1)} = x_{2i} - p_{2i}$$

$$x_{3(i+1)} = x_{3i} - p_{3i}.$$

The three numbers "p" are of course chosen by some rule in order that the procedure converge.

C. Related to the above idea are several procedures based on similar ideas. Such procedures as the Method of Descent are stated in the following chapter in detail. Most of these procedures are clever, but incomplete, in that they still require infinitely many approximations. They converge better than the relaxation procedures, but they also involve more work.

D. If by using Cramer's Rule or elimination procedures one can obtain the solution with a finite number of operations, why not in an iteration procedure also? In the last few years such procedures have been found. They are called N-step or minimized-iteration procedures. The N-step implies that ideally N approximations only need be made, where N is the number of equations. This thesis deals primarily with these procedures and hopes to show

a) that these procedures or variations on them are the best that can be obtained,

b) that all such procedures are related, and

c) a general expression for making further procedures has been obtained.

2.0 The Outline of the Thesis

In order to acquaint the reader with the content of this work before a complete reading is undertaken, the author feels obligated to describe the contents of each chapter briefly so that the reader may keep his eyes firmly fixed on the objectives. This is essential, primarily because most proofs given hereinafter are by contradiction or induction. While these are perfectly valid from the viewpoint of a mathematician, they yield little insight into the geometry (or physics) of the problem.

Chapter II will introduce the reader to a geometric interpretation of all procedures. The object here is to lift the uninteresting job of numerical analysis to a level of maps, pictures, and geometric objects. This affords insight which the author feels is valuable. To do this the convenience of matrix algebra is employed. Most of the necessary definitions and theorems on this topic are briefly contained in Appendix I at the end of this work. Before one proceeds it is suggested that this section (Appendix I) be given at least a cursory glance. Chapter II will conclude by stating the conditions for an N-step procedure.

Chapter III treats specifically with the N-step procedures of the author, Stiefel and Hestenes, and Lanczos, and attempts descriptive explanations of these.

Chapter IV presents a general proof of all the procedures stated in Chapter III and points out the fact that, in reality, all these procedures are intimately related to the general scheme due to Lanczos.

Chapter V states the general N-step procedure and gives examples of how more procedures can be constructed. By way of illustration, all the procedures are used in simple examples to clarify their operation.

Chapter VI deals with roundoff error. While these procedures yield exact answers in N steps, they do so only if enough significant figures can be carried. The procedures fall apart if sufficient accuracy is not maintained. This portion of the problem is very important and has not been solved by the author. The problem is discussed at length, but no definite conclusions are drawn. It is hoped that this aspect of the problem will be solved in the near future.

Chapter VII extends the author's procedure to non-linear equations and gives a few examples.

Appendix I is on matrix algebra, and Appendix II gives a short history of this type of iteration procedure, attempting to name some of the individuals who have made significant contributions to the problem.

Appendix III deals with the author's method exclusively. An extension is made to complex matrices, and a general proof for all variations on this procedure is given. This proof is complicated and is included primarily for those who wish to study the procedure in more detail. It is felt that some ideas concerning the influence of roundoff error on the procedure can be obtained from the proof. Further examples are given demonstrating the use of the procedures in the event the characteristic equation of a non-symmetric or complex matrix is desired.

CHAPTER II

GEOMETRIC INTERPRETATION OF ITERATION PROCEDURES

1.0 The Plan of Attack

The problem of the solution of a linear set of equations will be visualized as the task of locating a point in N dimensional space. The initial guess will consist of a point in this space, and rules for progressing toward the point which represents the answer will be given.

The N dimensional space will be assumed to consist of N mutually perpendicular axes, and each of the unknowns x represent one of these axes. Thus a certain set of N numbers may be considered to be the coordinates of a point in N space. For convenience, this point will be called x_k if it corresponds to the set of coordinates represented by the k^{th} approximation.

The convenience of matrix algebra is well suited to this purpose. One need merely express a set of N linear equations in N unknowns as

$$Ax = y.$$

Here A represents the coefficients of the various x, and y the set of N constants on the right side of the equations. In addition to a point in N space, the set of numbers can be visualized as an N-dimensional vector, represented by the line joining the origin of the axes and the point in question.

One proceeds to describe a given problem in terms of quadratic forms, which can be geometrically interpreted as maps.

A. Maps, what they are, and how to make them

The problem to be solved is

$$Ax = y,$$

where A is an N^{th} order non-singular square matrix, x an unknown column matrix, and y a known column matrix. One desires that x which, when substituted in the equation, yields y. Specifically, an exact answer is one which yields y exactly. It is presumed that the y is known accurately, as are all the elements of the matrix A.

Now it is known that if x_k (the k^{th} approximation to x) is not x, then Ax_k will not be y. Thus, one can always determine whether or not x_k is the answer. One needs to know more, however; one needs to know how to go from x_k to x.

One defines the residual

$$r_k = Ax_k - y = Ax_k - Ax = A(x_k - x).$$

Thus if $e_k = x_k - x$, the error in the k^{th} approximation,

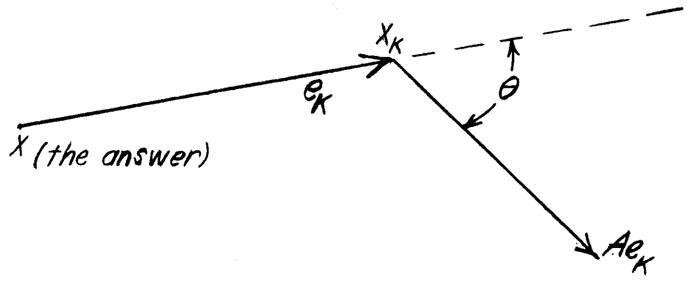
$$r_k = Ae_k.$$

Not only is it known that x_k is not the answer, but, since r_k is computable, one has a "weighted" measure of the error e_k.

Rather than consider the residual as a point in N space, one might visualize it as a direction. As such it might occur to the reader to use this direction to correct the approximation in x. Later some physical reasons for such a choice will be given, but for the moment it would appear that its only value is that it is a vector one has just computed and is related to the error vector e_k.

The relationship between the residual and the error can be pictured graphically. The premultiplication of the vector e_k by the matrix A has the general effect of rotating the direction of the vector and changing its length. The amount of this change cannot be predicted, unfortunately, for it depends on e_k which is not known.

Consider the following diagram in two dimensions:



If one is at x_k, and he wishes to go to x, then the direction to pursue is the negative e_k direction. He does not know what this is, but he does know the direction of Ae_k, which will, in general, be at an angle θ with e_k. θ will depend on where x_k is.

Evidently if θ is always less than 90°, motion in the negative $r_k(Ae_k)$ direction can bring him nearer to x if he does not move too far in this direction. θ will always be less than 90° if the dot product of e_k and Ae_k is always positive, i.e., if

$$e_{k_t} Ae_k > 0 \quad \text{for all} \quad e_k \neq 0.$$

This means that if $e_{k_t} Ae_k$ is a positive definite quadratic form, then x_{k+1}, a better approximation to x, can be obtained from x_k by a formula of the type

$x_{k+1} = x_k - m_k r_k$, where m_k is a constant so chosen that one does not move so far in the negative r_k direction that he gets farther away from the answer than he was before.

At this point an investigation of $e_{k_t} Ae_k$ where A is positive definite is undertaken. This expression is zero at the answer, and positive elsewhere. If one could compute this quadratic form, and at every step ensure that it becomes smaller, why he would have a measure of his nearness to x! This will be called a Map, or mapping function, for it is a device which, if properly used, will aid in the finding of x.

B. Definition

A Map (mapping function), M_i, will be considered to be a computable positive definite quadratic form in e_k, such that it will be zero when $e_k = 0$, and positive elsewhere. By computable is meant that the value of the quadratic form M_i can either be determined for a given x_k, or be an unknown constant which depends on the solution x plus a quantity which is determinable for any x_k. The contours given by M_i = constant in N space will cause the space to appear like a relief map, the lowest point of which is the answer.

It is not difficult to write down two expressions which satisfy these conditions:

$$M_1 = e_{k_t}Ae_k \qquad \text{A positive definite and symmetric.}$$

$$M_2 = e_{k_t}A_tAe_k \qquad \text{A non-singular.}$$

The restriction of symmetry on A in M_1 follows from the fact that M_1 is not computable if A is not symmetric.

$$M_1 = e_{k_t}Ae_k = (x_k - x)_t\, A\, (x_k - x)$$

$$= x_{k_t}\, Ax_k - x_tAx_k - x_{k_t}Ax + x_tAx.$$

x_tAx is a constant, call it C, and $Ax = y$, so

$$M_1 - C = x_{k_t}Ax_k - x_{k_t}y - x_{k_t}A_tx$$

where $x_{k_t}A_tx = x_tAx_k$.

Unless $A = A_t$, A_tx is not computable; hence, if A is symmetric:

$$M_1 - C = x_{k_t}Ax_k - 2x_{k_t}y.$$

$M_1 - C$ is evidently computable; hence, M_1 is usable and can be minimized.

M_2 is evidently computable, since $r_k = Ae_k$, so

$$M_2 = (Ae_k)_t(Ae_k) = r_{k_t}r_k.$$

A_tA is positive definite and symmetric; hence, M_2 is similar in form to M_1.

To prove the symmetry of A_tA:

$$(A_tA)_t = A_tA. \qquad \text{Q.E.D.}$$

To prove the positive definiteness:

Let $x' = Ax$. Then $x'_tx' = x_tA_tAx$ is a sum of squares, hence surely positive definite.

C. Investigation of the Maps M_1 and M_2

Since both maps involve symmetric positive definite matrices, one can solve both of these with one step by letting B, a positive definite symmetric matrix, represent A in M_1, and A_tA in M_2. (Note that the A in M_2 need not be either positive definite or symmetric.) Then the maps are of the general type

$$e_{k_t}Be_k.$$

Every symmetric matrix possesses an orthogonal modal matrix

$L = L_t^{-1}$ such that

$L_tBL = \Lambda$, where Λ is a diagonal matrix of the characteristic or latent roots of B. Since B is positive definite, all these roots will be positive.

If the substitution

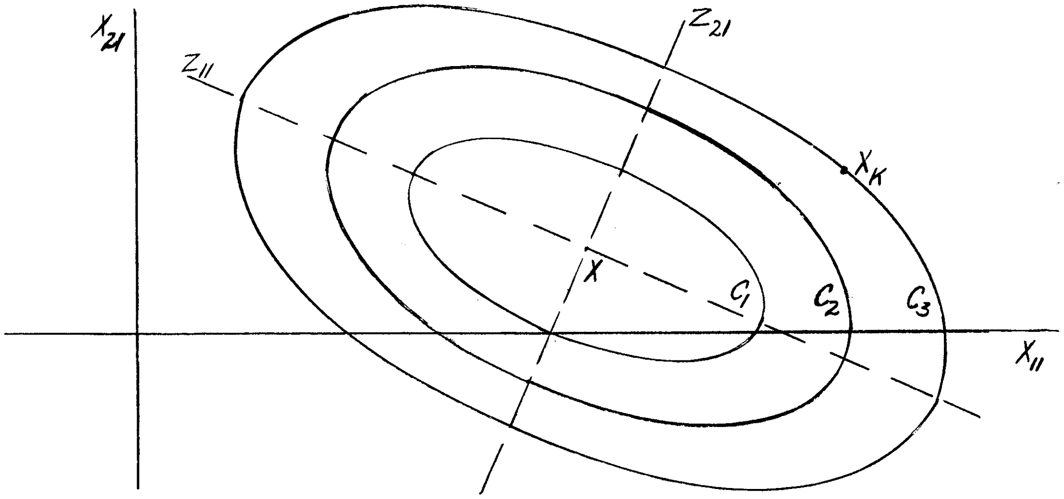
$e_k = Lz_k$, is made,

$$e_{k_t}Be_k = z_{k_t}L_tBLz_k = z_{k_t}\Lambda z_k.$$

$$e_{k_t}Be_k = \lambda_1z_{11}^2 + \lambda_2z_{21}^2 + \dots\dots\dots + \lambda_Nz_{nl}^2$$

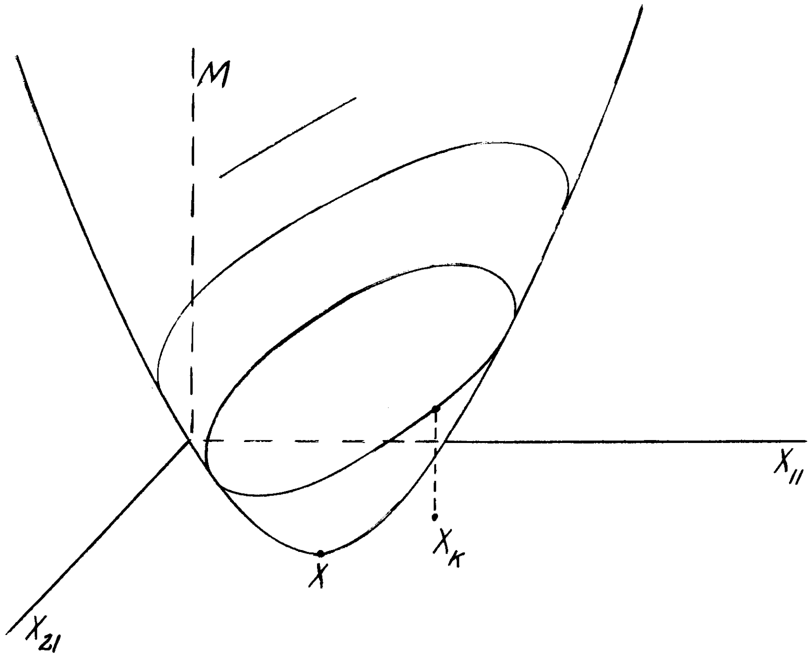
where $z_{k_t} = [z_{11} \;\; z_{21} \;\; \dots \;\; z_{Nl}]$ and $\lambda_j > 0$, $j = 1, 2, \dots, N$.

This is the equation of an elliptic quadric surface when $e_{k_t}Be_k$ = constant. For M_1 or M_2 equal to C_1, C_2, C_3, etc., one obtains a set of similar concentric ellipses of equal eccentricities. At the common center of these ellipses lies the answer to the problem. In two dimensions these contours appear as those in the figure below.



x is the answer, x_k may be the k^{th} approximation, and the contours for M = C_1, C_2, C_3 are as labelled, where $C_3 > C_2 > C_1$.

If one more dimension is added to the above picture and M is plotted vertically, a surface is obtained which is a paraboloid such as that pictured in the figure at the top of the next page.



An approximation x_k can now be visualized as a point on this surface. The method of Descent is concerned with moving down in this cavity toward the answer x which lies at the bottom.

D. <u>The Method of Descent</u>

or Let's get to the bottom of this!

The first idea in this connection was to choose the gradient of the map M and to move in the direction of the negative gradient a distance which minimizes M. Specifically, for M_1:

$$M_1 = e_{k_t} A e_k.$$

Grad $M_1 = 2r_k$. This follows immediately from the expansion of the quadratic form and performing the partial differentiation. x_{k+1} is then obtained from x_k by the formula

$x_{k+1} = x_k - m_k r_k$. Subtracting x from both sides gives

$$x_{k+1} - x = x_k - x - m_k\, r_k \qquad\qquad \text{or}$$

$$e_{k+1} = e_k - m_k\, r_k.$$

Thus $e_{{k+1}_t} A e_{k+1} = (e_k - m_k r_k)_t A\, (e_k - m_k r_k)$

$$= e_{k_t} A e_k - 2m_k r_{k_t} A e_k + m_k{}^2 r_{k_t} A r_k.$$

Differentiating twice with respect to m_k:

$$\frac{d(e_{{k+1}_t} A e_{k+1})}{dm_k} = -\,2\, r_{k_t} r_k + 2m_k r_{k_t} A r_k$$

Setting the first derivative equal to zero yields:

$$m_k = r_{k_t} r_k / r_{k_t} A r_k.$$

$$\frac{d^2(e_{{k+1}_t} A e_{k+1})}{dm_k{}^2} = 2m_k r_{k_t} A r_k > 0 \quad \text{if } A \text{ is positive definite,}$$

hence the quadratic form is truly minimized.

This procedure has been referred to as the Method of Steepest Descent, a name which is misleading since it implies that this is the best that can be done. Actually for a set of ill-conditioned equations the method is so slow as to be virtually useless. It is almost better to choose directions arbitrarily instead of using the r_k or residual as it is called. If this is done, one obtains a more general form of Descent. This is done as follows:

Choose any vector p_k. Using the formulae

$$x_{k+1} = x_k - m_k p_k$$

and $\qquad m_k = p_{k_t} r_k / p_{k_t} A p_k$ one has a procedure which is entirely general, i.e., the new approximation will be the one obtained which minimizes M_1 on the vector p_k emanating from x_k.

For M_2:
$$M_2 = e_kA_tAe_k$$

$$\text{Grad } M_2 = 2A_tr_k.$$

Here the gradient method or "steepest descent" uses the formulae

$$x_{k+1} = x_k - m_kA_tr_k$$

$$m_k = r_{k_t}AA\, r_k/r_{k_t}AA_tAA_tr_k.$$

If this method is generalized for any vector p_k, one obtains

$$x_{k+1} = x_k - m_kA_tp_k$$

$$m_k = p_{k_t}AA_tr_k/p_{k_t}AA_tAA_tp_k$$

or

$$x_{k+1} = x_k - m_kp_k$$

$$m_k = p_{k_t}A_tr_k/p_{k_t}A_tAp_k.$$

By suitable choice of the p_k these methods can be made to converge in N steps, and this is the work of Stiefel and Hestenes which is discussed in Chapter III.

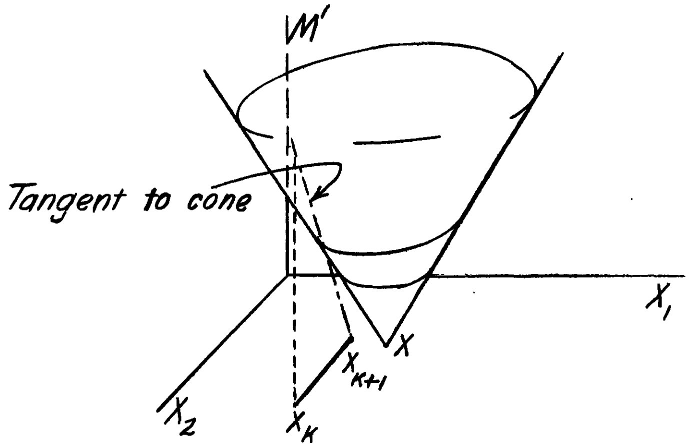
E. A minimized error procedure

Actually the considerations of the previous section can be summarized by stating that under certain restrictive conditions the best corrections to x seem to be obtained by minimizing the quadratic forms $e_{k_t}Ae_k = e_{k_t}r_k$, and $r_{k_t}r_k$. By this time the reader may wonder why the simplest procedure has not been investigated, namely minimizing the error e_k.

Stated in another way, when one is at x_k, and he chooses to move in some direction p_k to improve his answer, why not minimize his distance from the answer? This is a substantial part of the author's contribution which (he believes) has not been successfully accomplished before.

The reason this appears so difficult is that since one does not know where the answer is, how can one get as near as possible?

Consider the function $M' = (M)^{\frac{1}{2}} = \sqrt{e_{k_t} A_t A e_k} = |r_k|$. When, with two equations in two unknowns, one plots the surface represented by M' with the two unknowns x_1 and x_2 as the axes, he obtains a cone with elliptical cross-sections instead of the paraboloid on page 14. If at some point x_k on this surface, one erects a tangent to the surface whose projection on the x_1-x_2 plane is the gradient of M', then this tangent, if extended, intersects the x_1-x_2 plane at some point which will be called x_{k+1}. It is not obvious that this new x is better than the old, but it looks plausible in the figure below that such is the case.



When the mathematics of this are completed, (it is done just as the Newton-Raphson procedure in Appendix II) it appears that:

$$x_{k+1} = x_k - m_k A_t r_k$$

$$m_k = r_{k_t} r_k / r_{k_t} A A_t r_k.$$

It was not until several examples were done by the author that it was finally observed that the error in x is minimized by this procedure. To show this let

$$x_{k+1} = x_k - m_k A_t p_k$$

therefore $e_{k+1} = e_k - m_k A_t p_k$.

Then $e_{k+1_t} e_{k+1} = (e_k - m_k A_t p_k)_t (e_k - m_k A_t p_k)$

$$= e_{k_t} e_k - 2m_k p_{k_t} Ae_k + m_k^2 p_{k_t} AA_t p_k.$$

Differentiating twice as before with respect to m_k yields

$$\frac{d(e_{k+1_t} e_{k+1})}{dm_k} = -2p_{k_t} r_k + 2m_k p_k AA_t p_k$$

and $$\frac{d^2(e_{k+1_t} e_{k+1})}{dm_k^2} = 2p_{k_t} AA_t p_k > 0.$$

Setting the first derivative equal to zero and noting that this does minimize e_{k+1} since the second derivative is positive, one obtains

$$m_k = p_{k_t} r_k / p_{k_t} AA_t p_k. \qquad \text{Q.E.D.}$$

In particular if p_k is replaced by r_k these formulae reduce to the tangent gradient method discussed at the bottom of the previous page.

F. General

As nice as these procedures appear, none of them are very practical since even though convergence is assured, the rate of convergence is slow. This leads the reader to the N step procedures of the next chapter. The author would like to digress a moment to compare the minimized error and method of descent procedures. If they are compared on the basis of $M_2(e_{k_t} A_t Ae_k)$ and if the vectors, p_k, are chosen as the residuals, one obtains

$$x_{k+1} = x_k - m_k A_t r_k$$

1. Minimized error: $m_k = r_{k_t} r_k / r_{k_t} B r_k$ where $B = AA_t$.

2. Descent: $m_k = r_{k_t} B r_k / r_{k_t} B^2 r_k$

3. General: $m_k = r_{k_t} B^{n-1} r_k / r_{k_t} B^n r_k$.

The limit of the general expression as n becomes infinite can be shown to be the reciprocal of the largest characteristic number of B. In fact, all the m_k above lie in the range of the reciprocals of the smallest and largest characteristic numbers. This is important only since it can be shown that if the m_k are in turn the reciprocals of each of the characteristic numbers of B, the procedure will converge in N steps. Of course the characteristic numbers are not known, so this is not of much help practically, but some measure of the value of a procedure can be gained by noting how near the m_k comes to the reciprocal of a latent root of B. In any event it appears that the Method of Descent has a slight edge on the minimized error technique with regard to speed of convergence. The main difficulty with all these procedures is that the choice of directions is poor. The N step procedures which follow specify the directions to speed up the convergence.

CHAPTER III

N-STEP PROCEDURES

Chronologically Fox, Huskey, and Wilkinson[8] were the first to suggest an N-step procedure for the solution of N simultaneous equations. For reasons of practicality not much was done about this until the work of Stiefel and Hestenes[24] made the procedure workable.

In the meantime, Lanczos,[9] while attempting to find an iterative scheme for obtaining the characteristic polynomial of a matrix, found an orthogonalization scheme which all other procedures have adopted.

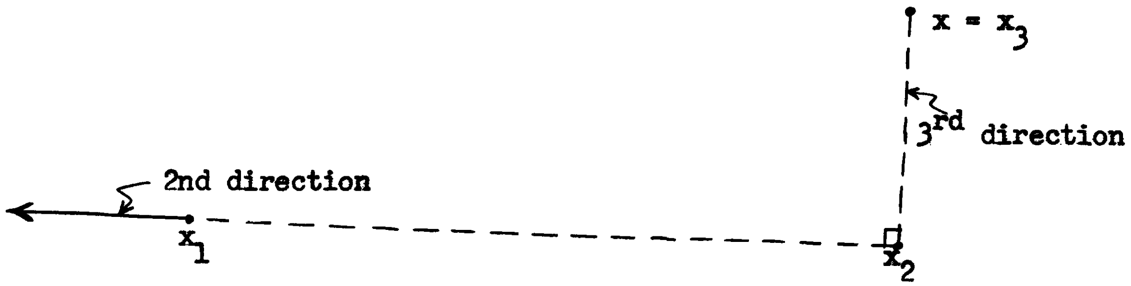
The object of this chapter is to describe the author's N-step procedure first, since it is believed to be the simplest conceptually. Descriptions of the work of the other men will then be given, and a general proof of all the methods appears in the next chapter.

1.0 N-Step Minimized Error Procedure

If one minimizes the error in x at each step of the iteration, then why not choose a mutually orthogonal set of directions for the steps, and on each minimize the error? Almost intuitively this procedure must converge in N steps!

Let the reader imagine that, in three-dimensional space, the answer to a problem lies in the plane of the paper, as in the figure on the next page. Imagine further that the first guess x_o is directly above x_1 in the figure, so that the direction from x_o to x_1 is perpendicular to the paper. Then, if one is at x_o, and the first direction is perpendicular to the plane of the paper, movement along this direction such that one gets as near to the answer as possible means stopping at x_1 on the paper.

If one now chooses a direction at x_1 perpendicular to $x_1 - x_o$, this direction must lie in the plane of the paper. One such direction is chosen, and x_2 is obtained by moving along the negative direction (since the wrong direction was chosen) until one is nearest to x. Obviously there remains but <u>one</u> line which is perpendicular to both of the previous directions, and this goes through the answer x. Movement along this direction until the error is minimized means arriving at the answer.



(x_o on the line prependicular to paper through x_1.)

There are two problems which now need to be settled. Any old set of orthogonal directions will not do. It will be recalled that to minimize the error by using vectors p_k one needed to know both p_k and $A_t p_k$. The equations are repeated below

$$x_{k+1} = x_k - m_k A_t p_k$$

$$m_k = r_{k_t} p_k / p_{k_t} A A_t p_k.$$

Notice that the correction vectors to the x are $A_t p_k$ and not p_k, hence one must find a set of vectors p_k such that the N vectors $A_t p_k$ form a mutually orthogonal set. p_k needs to be known, since it appears in the numerator of m_k.

Mathematically one may state that the p_k must have the following properties:

$$(A_tp_k)_t(A_tp_j) = 0 \qquad j \neq k$$

or

$$p_{k_t}AA_tp_j = 0 \qquad j \neq k.$$

There is a straightforward way in which this can be done. It is an adaptation of some of the work of Fox, Huskey, and Wilkinson and is really the Gram-Schmidt procedure.

a. Choose p_o arbitrarily.

b. Choose $b_1 \neq p_o$ and let

$$p_1 = b_1 - \alpha_o p_o.$$

Since $p_{1_t}AA_tp_o = 0$,

$$p_{1_t}AA_tp_o = b_{1_t}AA_tp_o - \alpha_o p_{o_t}AA_tp_o$$

or $\alpha_o = b_{1_t}AA_tp_o/p_{o_t}AA_tp_o$.

c. Choose b_k different from $p_o, p_1 \dots\dots p_{k-1}$ and

let $p_k = b_k - \alpha_{k-1}p_{k-1} - \beta_{k-2}p_{k-2} - \dots - \xi_o p_o$

choosing the constants so that A_tp_k is orthogonal to all previous A_tp_j.

Unfortunately this is a tremendous task, for at each step a piece of each of the previous directions must be removed. The solution of this dilemma is an adaptation from the procedure of Stiefel and Hestenes, which is discussed below. It appears that the following iterative scheme automatically orthogonalizes the directions with the advantage computationally that at each step only two constants need to be evaluated. Choosing the first residual as p_o, and each of the b_k above as the successive residuals, only the last direction needs to be removed from the b_k. This is not intended to be obvious, and the reasons for this will be made clearer as the thesis progresses.

In equation form the entire procedure is:

$$x_{k+1} = x_k - m_k A_t p_k$$

$$r_k = Ax - y$$

$$p_o = r_o$$

$$m_k = r_{k_t} r_k / p_{k_t} AA_t p_k$$

$$p_k = r_k + \varepsilon_{k-1} p_{k-1}$$

$$\varepsilon_{k-1} = r_{k_t} r_k / r_{k-1_t} r_{k-1}.$$

An additional result of this procedure, which will be proved in the following chapter, is the fact that the residuals form an orthogonal set as well as the directions. It is well to point out again that this procedure will work with any non-singular matrix.

2.0 The Method of Conjugate Directions*

Let $m_k p_k$ be the vector correction applied to x_k to obtain x_{k+1}, viz.

$$x_{k+1} = x_k - m_k p_k.$$

Choose the constant m_k so that the quadratic form $e_{k+1_t} A e_{k+1}$ is minimized, A being a symmetric, positive definite matrix. Thus

$$m_k = r_{k_t} p_k / p_{k_t} A p_k.$$

* This name is due to Stiefel and Hestenes, the procedure to Fox, Huskey, and Wilkinson.

Then if all the directions p_k, $k = 0, 1, 2, \ldots, N-1$ are "A-orthogonal" or "conjugate" (not to be confused with conjugate complex numbers), i.e., if

$$p_{k_t} A p_j = 0 \qquad j \neq k,$$

then x_N, the N^{th} approximation will be the solution x of the problem

$$Ax = y.$$

Suppose one digresses a moment to attempt a visualization of this procedure. One notes that the recursion formula for e_k can be obtained by subtracting x from both sides of the first equation:

$$x_{k+1} - x = x_k - x - m_k p_k$$

$$\text{or} \qquad e_{k+1} = e_k - m_k p_k.$$

The "A-orthogonality" of the p_k then suggests that the corrections to the error vector, $m_k p_k$, are orthogonal to all the vectors Ap_j where j is any number different from k. That is, if p_5 is the direction taken in going from e_5 to e_6, then this vector is perpendicular to Ap_o, Ap_1, ..., Ap_4, Ap_6, etc., all except Ap_5.

It is clear that the p_k form an independent set, and this can be seen in the following manner. Ap_k is orthogonal to all p_j except p_k. Since A is positive definite, Ap_k cannot be orthogonal to p_k, i.e., $p_{k_t} Ap_k \neq 0$. Therefore, p_k has a component of Ap_k which is orthogonal to all the other p_j, hence the p_k form an independent set. In an entirely similar way, one can show that the Ap_k are an independent set, and hence each set of vectors must span the entire N-space if $k = 0, 1, 2, 3, \ldots, N-1$.

Now let the reader suppose that an initial guess is made, thereby establishing e_o, though it is unknown. In general e_o will have a projection on each of the Ap_k, i.e., the dot product of e_o with the Ap_k will in general be non-zero. Suppose the successive corrections to e_o are such that these projections on the Ap_k are reduced to zero one at a time. Since the Ap_k span N space, after N such reductions the error must vanish for its projections on an independent set of vectors are all zero. Notice that in order to do this effectively, one must remove the projection of e_o on Ap_o (say) in such a way that the projections on all the other Ap_k do not change. If this is not done, one will have a relaxation procedure such as described in Chapter I. (With regard to the example of the mother putting the six children to bed, these are the more drastic means. She chains the children in bed one at a time!)

Starting at the beginning, one wishes to remove from e_o a vector such that e_1 is perpendicular to Ap_o, i.e., so that $e_{1_t}Ap_o = 0$. At the same time it is desired that $e_{1_t}Ap_k = e_{o_t}Ap_k$ for all $k \neq 0$, i.e., the new error vector has the same projections as the former on the N-1 vectors Ap_1, Ap_2, ..., Ap_{N-1}, but <u>no</u> projection on Ap_o.

Since

$$e_1 = e_o - m_o p_o$$

$$e_{1_t}Ap_k = e_{o_t}Ap_k - m_o p_{o_t}Ap_k.$$

Evidently the above conditions will be satisfied if

$$p_{o_t}Ap_k = 0 \qquad k \neq 0,$$

and for k = 0: $$e_{1_t}Ap_o = 0 = e_{o_t}Ap_o - m_o p_{o_t}Ap_o,$$

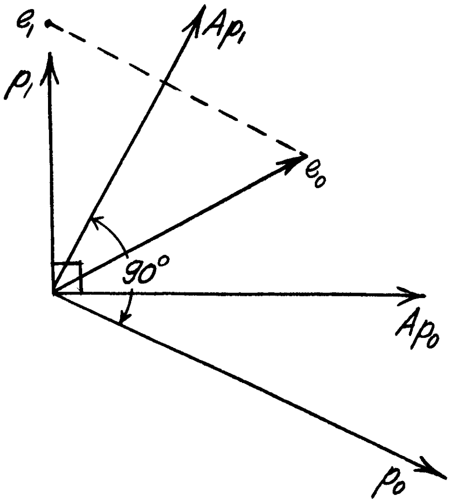
$$\text{or } m_o = e_{o_t}Ap_o/p_{o_t}Ap_o.$$

But $e_{o_t}A = (Ae_o)_t = r_{o_t}$, so

$$m_o = r_{o_t}p_o/p_{o_t}Ap_o.$$

But this is the same m_o one obtains by minimizing the quadratic form $e_{1_t}Ae_1$, and so the procedure above stated does indeed remove the projections of the e vectors on each of the Ap_k one at a time, without destroying the other projections in the process.

This is the same thing as the minimized error technique, with the exception that in the latter procedure one minimized the projections of the error vector on A_tp_k, and that the A_tp_k formed an orthogonal set. In the method of conjugate directions the Ap_k did not form an orthogonal set, and A furthermore had to be symmetric and positive definite. A two-dimensional picture of this is thought to be helpful.



It is presumed that p_o, and p_1, Ap_o and Ap_1 are known, and the reader will note that p_o is perpendicular to Ap_1, p_1 perpendicular to Ap_o. Notice that as p_o is subtracted from e_o, the projection of the result on Ap_1 does not change since p_o is normal to Ap_1. Clearly, if one moves along the dashed line parallel to p_o from e_o until the resultant is normal to Ap_o, then the answer will be obtained in two steps.

The author submits that the name of this procedure would be more appropriate if it were called a minimized projection procedure. Since the author did not find this procedure first, why there is little he can do but suggest.

For the reader who prefers to see a somewhat more rigorous demonstration of convergence the author submits the following proof. The result, i.e., convergence in N steps, follows from two considerations, one being the fact that p_o, p_1, ... p_{N-1} form an independent set of vectors, and the other being the fact that the N^{th} residual is normal to each of these vectors, hence it must be zero.

a. The p_j form an independent set.

Proof: By contradiction.

Assume there is at least one p_j, call it p_q, which is a linear combination of all the rest. That is, if the c_j are constants:

$$p_q = c_o p_o + c_1 p_1 + \cdots\cdot + c_{q-1} p_{q-1} + c_{q+1} p_{q+1} + \cdot\cdot + c_{N-1} p_{N-1}.$$

At least one c_j must be non-zero. For this particular c_j, premultiply the above equation by $p_{j_t}A$, where p_j is the vector associated with c_j. For all $j \neq k$, $p_{j_t}Ap_k = 0$ by hypothesis, then all terms will vanish except

$$c_j p_{j_t} Ap_j \quad \text{which must be zero.}$$

But this is impossible since c_j is non-zero, and A positive definite. Hence the contradiction, and thus the p_j do form an independent set.

b. r_N is orthogonal to all p_j.

Proof: Let the residual r_k be defined as

$$r_k = Ax_k - y. \tag{1}$$

Remark: Since A is non-singular, $r_k = 0$ implies that $x_k = x$.

Further

$$x_k = x_{k-1} - m_{k-1}p_{k-1} \tag{2}$$

where, from page 15, Chapter II

$$m_{k-1} = r_{k-1_t}p_{k-1}/p_{k-1_t}Ap_{k-1} \tag{3}$$

Premultiplying equation (2) by A, and subtracting y from both sides:

$$Ax_k - y = Ax_{k-1} - y - m_{k-1}Ap_{k-1}$$

or

$$r_k = r_{k-1} - m_{k-1}Ap_{k-1} \tag{4}$$

Transposing this, and postmultiplying by p_{k-1} and substituting the value of m_{k-1} given in (3) above:

$$r_{k_t}p_{k-1} = r_{k-1_t}p_{k-1} - \frac{r_{k-1_t}p_{k-1}}{p_{k-1_t}Ap_{k-1}}\, p_{k-1_t}Ap_{k-1} = 0. \tag{5}$$

From equation (4), postmultiplying its transpose by p_j:

$$r_{k_t}p_j = r_{k-1_t}p_j - m_{k-1}p_{k-1_t}Ap_j \tag{6}$$

If $j \neq k-1$, the coefficient of m_{k-1} vanishes, hence

$$r_{k_t}p_j = r_{k-1}p_j \qquad j \neq k-1. \tag{7}$$

Equation (7) is true, therefore, for $j = k-2$, so

$$r_{k_t}p_{k-2} = r_{k-1_t}p_{k-2} \tag{8}$$

But equation (5) states that the right side of (8) vanishes, hence

$$r_{k_t}p_{k-2} = 0. \tag{9}$$

Substituting $j = k-3$ in equation (7) then yields

$$r_{k_t}p_{k-3} = 0 \text{ and so forth.} \tag{10}$$

In particular, for k = N,

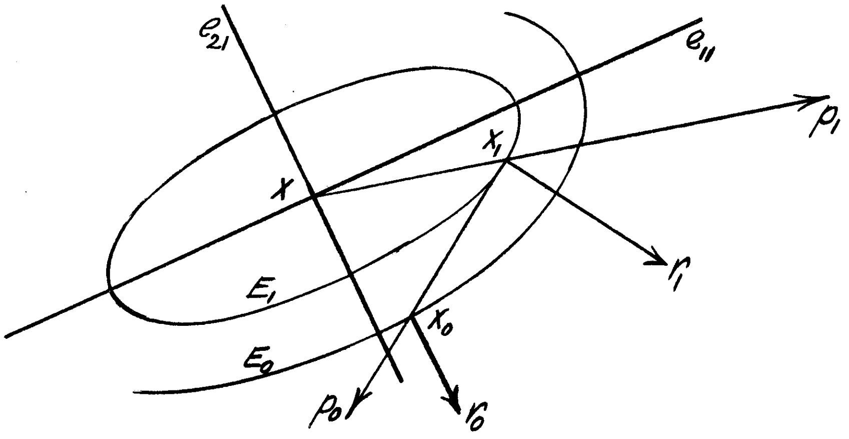
$$r_{N_t}p_j = 0 \quad \text{for all } j = 0, 1, 2, \ldots., N-1 \qquad \text{Q.E.D.}$$

Hence, as was asserted $r_N = 0$, and the procedure converges in N steps.

Another geometric picture is somewhat more difficult to obtain. In two dimensions, however, a picture can be constructed. It will be remembered that the map representing a positive definite quadratic form is a set of concentric ellipses and that the r_k represents the gradient of this map at the point x_k. Equation (5) indicates that $r_{k_t}p_{k-1} = 0$, that is, the m_{k-1} is so chosen that the last direction taken, p_{k-1}, is perpendicular to the gradient at the new approximation. In two dimensions, if the solution is to be obtained in two steps, the next direction must point to (or directly away from) the center of the ellipses.

Now if p_{k-1} is normal to the gradient, it must be <u>tangent</u> to the particular ellipse which contains x_k, and at the point x_k. The new direction p_k must be the radius vector from the center of the ellipses to x_k.

For example, consider the diagram below in two dimensions:



The initial guess is x_o which is on ellipse E_o. r_o represents the gradient at x_o, which is normal to the curve at x_o. p_o is arbitrarily chosen and x_1 is obtained by moving in the negative p_o direction a distance which minimizes the quadratic form. p_o is tangent to the ellipse E_1 at x_1, and the gradient at x_1, r_1, is normal to p_o. To obtain the solution x in one more step, it is necessary to follow the direction of the negative radius vector, here indicated by p_1.

The question arises, how does one find p_1? What is the relationship between the radius vector of an ellipse and the tangent? It turns out that

$$p_{1t}Ap_o = 0.$$

The simplest way to demonstrate this is to reduce the ellipse to its normal coordinates. As indicated in Appendix I this is always possible. As a matter of interest, all procedures are better understood when viewed in terms of their normal coordinates, hence the author will digress a minute to make this clearer.

If A is a symmetric matrix, then there exists an orthogonal modal matrix L such that A is reduced to a diagonal matrix, Λ, of its latent roots by

$$L_tAL = \Lambda .$$

This can be visualized as a rotation of axes as in analytic geometry, where the variables x_{k1} are replaced by x'_{k1} by the transformation

$$x = Lx'.$$

Thus $$r_k = Lr_k', \; p_k = Lp_k', \; e_k = Le_k', \text{ etc.}$$

The quadratic form $e_{kt}Ae_k$ written in the new variables becomes $e_{kt}'L_tALe_k' = e_{kt}'\Lambda e_k'$. If λ_1, λ_2, ... , λ_N are the latent roots of A, and e_1, e_2, etc., are the elements of the vector e_k', then the equation of a particular contour ellipse is

$$\lambda_1 e_1^2 + \lambda_2 e_2^2 = c.$$

The slope of the tangent of this ellipse at any point (e_1, e_2) is given by $de_2/de_1 = - \lambda_1 e_1/ \lambda_2 e_2$, that is, it has the direction of

$$p_o' = \begin{bmatrix} \lambda_2 e_2 \\ -\lambda_1 e_1 \end{bmatrix} .$$

The radius vector has the direction $p_1' = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$.

Note that

$$\begin{aligned} {p_1'}_t \Lambda p_o' &= \begin{bmatrix} e_1 & e_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \lambda_2 e_2 \\ -\lambda_1 e_1 \end{bmatrix} \\ &= \begin{bmatrix} e_1 & e_2 \end{bmatrix} \begin{bmatrix} \lambda_1 \lambda_2 e_2 \\ -\lambda_1 \lambda_2 e_1 \end{bmatrix} = \lambda_1 \lambda_2 e_1 e_2 - \lambda_1 \lambda_2 e_1 e_2 \\ &= 0. \end{aligned}$$

Note further that

$${p_1'}_t \Lambda p_o' = p_{1_t} A p_0.$$

It remains to be shown how one might obtain a set of conjugate directions. Fortunately this is quite simple and is explained in the work of Fox, Huskey, and Wilkinson and is entirely analogous to the previously mentioned orthogonalization scheme.

1. Choose p_o arbitrarily.
2. Choose $b_1 \neq p_o$ and let

$$p_1 = b_1 - \alpha_o p_o.$$

Since $p_{1_t}Ap_o = 0$,

$$p_{1_t}Ap_o = b_{1_t}Ap_o - \alpha_o p_{o_t}Ap_o = 0$$

or $$\alpha_o = b_{i_t}Ap_o/p_{o_t}Ap_o$$

3. Choose b_k different from p_o, p_1, ..., p_{k-1} and let

$$p_k = b_k - \alpha_{k-1}p_{k-1} - \beta_{k-2}p_{k-2} - \cdots\cdots - \xi_o p_o$$

choosing the constants so that p_k is conjugate to all previous p_j.

As previously pointed out, this scheme has serious defects.

3.0 Stiefel and Hestenes' Conjugate Gradients Procedure

At this point, independently, Stiefel and Hestenes made this procedure workable. Essentially they found a simple way to make the p_j conjugate. For reasons which they explain best, they call this procedure the Method of Conjugate Gradients.

In a nutshell the procedure is this:

$$x_{k+1} = x_k - m_k p_k$$

$$r_k = Ax_k - y$$

$$p_o = r_o$$

$$m_k = r_{k_t}r_k/p_{k_t}Ap_k$$

$$p_k = r_k + \varepsilon_{k-1}p_{k-1}$$

$$\varepsilon_{k-1} = r_{k_t}r_k/r_{k-1_t}r_{k-1}.$$

The reader will note that at each step two constants are evaluated, as in the minimized error procedure. The above sequence automatically ensures the mutual orthogonality of the r_k, as well as the A-orthogonality of the p_k. Again, the proof is deferred until Chapter IV.

Stiefel and Hestenes extend their procedure to any non-singular matrix by solving the new problem

$$A_t Ax = A_t y.$$

This is not the same as the author's procedure, for the formulae become:

$$\begin{aligned} x_{k+1} &= x_k - m_k A_t p_k \\ r_k &= Ax_k - y \\ p_o &= r_o \\ m_k &= r_{k_t} AA_t r_k / p_{k_t} AA_t AA_t p_k \\ p_k &= r_k + \varepsilon_{k-1} p_{k-1} \\ \varepsilon_{k-1} &= r_{k_t} AA_t r_k / r_{k-1_t} AA_t r_{k-1}. \end{aligned}$$

It is evident that the two are not equivalent, and that the latter extension by Stiefel and Hestenes looks more difficult. Which is actually the more complicated is a question for experiment to settle.

4.0 Lanczos Procedure

Lanczos was concerned with obtaining the solution to the eigenvalue problem, viz.

$$Ax = \lambda x.$$

His work will not be discussed in detail here, only that which is pertinent. He constructs a set of mutually orthogonal vectors in the following manner.

Starting with an initial vector r_o, which is arbitrary, he chooses

$$r_1 = Ar_o - \alpha_o r_o$$

where A is a symmetric matrix and α_o so chosen that $r_{1_t}r_o = 0$, hence

$$r_{1_t}r_o = 0 = r_{o_t}Ar_o - \alpha_o r_{o_t}r_o$$

$$\text{or } \alpha_o = r_{o_t}Ar_o/r_{o_t}r_o.$$

Then

$$r_2 = Ar_1 - \alpha_1 r_1 - \beta_o r_o \qquad (2)$$

choosing α_1 so that $r_{2_t}r_1 = 0$ and β_o so that $r_{2_t}r_o = 0$.

$$r_{2_t}r_1 = 0 = r_{1_t}Ar_1 - \alpha_1 r_{1_t}r_1 - \beta_o r_{o_t}r_1.$$

The coefficient of β_o is zero by choice of α_o, so

$$\alpha_1 = r_{1_t}Ar_1/r_{1_t}r_1.$$

Similarly

$$\beta_o = r_{1_t}Ar_o/r_{o_t}r_o.$$

Further

$$r_3 = Ar_2 - \alpha_2 r_2 - \beta_1 r_1 - \gamma_o r_o \qquad (3)$$

where

$$\alpha_2 = r_{2_t}Ar_2/r_{2_t}r_2$$

$$\beta_1 = r_{2_t}Ar_1/r_{1_t}r_1$$

and surprisingly enough

$$\gamma_o = 0.$$

The reason for this is based on the symmetry of A. It means that r_3 contains no component of r_o. Since r_2 and r_1 contain no component of r_o, this must mean from equation (3) that Ar_2 contains no component of r_o, or that

$$r_{o_t}Ar_2 = r_{2_t}Ar_o = 0.$$

But Ar_o from equation (1) is

$$Ar_o = r_1 + \alpha_o r_o \quad \text{so}$$

$$r_{2_t}Ar_o = r_{2_t}r_1 + \alpha_o r_{2_t}r_o.$$ This is zero since r_2 and r_1 and r_2 and r_o were made orthogonal.

In fact the general scheme

$$r_k = Ar_{k-1} - \alpha_{k-1}r_{k-1} - \beta_{k-2}r_{k-2} \tag{4}$$

where $$\alpha_{k-1} = r_{k-1_t}Ar_{k-1}/r_{k-1_t}r_{k-1} \tag{5}$$

and $$\beta_{k-2} = r_{k-1_t}Ar_{k-2}/r_{k-2_t}r_{k-2} \tag{6}$$

is sufficient to ensure that all r_k form a mutually orthogonal set. This is proved in the next chapter.

The point that the author wishes to emphasize is that two constants only need to be evaluated at each step. This is so similar to the method of the author and that of Stiefel and Hestenes that it cannot be coincidental. It is the object of Chapter IV to tie these procedures together.

However, to continue with Lanczos' procedure,

$$r_o$$

$$r_1 = Ar_o - \alpha_o r_o = (A - \alpha_o I)r_o$$

$$r_2 = Ar_1 - \alpha_1 r_1 - \beta_o r_o = A(A - \alpha_o I)r_o - \alpha_1(A - \alpha_o I)r_o - \beta_o r_o$$

$$= \left[A^2 - (\alpha_1 + \alpha_o)A + (\alpha_1\alpha_o - \beta_o)I\right] r_o$$

and similarly it can be shown that

$$r_3 = \left[A^3 - (\alpha_2+\alpha_1+\alpha_o)A^2 + (\alpha_1\alpha_o+\alpha_2\alpha_1+\alpha_2\alpha_o-\beta_1-\beta_o)A \right.$$
$$\left. -(\alpha_2\alpha_1\alpha_o-\alpha_2\beta_o-\beta_1\alpha_o)I\right] r_o.$$

Evidently the successive r_k are equal to the product of a polynomial in A times r_o. Since the r_k form an orthogonal set, then if A is of order N, $r_N = 0$, since in N-dimensional space r_N is perpendicular to N mutually perpendicular vectors. In general then,

$$r_k = P_k(A)r_o, \text{ where } P_k \text{ is a polynomial in A of the } k^{th} \text{ degree.}$$

It should be pointed out that r_M might vanish where $M < N$. This happens, as will be seen, if r_o does not contain all the eigenvectors of A.

The Cayley-Hamilton Theorem, it will be recalled, states that a matrix satisfies its own characteristic equation, hence if $P_N(A)$ is the characteristic equation of A, $P_N(A) = 0$, and the above equation is satisfied for $k = N$. It does not follow, however, that if the equation is satisfied, $P_N(A)$ is the characteristic equation.

It can be proved, however, that if the procedure terminates at the M^{th} step, the M^{th} degree polynomial in A is a factor of the characteristic equation and is the characteristic equation if $M = N$.

Proof: Let $v_1, v_2, \ldots\ldots, v_N$ be the mutually orthogonal eigenvectors of A. That these vectors are orthogonal and do span N-space can be shown. See for example Guillemin[11], page 141.

Then, in general, r_o will be a linear combination of M of these, $M \leqslant N$.

$$r_o = c_1v_1 + c_2v_2 + \ldots\ldots\ldots + c_Mv_M.$$

Remark: No linear combination of the v_j can possibly vanish since the vectors are orthogonal.

It is now noted that

$$Av_i = \lambda_i v_i$$

$$A^n v_i = \lambda^n_i v_i$$

and $P_M(A)v_i = P_M(\lambda_i)v_i$.

Hence $P_M(A)r_o = c_1P_M(\lambda_1)v_1 + \ldots\ldots + c_MP_M(\lambda_M)v_M$.

If $P_M(A)r_o = 0$, then this can only occur if all coefficients of the v_M vanish, i.e.,

$$P_M(\lambda_1) = P_M(\lambda_2) = \ldots\ldots = P_M(\lambda_M) = 0.$$

Then $P_M(\lambda)$ contains as many factors of the characteristic equation as r_o contains eigenvectors of A. It is still conceivable that $P_M(\lambda)$ has extraneous factors. It does not actually. If r_o is composed of M eigenvectors, then r_M is a linear combination of the same M eigenvectors. It is also normal to M other linear combinations of the same M eigenvectors, which is impossible. Therefore $r_M = 0$, and so the procedure will converge in exactly as many steps as r_o has components which are eigenvectors of A. Hence $P_M(\lambda)$ is a factor of the characteristic equation of A.

It will be pointed out subsequently that the characteristic equation can also be obtained from Stiefel's method,or the minimized error technique. Further study should be made of Lanczos' work if the reader wishes amplification of any points about this procedure, or if the reader is interested in studying the procedure used when A is not symmetric.

CHAPTER IV

PROOF OF THE N-STEP PROCEDURES

Three methods have now been presented which establish an orthogonal set of vectors in N-steps, and at each step only two constants are evaluated. It is reasonable to suppose that these procedures are more than similar, that they are probably one and the same. Such is indeed the case, and it is the object of this chapter to prove the orthogonalization scheme of Lanczos and to show that the procedure of Stiefel and Hestenes is a clever adpatation of this. The method of the author which was at first thought quite different from the others turns out to be but a simple extension of the procedure of Stiefel and Hestenes.

A. Theorem

Given any symmetric matrix A and non-zero vector r_o, then the r_k defined by the iterative scheme

$$c_k r_k = Ar_{k-1} - \alpha_{k-1} r_{k-1} - \beta_{k-2} r_{k-2} \tag{1}$$

form an orthogonal set if

1. c_k is any constant $\neq 0$.

2. $\alpha_{k-1} = \dfrac{r_{k-1_t} A r_{k-1}}{r_{k-1_t} r_{k-1}} \qquad \alpha_{-1} = 0.$

3. $\beta_{k-2} = \dfrac{r_{k-1_t} A r_{k-2}}{r_{k-2_t} r_{k-2}} \qquad \beta_{-2} = \beta_{-1} = 0.$

Proof: Given

$$c_k r_k = Ar_{k-1} - \alpha_{k-1} r_{k-1} - \beta_{k-2} r_{k-2}$$

Choose α_{k-1} such that $r_{k_t} r_{k-1} = 0.$

Premultiplying (1) by r_{k-1_t} one obtains

$$c_k r_{k-1_t} r_k = 0 = r_{k-1_t} A r_{k-1} - \alpha_{k-1} r_{k-1_t} r_{k-1} - \beta_{k-2} r_{k-1_t} r_{k-2}$$

The coefficient of β_{k-2} is zero since one presumes α_{k-2} was chosen so that $r_{k-1_t} r_{k-2} = 0$, hence

$$\alpha_{k-1} = \frac{r_{k-1_t} A r_{k-1}}{r_{k-1_t} r_{k-1}} .$$

Choose β_{k-2} such that $r_{k_t} r_{k-2} = 0$.

Premultiply (1) by r_{k-2_t} obtaining

$$c_k r_{k-2_t} r_k = 0 = r_{k-2_t} A r_{k-1} - \alpha_{k-1} r_{k-2_t} r_{k-1} - \beta_{k-2_t} r_{k-2}$$

The coefficient of α_{k-1} is zero by choice of α_{k-2}, hence

$$\beta_{k-2} = \frac{r_{k-2_t} A r_{k-1}}{r_{k-2_t} r_{k-2}} = \frac{r_{k-1_t} A r_{k-2}}{r_{k-2_t} r_{k-2}}$$

since the numerator is a scalar and A symmetric.

It has now been proved that r_k is normal to r_{k-1} and r_{k-2}. It is not evident that r_k is normal to r_j for all $j = 0, 1, 2, \ldots, k-3$. The proof is by induction.

a. The statement is true for $k = 1$ and 2. This is true since α_o is chosen so that $r_{1_t} r_o = 0$, and α_1 and β_o chosen so that $r_{2_t} r_1 = 0$ and $r_{2_t} r_o = 0$ respectively.

b. Assume that it is true for $k \leq q$, that is

$r_{k_t} r_j = 0$ for all $k \leqslant q$, and all $j < k$, hence in particular, assume

$$r_{q_t} r_j = 0 \text{ for all } j < q.$$

From equation (1) one has

$$c_j r_j = Ar_{j-1} - \alpha_{j-1} r_{j-1} - \beta_{j-2} r_{j-2}$$

so $$c_j r_{q_t} r_j = 0 = r_{q_t} A r_{j-1} - \alpha_{j-1} r_{q_t} r_{j-1} - \beta_{j-2} r_{q_t} r_{j-2}.$$

But the coefficients of α_{j-1} and β_{j-2} are zero by hypothesis, hence

$$r_{q_t} A r_{j-1} = r_{j-1_t} A r_q = 0 \qquad j < q \tag{2}$$

Again using equation (1):

$$c_{q+1} r_{q+1} = Ar_q - \alpha_q r_q - \beta_{q-1} r_{q-1}$$

Premultiplying by r_{j-1_t} one obtains

$$c_{q+1} r_{j-1_t} r_{q+1} = r_{j-1_t} A r_q - \alpha_q r_{j-1_t} r_q - \beta_{q-1} r_{j-1_t} r_{q-1}.$$

But again, the coefficients of α_q and β_{q-1} vanish by hypothesis, so

$$c_{q+1} r_{j-1_t} r_{q+1} = r_{j-1_t} A r_q. \qquad j < q \tag{3}$$

From equation (2) the right side of (3) vanishes if $j<q$, so

$$r_{q+1_t} r_{j-1_t} = 0 \text{ for } j = 0, 1, 2, \ldots., q-1.$$

Since α_q and β_{q-1} were chosen so that

$$r_{q+1_t} r_q = 0 \quad \text{and} \quad r_{q+1_t} r_{q-1} = 0$$

then it is true therefore that

$$r_{q+1_t} r_j = 0 \quad \text{for} \quad j < q+1 \qquad \text{Q.E.D.}$$

B. The N-step procedures and the Lanczos scheme

It is the task of this section to show that the author's procedure and that of Stiefel and Hestenes are equivalent to each other and are in fact related to the orthogonalization scheme of the previous section.

The iterative formula for x_k in both instances is

$$x_{k+1} = x_k - m_k A_t p_k$$

Thus if $$Ax_{k+1} - y = Ax_k - y - m_k AA_t p_k$$

then $$r_{k+1} = r_k - m_k AA_t p_k. \tag{4}$$

In the conjugate gradient method the formula is the same as this but with A only and not AA_t. In both cases AA_t and A were symmetric, so with the reader's permission, equation (4) will be written simply as

$$r_{k+1} = r_k - m_k Bp_k \tag{5}$$

where B will signify A for the procedure of Stiefel and Hestenes, and it will mean AA_t for the author's procedure.

One more equation is used in each of these procedures, namely

$$p_k = r_k + \varepsilon_{k-1} p_{k-1}. \tag{6}$$

Note that equation (5) can be written

$$-\frac{1}{m_k} r_{k+1} = Bp_k - \frac{1}{m_k} r_k \tag{7}$$

If one replaces k by k-1 in equation (5) and solves for Bp_{k-1}:

$$Bp_{k-1} = \frac{1}{m_{k-1}} (r_{k-1} - r_k) \tag{8}$$

If equation (6) is substituted in (7) for p_k, one obtains

$$-\frac{1}{m_k} r_{k+1} = Br_k + \varepsilon_{k-1} Bp_{k-1} - \frac{1}{m_k} r_k \tag{9}$$

One can now use (8) to eliminate Bp_{k-1} from (9) yielding

$$-\frac{1}{m_k} r_{k+1} = Br_k - \left(\frac{1}{m_k} + \frac{\varepsilon_{k-1}}{m_{k-1}} \right) r_k + \frac{\varepsilon_{k-1}}{m_{k-1}} r_{k-1} \tag{10}$$

Equation (10) is the same as equation (1) with k replaced by k+1 if

$$c_{k+1} = - \frac{1}{m_k}$$

$$\alpha_k = \frac{1}{m_k} + \frac{k-1}{m_{k-1}} = \frac{r_{k_t}Br_k}{r_{k_t}r_k} \tag{11}$$

and $$\beta_{k-1} = - \frac{\varepsilon_{k-1}}{m_{k-1}} = \frac{r_{k_t}Br_{k-1}}{r_{k-1_t}r_{k-1}} \, . \tag{12}$$

Therefore the r_k will form an orthogonal set if m_k and ε_{k-1} are so chosen that the above relations are satisfied.

Adding (12) to (11) gives

$$\frac{1}{m_k} = \frac{r_{k_t}Br_k}{r_{k_t}r_k} + \frac{r_{k_t}Br_{k-1}}{r_{k-1_t}r_{k-1}} \tag{13}$$

Replacing k by k-1 and substituting in (12)

$$\varepsilon_{k-1} = - \frac{r_{k_t}Br_{k-1}}{r_{k-1_t}r_{k-1}} \cdot \frac{1}{\left(\frac{r_{k-1_t}Br_{k-1}}{r_{k-1_t}r_{k-1}} + \frac{r_{k-1_t}Br_{k-2}}{r_{k-2_t}r_{k-2}}\right)} \tag{14}$$

Evidently this is always possible even though the expressions for the constants are messy. If one chooses the m_k and ε_{k-1} according to (13) and (14) then the r_j do form an orthogonal set. Knowing this it is possible to simplify the expressions for these constants.

From equation (5)

$$r_{k+1} = r_k - m_kBp_k .$$

Premultiplying by r_{j_t}, $j \neq k$ or $k+1$ one obtains

$$r_{j_t}Bp_k = 0 \qquad j \neq k \text{ or } k+1 \tag{15}$$

Since $p_k = r_k + \varepsilon_{k-1}p_{k-1}$

$$r_{j-1_t}Bp_k = r_{j-1_t}Br_k \qquad j \neq k \text{ or } k+1 \tag{16}$$

Also since $r_{{k+1}_t} r_k = 0$

$$r_{{k+1}_t} r_k = 0 = r_{k_t} r_k - m_k r_{k_t} B p_k$$

$$\text{or} \qquad m_k = \frac{r_{k_t} r_k}{r_{k_t} B p_k} \tag{17}$$

and
$$r_{{k+1}_t} r_{k+1} = r_{{k+1}_t} r_k - m_k r_{{k+1}_t} B p_k$$

$$\text{or} \qquad m_k = \frac{-r_{{k+1}_t} r_{k+1}}{r_{{k+1}_t} B p_k} \tag{18}$$

From (16), with j = k+2

$$r_{{k+1}_t} B p_k = r_{{k+1}_t} B r_k$$

$$\text{so} \qquad m_k = -\frac{r_{{k+1}_t} r_{k+1}}{r_{{k+1}_t} B r_k} \tag{19}$$

Replacing k+1 by k in (19) and substituting in (12) yields

$$\varepsilon_{k-1} = \frac{r_{k_t} r_k}{r_{{k-1}_r} r_{k-1}} \tag{20}$$

That equations (20) and (14) are equivalent can be demonstrated, but it is not of particular interest. Equation (17) can be altered further by noting that the vectors p_j are conjugate or B-orthogonal. That is

$$p_{k_t} B p_j = 0 \qquad j \neq k \tag{21}$$

This can be proved by simply noting that

$$p_j = r_j + \varepsilon_{j-1} p_{j-1} = r_j + \varepsilon_{j-1} r_{j-1} + \varepsilon_{j-1} \cdot \varepsilon_{j-2} r_{j-2} + \cdots + (\varepsilon_{j-1}\, \varepsilon_{j-2} \cdot\cdot\, \varepsilon_0) r_0$$

Therefore $$r_{{k+1}_t} p_j = 0 \qquad \text{for} \quad j < k+1 \tag{22}$$

But $$r_{k+1} = r_k - m_k B p_k$$

so $$r_{{k+1}_t} p_j = r_{k_t} p_j - m_k p_{k_t} B p_j \tag{23}$$

The left side is zero for $k+1 > j$, and the first term on the right is zero for $k > j$, hence both vanish for $k > j$.

Therefore $$p_{k_t} B p_j = 0 \qquad k > j. \qquad \text{Q.E.D.}$$

Since $$p_k = r_k + \varepsilon_{k-1} p_{k-1}$$

$$p_{k_t} B p_j = p_{k_t} B r_k \,. \tag{24}$$

If (24) is used in equation (17)

$$m_k = \frac{r_{k_t} r_k}{p_{k_t} B p_k}. \tag{25}$$

C. <u>Résumé</u>

It has now been demonstrated that any method employing the equations

$$r_{k+1} = r_k - m_k B p_k$$

$$p_k = r_k + \varepsilon_{k-1} p_{k-1}$$

$$m_k = \frac{r_{k_t} r_k}{p_{k_t} B p_k}$$

$$\varepsilon_{k-1} = \frac{r_{k_t} r_k}{r_{{k-1}_t} r_{k-1}}$$

where B is symmetric will converge in N steps. The requirement of positive definiteness seems unnecessary in the light of the foregoing proof.

The disadvantage when B is not definite is that the m_k may become large, even infinite. Hence the observation that B should be positive definite is a practical consideration rather than theoretical.

It is now seen that all three methods are fundamentally the same. It is well to point out that in the author's method

$$p_{k_t}Bp_j = p_{k_t}AA_tp_j = (A_tp_k)_t(A_tp_j) = 0 \qquad j \neq k.$$

Hence the vectors which represent the corrections in the x_k do - as previously asserted - constitute an orthogonal set.

CHAPTER V

VARIATIONS OF THE THEME, ANOTHER PROCEDURE

By this time it should be apparent to the reader that variations on this theme should not be hard to produce. It is curious how geometric reasoning leads to a mathematical formulation. Once the latter is achieved, it is child's play to manipulate the equations to obtain new procedures. These new procedures may be of questionable value, and then again the variations might be those which make roundoff error smaller.

To illustrate the point let us write Lanczos' formula again.

$$c_{k+1}r_{k+1} = AA_t r_k - \alpha_k r_k - \beta_{k-1}r_{k-1}$$

where α_k and β_{k-1} are chosen so that $r_{k+1_t}r_k$ and $r_{k+1_t}r_{k-1}$ are zero respectively. (Note that if A is symmetric, A may be written for AA_t.) This ensures an orthogonal set for the r_k. Now if one substitutes $Ax_j - y$ for r_j:

$$c_{k+1}(Ax_{k+1} - y) = AA_t r_k - \alpha_k(Ax_k - y) - \beta_{k-1}(Ax_{k-1} - y).$$

Premultiplying by A^{-1}, and noting that $A^{-1}y = x$ one obtains

$$c_{k+1}x_{k+1} - c_{k+1}x = A_t r_k - \alpha_k x_k + \alpha_k x - \beta_{k-1}x_{k-1} + \beta_{k-1}x.$$

The x is unknown, so it can be eliminated by setting $-c_{k+1} = \alpha_k + \beta_{k-1}$ so that

$$-(\alpha_k + \beta_{k-1})x_{k+1} = A_t r_k - \alpha_k x_k - \beta_{k-1}x_{k-1}.$$

It would appear that an N-step procedure can be established where the new approximation can be made up of the last two x_k and the gradient. It takes a little rearrangement at this point to simplify the scheme. If $m_k = 1/\alpha_k$, and the above equation is multiplied through by $-m_k$ and rearranged,

$$m_k(\propto_k+\beta_{k-1})x_{k+1} = x_k + m_k\beta_{k-1}x_{k-1} - m_kA_tr_k.$$

If n_{k-1} is defined as $m_k\beta_{k-1}$, one obtains

$$x_{k+1} = \frac{1}{1 + n_{k-1}}(x_k+n_{k-1}x_{k-1} - m_kA_tr_k).$$

where $$m_k = \frac{r_{k_t}r_k}{r_{k_t}AA_tr_k}$$

and $$n_{k-1} = \frac{r_{k_t}AA_tr_{k-1}}{r_{k-1_t}r_{k-1}}m_k$$

This procedure looks all right; one must be certain that n_k never gets near -1. Note that AA_t is positive definite, so that all m_j are positive. The iterative formula for the residuals can be written

$$(1+n_{k-1})r_{k+1} = r_k + n_{k-1}r_{k-1} - m_kAA_tr_k.$$

If n_{k-1} should equal -1, then substitution in the above equation yields

$$r_k - r_{k-1} = m_kAA_tr_k.$$

Squaring both sides (i.e., multiplying each by its transpose)

$$r_{k_t}r_k - 2r_{k_t}r_{k-1} + r_{k-1_t}r_{k-1} = m_k^2r_{k_t}(AA_t)^2r_k.$$

The product of different subscripted residuals is zero, so

$$r_{k_t}r_k + r_{k-1_t}r_{k-1} = m_k^2r_{k_t}(AA_t)^2r_k.$$

If one now writes the equation again and premultiplies both sides by $(r_k + r_{k-1})$, one has

$$r_{k_t}r_k + r_{k-1_t}r_{k-1} = m_kr_{k_t}AA_tr_k + m_kr_{k-1_t}AA_tr_k.$$

Setting the right sides of the last two equations equal to one another and cancelling m_k from both sides, one has

$$r_{k_t} AA_t r_k + r_{k-1_t} AA_t r_k = m_k r_{k_t} (AA_t)^2 r_k$$

$$r_{k-1_t} AA_t r_k = \frac{r_{k_t} r_k x r_{k_t} (AA_t)^2 r_k}{r_{k_t} AA_t rk} - \frac{(r_{k_t} AA_t r_k)^2}{r_{k_t} AA_t r_k}.$$

It can now be shown that the numerator is always zero or positive, hence the left side of the equation is greater than or equal to zero.

Proof: Since AA_t is positive definite and symmetric, there exists an orthogonal modal matrix L which will reduce the quadratic forms to a sum of squares. Letting $r'_k = Lr_k$ and defining the components of r_k' as $r_{11}, r_{21}, \ldots, r_{N1}$, the numerator becomes

$$(r_{11}^2 + r_{21}^2 + .. + r_{N1}^2)(\lambda_1^2 r_{11}^2 + .. + \lambda_N^2 r_{N1}^2)$$
$$-(\lambda_1 r_{11}^2 + \lambda_2 r_{21}^2 + .. + \lambda_N r_{N1}^2)^2.$$

Multiplying these out one gets

$$\sum_{j=1}^{N} \lambda_j^2 r_{j1}^4 + \sum_{j \neq k} (\lambda_j^2 + \lambda_k^2) r_{j1}^2 r_{k1}^2 - \sum_{j=1}^{N} \lambda_j^2 r_{j1}^4 - \sum_{j \neq k} 2 \lambda_j \lambda_k r_{j1}^2 r_{k1}^2.$$

The first and third summations cancel, and the second and third combine to give $\sum_{j \neq k} (\lambda_j - \lambda_k)^2 r_{j1}^2 r_{k1}^2$ which is obviously non-negative.

Note: Before leaving this point note that one has also shown that this statement is true even if the matrix is not positive definite and also that what has been shown can be stated equivalently as

$$\frac{\dfrac{r_{k_t}(AA_t)^2 r_k}{r_{k_t}AA_t r_k}}{\dfrac{r_{k_t}AA_t r_k}{r_{k_t} r_k}} \geqslant 1$$

or that $\dfrac{r_{k_t}(AA_t)^2 r_k}{r_{k_t}AA_t r_k} \geqslant \dfrac{r_{k_t}AA_t r_k}{r_{k_t} r_k}$. This is what was alluded to at the bottom of page 19 when the method of descent was discussed generally.

Since one has now shown that $r_{k-1_t}AA_t r_k$ is non-negative, then

$n_{k-1} = m_k \dfrac{r_{k-1_t}AA_t r_k}{r_{k-1_t} r_{k-1}}$ is non-negative also. But its was assumed that n_{k-1} was -1, hence the conclusion that n_{k-1} cannot be -1.

While n_{k-1} cannot be -1, it can unfortunately get close to this value, and it is a matter of experience whether the procedure is of practical value. In any event, it is clear that this study has not been ended. It is really only beginning.

1.0 Skew Symmetric Matrices

Though what is to follow may not have practical implications, it is possible to set up an N-step procedure for skew symmetric matrices. In the back of the author's mind is a procedure by which non-symmetric matrices can be partitioned into their symmetric and skew symmetric parts. Actually there is little reason to suppose that this would work.

While it is true that one is apt to outlive Methuseleh before he finds a practical example which is skew symmetric, it is of mathematical interest at any rate.

The large difficulty with skew symmetric matrices is the fact that half the time they are singular. Since all characteristic roots are imaginary, and all the coefficients of the characteristic equation are real, all roots must occur in conjugate pairs. If the order of the matrix is odd, then one imaginary root must be its own conjugate and that is, of course, zero. Then all odd-order skew symmetric matrices are singular. Even order skew symmetric matrices have a rank which is an even number, and chances are that one chosen at random will be non-singular.

Since $S = -S_t$ is the condition for skew symmetry, all quadratic forms in S or any odd power of S are zero. Thus $x_t Sx = x_t S_t x = -x_t Sx$ implies that this is true. This is a big help, for if one starts out as Lanczos did:

$$r_o$$

$$r_1 = Sr_o \qquad r_1 \text{ is orthogonal to } r_o.$$

$$r_2 = Sr_1 - \alpha_o r_o$$

One now chooses α_o so that $r_{2_t} r_o = 0$. Note that $r_{2_t} r_1$ is zero automatically since $r_{1_t} Sr_1$ vanishes, and $r_{1_t} r_o$ is zero.

$$\alpha_o = r_{1_t} S_t r_o / r_{o_t} r_o .$$

Note at this point that $r_1 = Sr_o = -S_t r_o$, hence

$$\alpha_o = -r_{1_t} r_1 / r_{o_t} r_o .$$

This is evidently a negative number, so let $a_o = - \alpha_o$.

One now wonders if an iteration formula such as

$r_{k+1} = Sr_k + a_{k-1} r_{k-1}$ is sufficient to establish an orthogonal set. One might attempt a proof of this by induction.

Proof: The statement is true for k = 0 and 1. Assume it is true for all k up to and including q-1, i.e., $r_{q_t} r_j = 0$ for $j < q$, and all $r_{p_t} r_j$ are also zero for $j \neq p$, and j and p less than q.

Choose a_{q-1} so that $r_{q+1_t} r_{q-1} = 0$. Thus

$$r_{q+1} = Sr_q + a_{q-1} r_{q-1} \tag{1}$$

$$r_{q+1_t} r_{q-1} = r_{q_t} S_t r_{q-1} + a_{q-1} r_{q-1_t} r_{q-1} = 0$$

$$a_{q-1} = -r_{q_t} S_t r_{q-1} / r_{q-1_t} r_{q-1} \tag{2}$$

Formulating the product of the transpose of (1) and r_j:

$$r_{q+1_t} r_j = r_{q_t} S_t r_j + a_{q-1} r_{q-1_t} r_j.$$

The last term on the right is zero for all $j \leqslant q$, except q-1, by hypothesis, so

$$r_{q+1_t} r_j = r_{q_t} S_t r_j \quad \text{for all } j \leqslant q \text{ except } q-1. \tag{3}$$

But

$$r_{j+1} = Sr_j + a_{j-1} r_{j-1} \quad \text{so}$$

$$r_{q_t} r_{j+1} = r_{q_t} Sr_j + a_{j-1} r_{q_t} r_{j-1}.$$

The last term on the right is zero for all $j \leqslant q$, hence

$$r_{q_t} r_{j+1} = r_{q_t} Sr_j = - r_{q_t} S_t r_j = - r_{q+1_t} r_j \quad \text{for } j < q-1. \tag{4}$$

Equation (4) is obtained by substitution from (3) into (4).

Thus

$$r_{q+1_t} r_j = - r_{q_t} r_{j+1} \quad \text{for } j < q-1.$$

But the right side is zero by hypothesis for all $j < q-1$, and so

$$r_{q+1_t} r_j = 0 \quad \text{for } j < q-1. \tag{5}$$

Since $r_{q+1_t} r_{q-1}$ was made zero by choice of a_{q-1}, then

$$r_{q+1_t} r_j = 0 \quad \text{for } j < q.$$

What about $j = q$; is this zero? Premultiplying equation (1) by r_{q_t}

$$r_{q_t} r_{q+1} = r_{q_t} S r_q + a_{q-1} \, r_{q_t} r_{q-1} = 0.$$

Thus one has shown that

$$r_{q+1_t} r_j = 0 \quad \text{for } j < q+1. \tag{6}$$

The procedure for skew symmetric matrices can now be summed up as follows:

$$r_{k+1} = Sr_k + a_{k-1} r_{k-1} \tag{7}$$

$$a_{k-1} = - \frac{r_{k_t} S_t r_{k-1}}{r_{k-1_t} r_{k-1}}$$

Since $r_k = Sr_{k-1} + a_{k-2} r_{k-2}$, $r_{k_t} r_k = r_{k_t} S r_{k-1}$ and so

$$a_{k-1} = \frac{r_{k_t} r_k}{r_{k-1_t} r_{k-1}} . \tag{8}$$

This is seen to be much simpler than the symmetric case.

2.0 The Characteristic Equation of a Matrix

Since each procedure actually constructs the characteristic equation of the matrix - A, AA_t, S, depending on the procedure - one can obtain the characteristic equation by using the m_k and ε_{k-1} obtained im the solution of $Ax = y$ in the following polynomial difference equations:

$$P_o = Q_o = 1$$

$$P_{k+1} = P_k - m_k \lambda Q_k$$

$$Q_k = P_k + \varepsilon_{k-1} Q_{k-1}.$$

If this scheme is used with the Stiefel-Hestenes procedure, one obtains the characteristic equation of A. If it is used with the author's procedure, one obtains the characteristic equation of AA_t, which is the same as the characteristic equation of A_tA. In the latter event it is quite simple to show that

$$r_k = P_k(AA_t)r_o \quad \text{and}$$

$$p_k = Q_k(AA_t)r_o.$$

Should the procedure terminate in M steps instead of N, where $M < N$, then $P_M(\lambda)$ is a factor of the characteristic equation.

3.0 Examples of the Iteration Procedures Discussed

To make the ideas clearer, several examples of the author's iteration procedures are given below. All steps are given, but some additions are absent. Simple examples are taken so that roundoff error does not affect results.

A. Non-symmetric, non-definite

$$Ax = y \quad \text{where}$$

$$A = \begin{bmatrix}1 & 1 & -1\\ 2 & -2 & 0\\ 1 & 0 & 1\end{bmatrix} \quad y = \begin{bmatrix}1\\ 0\\ 2\end{bmatrix}. \quad x_o = \begin{bmatrix}0\\ 0\\ 0\end{bmatrix} \quad r_o = \begin{bmatrix}-1\\ 0\\ -2\end{bmatrix} \quad A_t r_o = \begin{bmatrix}-3\\ -1\\ -1\end{bmatrix} \quad m_o = \frac{5}{11} \; x_1 = \frac{5}{11}\begin{bmatrix}3\\ 1\\ 1\end{bmatrix}$$

$$r_1 = \frac{2}{11}\begin{bmatrix}2\\ 10\\ -1\end{bmatrix} \quad \mathcal{E}_o = \frac{84}{121} \quad p_1 = \frac{10}{121}\begin{bmatrix}-4\\ 22\\ -19\end{bmatrix} \quad A_t p_1 = \frac{30}{121}\begin{bmatrix}7\\ -16\\ -5\end{bmatrix} \quad m_1 = \frac{77}{450} \quad x_2 = \frac{1}{15}\begin{bmatrix}16\\ 17\\ 10\end{bmatrix}$$

$$r_2 = \frac{2}{15}\begin{bmatrix}4\\ -1\\ -2\end{bmatrix} \quad \mathcal{E}_1 = \frac{121}{1125} \quad p_2 = \frac{14}{225}\begin{bmatrix}8\\ 1\\ -7\end{bmatrix} \quad A_t p_2 = \frac{14}{75}\begin{bmatrix}1\\ 2\\ -5\end{bmatrix} \quad m_2 = \frac{5}{14} \quad x_3 = \begin{bmatrix}1\\ 1\\ 1\end{bmatrix} \quad \text{Ans.}$$

If one now uses the m_k and $\mathcal{E}_k$ in the polynomial iteration:

$$P_o = Q_o = 1$$

$$P_{k+1} = P_k - m_k \lambda Q_k$$

$$Q_{k+1} = P_{k+1} + \mathcal{E}_k Q_k \qquad \text{one obtains}$$

$P_o = 1$, $Q_o = 1$, $P_1 = 1 - \frac{5}{11}\lambda$, $Q_1 = \frac{205}{121} - \frac{5}{11}\lambda$, $P_2 = 1 - \frac{67}{90}\lambda + \frac{7}{90}\lambda^2$,

$Q_2 = \frac{266}{225} - \frac{119}{150}\lambda + \frac{7}{90}\lambda^2$, and $P_3 = 1 - \frac{7}{6}\lambda + \frac{13}{36}\lambda^2 - \frac{1}{36}\lambda^3$.

Multiplying by 36 gives the characteristic equation of AA_t, viz.

$$\lambda^3 - 13\lambda^2 + 42\lambda - 36.$$

B. A non-symmetric matrix with the procedure introduced at the beginning of this chapter.

For comparison the problem of example A will be taken.

$$A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & -2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \quad . \quad x_o = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad r_o = \begin{bmatrix} -1 \\ 0 \\ -2 \end{bmatrix} \quad A_t r_o = \begin{bmatrix} -3 \\ -1 \\ -1 \end{bmatrix} \quad m_o = \frac{5}{11} \quad x_1 = \frac{5}{11}\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}$$

$$r_1 = \frac{2}{11}\begin{bmatrix} 2 \\ 10 \\ -1 \end{bmatrix} \quad A_t r_1 = \frac{6}{11}\begin{bmatrix} 7 \\ -6 \\ -1 \end{bmatrix} \quad m_1 = \frac{35}{258} \quad n_o = -\frac{98}{473} \quad 1+n_o = \frac{375}{473} \quad x_2 = \frac{1}{15}\begin{bmatrix} 16 \\ 17 \\ 10 \end{bmatrix}$$

Notice that the x_2 is the same as in example A. It would appear that the steps are the same for

$$r_2 = \frac{2}{15}\begin{bmatrix} 4 \\ -1 \\ -2 \end{bmatrix} \quad A_t r_2 = \frac{12}{15}\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \quad m_2 = \frac{7}{24} \quad n_1 = -\frac{11}{60} \quad 1+n_1 = +\frac{49}{60} \quad x_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{Ans.}$$

From a comparison of examples A and B it would appear that the latter is superior numerically since it was much easier to compute. This is something for experience to settle.

C. Skew symmetric matrix

It will be recalled that a simple orthogonalization scheme was presented for skew symmetric matrices. This procedure can be used in even-order skew symmetric simultaneous equations by evaluating all the even-subscripted approximations. The reason that the odd-subscripted approximations are not evaluated is simply because the iteration formula for x_{k+1} is

$x_{k+1} = x_{k-1}+m_k r_k$, and setting $k = 0$, one needs x_{-1} to evaluate x_1. Thus x_1, x_3, etc., cannot be evaluated. This is all right since only even-ordered systems have answers, if then.

The procedure:

$$x_o$$

$$r_o = Sx_o - y$$

$$r_1 = Sr_o$$

$$m_1 = \frac{r_{o_t} r_o}{r_{1_t} r_1}$$

$$x_2 = x_o + m_1 r_1$$

$$r_2 = Sx_2 - y$$

$$m_2 = \frac{r_{o_t} r_o}{r_{2_t} r_2}$$

$$r_3 = r_1 + m_2 S r_2$$

$$m_3 = \frac{r_{o_t} r_o}{r_{3_t} r_3}$$

$$x_4 = x_2 + m_3 r_3 \text{ , etc.}$$

An example:

Sx = y where

$$S = \begin{bmatrix} 0 & 2 & -1 & 0 \\ -2 & 0 & 2 & 1 \\ 1 & -2 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \end{bmatrix} \quad . \quad x_o = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad r_o = \begin{bmatrix} -1 \\ -1 \\ 2 \\ 0 \end{bmatrix} \quad r_1 = \begin{bmatrix} -4 \\ 6 \\ 1 \\ 3 \end{bmatrix} \quad m_1 = \frac{3}{31} \quad x_2 = \frac{3}{31}\begin{bmatrix} -4 \\ 6 \\ 1 \\ 3 \end{bmatrix}$$

$$r_2 = \frac{1}{31}\begin{bmatrix}2\\8\\5\\-15\end{bmatrix} \quad m_2 = \frac{31^2}{53} \quad r_3 = \frac{3}{53}\begin{bmatrix}43\\13\\28\\22\end{bmatrix} \quad m_3 = \frac{53}{93} \quad x_4 = \begin{bmatrix}1\\1\\1\\1\end{bmatrix} \qquad \text{Ans.}$$

The characteristic equation for S can be obtained from the following procedure: $P_o = 1$, $P_1 = \lambda$, $P_{k+1} = P_{k-1} + m_k \lambda P_k$. When the values of m_k above are used, one obtains $P_o = 1$, $P_1 = \lambda$, $P_2 = 1 + \frac{3}{31} \lambda^2$, $P_3 = \frac{1114}{53} \lambda + \frac{93}{53} \lambda^3$, $P_4 = 1 + 11 \lambda^2 + \lambda^4$.

4.0 Inverting a Matrix

As indicated by Fox, Huskey, and Wilkinson, a matrix may be inverted by these procedures by solving N problems, with

$$y_1 = \begin{bmatrix}1\\0\\0\\ \\0\end{bmatrix} \quad y_2 = \begin{bmatrix}0\\1\\0\\ \\0\end{bmatrix} \quad \circ \circ \circ \circ \quad y_N = \begin{bmatrix}0\\0\\0\\ \\1\end{bmatrix}$$

. Each answer constitutes a <u>column</u> of A^{-1}.

CHAPTER VI

EXPERIMENTS WITH LINEAR EQUATIONS, ROUNDOFF ERROR

The N-step procedures have some selling points which the author would like to review. So far as iteration procedures go, those which converge in N steps seem to be the best. The author's procedure pays no attention to symmetry and is convenient in this regard.

As far as the number of operations is concerned, the elimination methods are best. The N-step procedures have from three to six times as many multiplications, but nevertheless have other advantages. It is clear to the author that if a problem were to be solved by a hand computer, an elimination procedure is easiest. Since it takes so much time to perform an operation by hand, the time is worth money. On a high-speed computer, however, the difference in time may only be a matter of ten seconds. This is still worth money, but there are other advantages.

Elimination methods generally triangularize the original matrix, thus destroying it or requiring additional storage space. In many machines storage is more important than time. If, as a result of error accumulation due to the finite number of digits carried by the computer the answer obtained is not good, there is not much that can be done. The iteration methods have a distinct advantage in this regard. The matrix is not destroyed, and any answer can be checked.

The real problem involved here is to <u>know</u> <u>in</u> <u>advance</u> whether a set of equations is ill-conditioned. This last term warrants explanation.

First, one assumes that the matrix A and the known vector y are known accurately. If this is not true, then the whole problem is one of guesswork unless the matrix A is well-behaved or well-conditioned.

It is possible to have a set of equations and an approximation to the answer for which the residuals are very small, i.e.,

$Ax_k - y = r_k$, where r_k is small compared to y. At the same time $e_k = x_k - x$ - the error in this approximation - is large. For example, let

$$x_{11} + 3x_{21} = 4$$
$$33.33x_{11} + 100\, x_{21} = 133.33.$$

Here $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Let $x_k = \begin{bmatrix} 2.000 \\ 0.6666 \end{bmatrix}$, then $r_k = \begin{bmatrix} 0.0002 \\ 000.01 \end{bmatrix}$.

Most individuals would be quite satisfied that $x_k = x$ if they did not know x, simply because r_k is small compared to y. Evidently small residuals are not always an indication of good results.

In the process of analyzing this, we note that

$r_k = Ae_k$, and

$|r_k|^2 = r_{k_t} r_k = e_{k_t} A_t A e_k$. Hence the square of the ratio of the magnitude of the residual to the error is

$$\frac{|r_k|^2}{|e_k|^2} = \frac{e_{k_t} A_t A e_k}{e_{k_t} e_k}.$$

This is known as the Rayleigh Quotient of A_tA and is known to be bounded above by the largest characteristic number of A_tA and below by the smallest characteristic number of A_tA. Hence, if λ_n is the largest and λ_1 the smallest characteristic numbers of A_tA (which is positive definite and symmetric)

$$\lambda_1 \leqslant \frac{|r_k|^2}{|e_k|^2} \leqslant \lambda_n.$$

Apparently λ_n and λ_1 have some influence on iterative procedures.

It is known that if all the characteristic numbers were equal, convergence would be obtained in one step. Hence, the ratio $\frac{\lambda_n}{\lambda_1}$ is a figure of merit for a matrix with regard to an iterative procedure.

It would appear that the number of equations would somehow enter into the picture also. This is reasonable since the more multiplications the greater the accumulated error.

1.0 Diagonal Matrices

Some interesting results have been obtained by using the author's procedure on diagonal matrices. Since an iterative procedure cannot differentiate between matrices, a diagonal matrix is just as hard to solve as any other. It has distinct advantages for testing purposes.

It will be shown that only two things influence roundoff error in an iterative scheme: one, the spread of the characteristic numbers of A_tA, and two, the size of the component in the error vector, e_o, which is parallel to the eigenvector corresponding to the smallest characteristic number.

A diagonal matrix has its characteristic numbers and its eigenvectors in evidence. For these reasons it is possible to evaluate the effects of both the spread of characteristic numbers as well as that due to the initial error vector, e_o. In all the examples below the initial guess is the null vector, the characteristic vectors are the coordinate axes, and the answer is the vector whose components are all unity. In six dimensions, for example

$x_{o_t} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$, $x_t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$, and so

$e_{o_t} = -\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$.

If the eigenvectors have unit magnitude and are denoted by v_j:

$$v_{1_t} = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\end{bmatrix}, \quad v_{2_t} = \begin{bmatrix}0 & 1 & 0 & 0 & 0 & 0\end{bmatrix},$$
$$v_{3_t} = \begin{bmatrix}0 & 0 & 1 & 0 & 0 & 0\end{bmatrix}, \quad v_{4_t} = \begin{bmatrix}0 & 0 & 0 & 1 & 0 & 0\end{bmatrix},$$
$$v_{5_t} = \begin{bmatrix}0 & 0 & 0 & 0 & 1 & 0\end{bmatrix}, \quad v_{6_t} = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}.$$

A moment's reflection will show the reader that

$$e_o = -(v_1 + v_2 + v_3 + v_4 + v_5 + v_6).$$

Thus, with the above choice for x and x_o, one has a situation in which all components of e_o parallel to the eigenvectors are equal in magnitude.

Five examples of this type were done on the digital computer at M.I.T. - Whirlwind I. In each case the procedure was used four times; i.e., after x_6 was obtained, it was used as an initial guess and the procedure was used again in an effort to improve the answer. The x_{12} thus obtained was used as an initial guess again, etc. If we denote each of the following matrices by D_j, j some number, then the problems solved were $D_jx = y_j$, where the answer to be sought, x, was six ones. Since the diagonal matrix has its characteristic numbers on the diagonal, i.e.,

$$D_j = \begin{bmatrix}\lambda_1 & 0 & 0 & 0 & 0 & 0\\ 0 & \lambda_2 & 0 & 0 & 0 & 0\\ 0 & 0 & \lambda_3 & 0 & 0 & 0\\ 0 & 0 & 0 & \lambda_4 & 0 & 0\\ 0 & 0 & 0 & 0 & \lambda_5 & 0\\ 0 & 0 & 0 & 0 & 0 & \lambda_6\end{bmatrix},$$

the problem is totally specified by writing down the characteristic numbers. Since the answer is a group of ones, the vector y is merely the column matrix of the characteristic numbers, and so in the examples below, y, x_6, x_{12}, x_{18}, and x_{24} are given. The reader is to remember that x_{18} (say) is obtained by using x_{12} as an initial guess.

Example I

$$y = \begin{bmatrix} 10^3 \\ 10^2 \\ 1 \\ .1 \\ .01 \\ .001 \end{bmatrix} \quad x_6 = \begin{bmatrix} .9994 \\ 1.0016 \\ 1.0100 \\ .0101 \\ .0001 \\ .000001 \end{bmatrix}. \quad x_{12} = \begin{bmatrix} .9996 \\ 1.0014 \\ .5050 \\ .5100 \\ .0052 \\ .00005 \end{bmatrix} \quad x_{18} = \begin{bmatrix} .9997 \\ .9908 \\ 1.0049 \\ .5151 \\ .0053 \\ .00005 \end{bmatrix} \quad x_{24} = \begin{bmatrix} .999986 \\ 1.000073 \\ .749853 \\ .762821 \\ .010336 \\ .000104 \end{bmatrix}$$

Ratio of largest to smallest characteristic number of $D_tD = D^2$: 10^{12}. The computer carried 24 binary digits (about 7.2 decimal) and used floating point operations. Only those digits which are significant were written above, although the machine gives eight digits on the print-out.

Notice that the components of e_o along the eigenvectors v_1, v_2, and v_3 were removed almost completely in the first six steps. In the next six steps some of the v_3 was lost in an effort to remove more of v_4. Judging from the last two steps v_5 and v_6 might be removed from the error vector, if one wanted to wait long enough. Conclusion: the method is not of much use with a 10^{12} ratio. This, of course, is not strictly true. For example, if the initial guess were

$x_{o_t} = [0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1]$ it is clear good results would be obtained, probably in the first six steps. This is the same thing as knowing half the answer, which is, of course, wishful thinking.

Even though the x_{24} is not what one might deem a good answer, the residual r_{24} is

$$r_{24_t} = \begin{bmatrix} -.01373 & .00725 & -.25014 & -.02371 & -.00989 & -.00099 \end{bmatrix}$$

Example II

$$y = \begin{bmatrix} 100 \\ 10 \\ 5 \\ .5 \\ .1 \\ .01 \end{bmatrix} \quad x_6 = \begin{bmatrix} .990559 \\ .999987 \\ .999998 \\ 1.039188 \\ .042162 \\ .000422 \end{bmatrix} \quad x_{12} = \begin{bmatrix} 1.04992 \\ 1.00102 \\ .69472 \\ .84994 \\ .85871 \\ .00922 \end{bmatrix} \quad x_{18} = \begin{bmatrix} .998395 \\ .999997 \\ .999996 \\ 1.007805 \\ .864738 \\ .009639 \end{bmatrix} \quad x_{24} = \begin{bmatrix} 1.01132 \\ 1.00058 \\ .973295 \\ .908011 \\ .941039 \\ .015247 \end{bmatrix}$$

Ratio of largest to smallest characteristic number of D^2: 10^8.

Notice how the procedure seems to destroy some of the good answers in an attempt to improve some poor ones. Again one can hardly be overjoyed at the results.

Example III

$$y = \begin{bmatrix} 100 \\ 50 \\ 10 \\ 5 \\ 1 \\ .1 \end{bmatrix} \quad x_6 = \begin{bmatrix} .99902 \\ 1.40453 \\ .99921 \\ 1.01198 \\ .78848 \\ .00827 \end{bmatrix} \quad x_{12} = \begin{bmatrix} .99336 \\ .92994 \\ .97285 \\ .97390 \\ 1.01914 \\ .01957 \end{bmatrix} \quad x_{18} = \begin{bmatrix} 1.00073 \\ 1.05557 \\ .99867 \\ 1.02052 \\ 1.01161 \\ .02359 \end{bmatrix} \quad x_{24} = \begin{bmatrix} 1.00101 \\ 1.20335 \\ .97624 \\ .97721 \\ .95638 \\ .07241 \end{bmatrix}$$

Ratio of largest to smallest characteristic number of D^2: 10^6.

Example IV

$$y = \begin{bmatrix} 100 \\ 70 \\ 40 \\ 10 \\ 5 \\ 1 \end{bmatrix} \quad x_6 = \begin{bmatrix} 1.21394 \\ .94053 \\ .99409 \\ .99683 \\ 1.04826 \\ .10856 \end{bmatrix} \quad x_{12} = \begin{bmatrix} 1.00186 \\ 1.00171 \\ 1.00004 \\ .71797 \\ .71767 \\ .71761 \end{bmatrix} \quad x_{18} = \begin{bmatrix} 1.04098 \\ .99995 \\ .99780 \\ .99907 \\ 1.01412 \\ .73912 \end{bmatrix} \quad x_{24} = \begin{bmatrix} 1.00051 \\ .99999 \\ .99996 \\ .91747 \\ .91736 \\ .91735 \end{bmatrix}$$

Ratio of largest to smallest characteristic number of D^2: 10^4.

Example V

It has probably occurred to the reader by now that one large point has been overlooked, viz., who on earth would ever solve a problem with such a large discrepancy in element size? The answer to this comes in the next example. The point is that a matrix may appear to have elements of uniform size, and yet its characteristic numbers have a large spread or ratio. To demonstrate this, let us take the matrix of example IV and simply rotate axes. Thus, the problem will be changed only because our point of view is changed. To affect this, the author chose an orthogonal matrix L, where

$$L = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-1}{2} & \frac{-1}{2} & \frac{-1}{2\sqrt{3}} & \frac{-1}{2\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{1}{2} & \frac{-1}{2} & \frac{-1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{2} & \frac{1}{2} & \frac{-1}{2\sqrt{3}} & \frac{-1}{2\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{-1}{2} & \frac{1}{2} & \frac{-1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & 0 & 0 & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & 0 & 0 & \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{3}} \end{bmatrix}$$

The new matrix A was constructed by setting

$A = L_tDL$, where D is the diagonal matrix of example IV. Thus, $LAL_t = D$, and so A has the same characteristic numbers as D, and since A is symmetric, $A_tA = A^2$ has the same characteristic numbers as D^2. To seven significant figures A is

$$\begin{bmatrix} 37.66667 & 10.66667 & 0 & -24.49490 & -24.51304 & -6.12826 \\ 10.66667 & 37.66667 & -24.49490 & 0 & -6.12826 & -24.51304 \\ 0 & -24.49490 & 55. & 15. & 0 & 17.32051 \\ -24.49490 & 0 & 15. & 55. & 17.32051 & 0 \\ -42.51304 & -6.12826 & 0 & 17.32051 & 20.33333 & 6.33333 \\ -6.12826 & -24.51304 & 17.32051 & 0 & 6.33333 & 20.33333 \end{bmatrix}$$

If $x_t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$, then

$$y_t = \begin{bmatrix} -6.80286 & -6.80286 & 62.82561 & 62.82561 & 13.34587 & 13.34587 \end{bmatrix} .$$

It will be noticed that the eigenvectors of A are the rows of L, and it might be interesting to compute e_o in terms of these rows.

$$e_o = .76\ v_1 - 12.4\ v_3 - 1.97\ v_5.$$

Notice that three of the eigenvectors, including the smallest, are missing. This then is equivalent to solving the first problem with an initial guess of $x_{o_t} = \begin{bmatrix} 1.76 & 1 & -.24 & 1 & -.97 & 1 \end{bmatrix}$. It is clear that the answer should be good, and it is. For the matrix A above, the sixth step yields:

$$x_{6_t} = \begin{bmatrix} 1.0000001 & .99999976 & .99999994 & 1.0000001 & .99999988 & .99999982 \end{bmatrix} .$$

It is clear at this point that the spread in characteristic numbers, if large, must be accompanied by a shrewd guess at the answer for an initial step if good results are to be expected. Two points have not been covered as yet: the effect of dissymmetry, and the effect of roundoff due to the additional multiplications when the matrix is not diagonal.

The first question can be answered easily. A new matrix A' was constructed from A of the previous example by interchanging the first and third equations, and the sixth and fourth equations. This gives A' as

$$A' = \begin{bmatrix} 0 & -24.49490 & 55. & 15. & 0 & 17.32051 \\ 10.66667 & 37.66667 & -24.49490 & 0 & -6.12826 & -24.51304 \\ 37.66667 & 10.66667 & 0 & -24.49490 & -24.51304 & -6.12826 \\ -6.12826 & -24.51304 & 17.32051 & 0 & 6.33333 & 20.33333 \\ -24.51304 & -6.12826 & 0 & 17.32051 & 20.33333 & 6.33333 \\ -24.49490 & 0 & 15. & 55. & 17.32051 & 0 \end{bmatrix}$$

It will be demonstrated below that the characteristic numbers of $A'{}_tA'$ are the same as those of A^2 or D^2 (in example IV), and the eigenvectors of $A'{}_tA'$ are almost the same as those for A^2. Hence, no change in results should be expected, and this is corroborated by the results:

$$x_{6_t} = \begin{bmatrix} 1.0000002 & .99999964 & .99999934 & 1.0000004 & .99999940 & 1.0000001 \end{bmatrix}.$$

In the product of two n-dimensional vectors, n products are formed. Assume that in each of these the first k significant figures are retained and the rest discarded. Let @ represent the error due to rounding. This will be less than one half a unit in the last place, or, if k stands for decimal digits, $@ \leqslant .5 \times 10^{-k}$, assuming that the product is between .1 and 1.0. For example, if the product were .23760, and this were rounded to three figures (k = 3) yielding .238, the error is $.4 \times 10^{-3}$ or $.4 \times 10^{-k}$. Now the worst possible situation would arise if all n errors were cumulative, yielding a total error of $.5n \times 10^{-k}$. For n = 10, then, one would obtain the same accuracy with ten multiplications using k digits as with one multiplication with k-1 digits. This would occur if all product terms were of equal size. If one product term dominates, then its error is roughly the error of the total.

It would appear that the worst situation is not very serious, since even if one had 100 equations, only the last two digits would be, in effect, wasted. So far as a computer goes, it is generally not difficult to get around this by changing the arithmetic used. Of course, for machines with a small number of digits, this could be a problem.

2.0 Non-Diagonal Matrices

A few examples were done with matrices which were not diagonal. Some were symmetric, some non-symmetric. It hardly seems necessary to include them since they were used merely to demonstrate the fact that the procedure works and are contained in Memorandum M-2229, Digital Computer Laboratory Report, dated June 11, 1953, and written by the author. Some of the results were good, some very bad. The experiments are, however, of little scientific value since it is felt that the characteristic numbers, eigenvectors, etc., should be known in advance.

One conclusion can be drawn, but cannot be based on fact. It appears that the number of equations does not have too much to do with the efficiency of the procedure. If the number of equations is large, it is possible to have a greater number of small characteristic numbers. This possibility plus the accumulation of roundoff errors would have some effect on the speed of convergence.

In order that the author's method might be compared to that of Stiefel and Hestenes, two examples which they used in their paper[24] are given below as well as one simple fourth order non-symmetric matrix.

Example I (by Stiefel)

$$A = \begin{bmatrix} .263879 & -.014799 & .016836 & .079773 & -.020052 & .011463 \\ -.014799 & .249379 & .028764 & .057757 & -.056648 & -.134493 \\ .016836 & .028764 & .263734 & -.033628 & -.012128 & .084932 \\ .079773 & .057757 & -.033628 & .215331 & .090696 & -.037489 \\ -.020052 & -.056648 & -.012128 & -.090696 & .324486 & -.022484 \\ -.011463 & -.134493 & .084932 & -.037489 & -.022484 & .339271 \end{bmatrix}$$

$$y_t = \begin{bmatrix} .337100 & .129960 & .348510 & .372440 & .303870 & .241200 \end{bmatrix}$$

$$x_t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad x_{o_t} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Again with twenty-four binary digits:

$$x_{6_t} = \begin{bmatrix} 1.0000004 & 1.0000001 & .99999934 & 1.0000014 & 1.0000019 & .99999910 \end{bmatrix}.$$

Ratio of largest to smallest characteristic numbers of A_tA: 61.6.

Example II (by Stiefel)

$$A = \begin{bmatrix} 6 & 13 & -17 \\ 13 & 29 & -38 \\ -17 & -38 & 50 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ -3 \\ -2 \end{bmatrix} \quad x_o = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

This example was done twice. The x_3 below was used as an initial guess and x_6 was obtained.

$$x_3 = \begin{bmatrix} -1.0646319 \\ -1.9835417 \\ -1.9039863 \end{bmatrix} \quad \text{and} \quad x_6 = \begin{bmatrix} .99999475 \\ -2.9999986 \\ -2.0000007 \end{bmatrix}.$$

Ratio of largest to smallest characteristic numbers of A_tA: 2.075×10^6.

Example III (by the author)

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 2 & 2 \\ 2 & 1 & 2 & -2 \end{bmatrix} \quad y = \begin{bmatrix} 0 \\ 4 \\ 2 \\ 5 \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \quad x_o = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad x_4 = \begin{bmatrix} 1.0000041 \\ -.99999922 \\ 1.0000058 \\ -.99999594 \end{bmatrix}$$

Ratio of largest to smallest characteristic numbers of A_tA: 282.

3.0 A Priori Error Analysis

The large and unanswered question is, how does one know in advance whether the answers are good or bad? If this could be answered in one sentence, then iteration procedures would be unnecessary. The following sections talk about this problem, but no definite conclusions are drawn.

While this may sound like heresy, the author would like to raise this point. Why would anyone want an answer to a set of ill-conditioned equations? If the unknown which corresponds to the eigenvector associated with the smallest characteristic number is so ill-defined by the equations, what matter what it is? If this were a physical system, it would be a very unstable one, and hence the inability to get an answer gives a measure of the stability of the system.

Physicists have often remarked that they do not really want the solution to the matrix A anyway, since its elements are not accurate. Such an individual is usually the first to complain about the inefficient solutions. If one's mathematics or physics are sloppy, one cannot expect impeccable results.

With an iteration procedure, ill-conditioning can be discovered quite simply. Merely try two opposing initial guesses and compare results. If they are close, it is probably a good answer. If not, the opposite conclusion may be drawn.

Still, someone may insist on the solution of the set of equations. They must allow the computer to change the equations, for an ill-conditioned set can be made well-conditioned with a little effort. If this is done, some of the advantage of the iteration procedure over the elimination procedure is lost, but not the most important advantage, that which enables one to evaluate the answer.

The author and the reader would be happy at this point if such a procedure could be outlined. Some solution to this must be obtained in the near future, but the task is not simple. The author has a few suggestions which he will offer but he will not at this time vouch for them.

The equations should be rearranged so that the largest elements fall on the main diagonal. It is not necessary to do this; it is only necessary that the largest element of each row be in a different column from the largest element of the other rows. If this cannot be done, the equations are most likely ill-conditioned. The variables should now be changed in such a way as to affect a dominant diagonal. It might be appropriate to show at this point that the interchanging of rows of equations does not affect the characteristic numbers of A_tA.

Shuffling rows or columns of A does not have any effect on the characteristic numbers of $A.A_t$. This can be shown by realizing that shuffling rows of A is the same as shuffling columns of A_t. Since A_tA and AA_t have the same characteristic numbers, it is true that interchanging rows of A is permissible if it is permissible to interchange columns. Since any change of columns can be broken down into a finite number of single interchanges, it is only necessary to show that if the i^{th} and j^{th} columns of A are interchanged, then the characteristic numbers of A_tA are unchanged. Let the columns of A be called u_1, u_2, ..., u_N. Then the PQ element of A_tA is $u_{P_t}u_Q$. Now interchange the i^{th} and j^{th} columns of A. This new matrix is to be called A', and its columns are the vectors u_1, ..., u_j, u_i, ..., u_N. Then the PQ element of A'_tA' is $u_{P_t}u_Q$ for all P and Q different from i and j. Writing out the matrices A_tA and A'_tA' and indicating only the i^{th} and j^{th} columns and rows of each:

$$A_tA = \begin{bmatrix} & & & & u_{1_t}u_i & u_{1_t}u_j & & \\ & & & & \circ & \circ & & \\ u_{i_t}u_1 & u_{i_t}u_2 & \circ & \circ & u_{i_t}u_i & u_{i_t}u_j & \circ\circ\circ & u_{i_t}u_N \\ u_{j_t}u_1 & u_{j_t}u_2 & \circ & \circ & u_{j_t}u_i & u_{j_t}u_j & \circ\circ\circ & u_{j_t}u_N \\ & & & & \circ & \circ & & \\ & & & & u_{N_t}u_i & u_{N_t}u_j & & \end{bmatrix}$$

$$A'_tA' = \begin{bmatrix} & & & & u_{1_t}u_j & u_{1_t}u_i & & \\ & & & & \circ & \circ & & \\ u_{j_t}u_1 & u_{j_t}u_2 & \circ & \circ & u_{j_t}u_j & u_{j_t}u_i & \circ\circ\circ & u_{j_t}u_N \\ u_{i_t}u_1 & u_{i_t}u_2 & \circ & \circ & u_{i_t}u_j & u_{i_t}u_i & \circ\circ\circ & u_{i_t}u_N \\ & & & & \circ & \bullet & & \\ & & & & u_{N_t}u_j & u_{N_t}u_i & & \end{bmatrix}$$

Thus the determinants $|A_tA - \lambda I|$ and $|A'_tA' - \lambda I|$ are identical if the i-j columns and rows are interchanged. But the interchange of rows and columns does not affect the value of a determinant - it merely changes its sign. Thus two changes of rows and/or columns do not even change the sign (which is unimportant anyway since the determinant is to be equated to zero). Hence the determinants are equal and so are the characteristic numbers. This should not be suprising, for it is clear almost intuitively that one can expect little help from an interchange of rows or columns.

There are two tricks remaining. One is to multiply the i^{th} row by a constant and add it to the j^{th} row. In A_tA this will mean multiplying the i^{th} row and column by the constant, and adding them to the j^{th} row and column respectively. The only elements that will be affected are those in the j^{th} row and column. When λ is subtracted from all the elements on the trace of A_tA, and then the i^{th} row and column are multiplied by c and added to the j^{th} row and column, the result is not the same as when λ is subtracted from the diagonal elements of A'_tA'. Indeed, λ appears in some of the off-diagonal elements. Hence, as expected, the characteristic numbers of A_tA are changed by such a manipulation. How much and in what way one must investigate.

The other trick is to multiply one of the rows or columns by a constant. This will also change the characteristic numbers of A_tA. Then it is possible to improve the ill-conditioning of a set of equations by either of these two tricks. These are the only really simple maneuvers at one's disposal.

Just what rule one should follow is debatable. One procedure would be to rearrange the equations so that as many large numbers appear on the main diagonal as possible. These can be made approximately the same size by suitable multiplication, and the large off-diagonal elements remaining can be minimized by manipulating the equations.

Example:

$$A = \begin{bmatrix} -2 & 3 & -3 \\ 6 & -2 & 6 \\ 6 & -3 & 7 \end{bmatrix} \qquad A_tA = \begin{bmatrix} 76 & -36 & 84 \\ -36 & 22 & -42 \\ 84 & -42 & 94 \end{bmatrix}$$

The characteristic numbers of A are 1, -2, 4. The characteristic numbers of A_tA are 188.1, 3.9, .0853. The spread in characteristic numbers of A_tA is 2200 to one.

If rows one and two of A are interchanged, the matrix becomes symmetric, but with large off-diagonal elements. This matrix can be helped by adding twice the first row to the second and third rows. If this is done, the new characteristic numbers of A_tA are 37, 18.9, and .092 with a spread of 400 to one. This is not a large improvement, it is true, but it is surely an improvement. If the above operation is performed and the first and third columns are then interchanged, then

$$A' = \begin{bmatrix} -3 & 3 & -2 \\ 0 & 4 & 2 \\ 1 & 3 & 2 \end{bmatrix}$$ whose off-diagonal terms are smaller,

but still not small enough. If the first row is now added to the second and subtracted from the third, one obtains

$$A'' = \begin{bmatrix} -3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 4 & 1 \end{bmatrix} .$$ This looks as though it might be

an improvement, but one has to stop and think about what he is doing.

It is to be remembered that operations on rows are simple because one is actually attempting to solve a linear set of equations. These are lined up in rows to begin with; hence it is simple to multiply one equation by something and add it to another. It is also simple to rearrange columns since this amounts to a redefinition of the unknown x by simply rearranging its elements. But when columns are added, the x is changed, perhaps beyond recognition. For example, if the first column is added to the second, the new unknown x_{21} is now the sum of x_{11} which has not changed and the old x_{21}. This really amounts to a linear transformation and may involve the solution of another set of equations. This is not too serious for many of these transformations are trivial, and some are orthogonal. This means that the transpose of the transformation is its inverse, and hence it should be a simple matter to change from one set of variables back to the original set. See Guillemin[11], page 54 et seq.

In any event there are some things one can do to improve the equations. The only other thing left at his disposal is an improvement of his first guess. Oddly enough it is not necessary that his first guess be near the answer; it is only necessary that the error vector, e_o, contain small components of the eigenvectors corresponding to the small characteristic numbers. This can be seen from the previous examples or from the two dimensional ellipse. If the ellipse is long and thin like a cigar or worse, like a needle, then the gradient will be in the direction of the largest eigenvector, almost regardless of one's position on the ellipse excepting, of course, the very ends. Since $A_t r_k$ is the gradient of $e_{k_t} A_t A e_k$ at x_k, one is alsmot certain to wind up on one of these needles. Once there, the residuals are so small that the procedure falls to pieces. This will not matter if the component of the error along the long axis (the one corresponding to the smaller characteristic number) is small. This really does not help much since one does not know this vector. It is too difficult to find, and once found would not be of any use since one presumably does not know where the answer is. It does help one in this sense though; if one initial guess proves to give miserable results, then try another. Specifically, if all zeros are used for the first guess and the answers oscillate about plus ones, then try all twos as a first guess. It might be possible to bracket the answer in this fashion.

An approximate way to get the number of digits in the ratio of the largest to smallest characteristic numbers would prove helpful. It can be shown (see Hildebrand[26], pages 49-50, Section 1.18) that the sum of the characteristic numbers of a symmetric matrix is equal to the sum of the elements on the trace. Since the elements of the trace of A_tA are the squares of the lengths of the vectors represented by the rows of A, then the sum of the characteristic numbers of A_tA is equal to the sum of the squares of the elements of A. Since all characteristic numbers are positive, this is an upper bound on the characteristic numbers of A_tA. Thus one can obtain some estimate of the size of the largest characteristic number. Also, if the first estimate x_o is not near one of the longer axes of the ellipse, then $1/m_o \approx \lambda_{max}$. Actually $1/m_o$ will always be smaller than the largest characteristic number; hence one has an upper and lower bound on this number. As for the smallest characteristic number, one encounters much more difficulty. Since the product of all the characteristic numbers is the square of the determinant of A, then usually $\lambda_{min.} \approx |A|^2/\lambda$ max. This is rather hit or miss, but it is as simple a method as the author can devise.

Example: $A = \begin{bmatrix} -2 & 3 & -3 \\ 6 & -2 & 6 \\ 6 & -3 & 7 \end{bmatrix}$. The sum of the squares of all elements is 192, which is almost exactly the largest characteristic number. $|A|^2 = 64$, and $64/192 = .333$ which is about four times the smallest characteristic number. This gives a ratio of 578 instead of 2200, but the number of digits is only off by unity.

Another example:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 2 & -2 \\ 2 & 1 & 2 & -2 \end{bmatrix}.$$

The sum of the squares of the elements is 31, and the largest characteristic number is 16.95. $|A|^2 = 36$, so 36/31 = 1.16 while the smallest characteristic number is about .06. Here the ratio is 27 instead of 282. Again, the largest error seems to come from the smallest characteristic number.

3.0 More Equations than Unknowns

If there are fewer equations than unknowns, then it is not likely that any solution exists. Certainly no unique solution exists since several of the variables can be chosen arbitrarily, and different answers can be obtained.

If there are more than N equations and only N unknowns, then A is a MxN matrix with $M > N$. In many instances solutions of a type exist. In any event both the case for $M > N$ and $M < N$ can be covered along with the singular case.

An MxN matrix with $M < N$ can be made a NxN matrix by adding N-M rows of zeros to the bottom. One now obtains a matrix which is square, non-symmetric, and of rank less than or equal to M. This problem is now seen to be identical with the one in which A was an NxN matrix at the start, but of rank $M < N$. The two problems differ only in that the latter matrix can be symmetric while the former obviously cannot.

If A is of order N and rank M and symmetric (read A_tA for A if it is not symmetric), then there exist N-M eigenvectors of A which are mutually orthogonal to all the other eigenvectors of A, such that A times these vectors yields zero. In general an initial guess x_o will contain these eigenvectors as will the right side of the equation y. Ax_o will not possess these eigenvectors as components, however, and so the residual $r_o = Ax_o - y$ will contain as many of these eigenvectors as -y and in the same amount. It will now be recalled that $r_1 = r_o - m_oAr_o$, and so r_1 will have these same components of the original eigenvectors as -y. These eigenvectors proceed right through the procedure, so that the last residual is not zero, but contains these components of the null eigenvectors that -y had. Since $x_1 = x_o - m_or_o$, and x_o and r_o have components of these eigenvectors in general, then it is seen that something happens to these components in the approximations x_k. Just what happens to them depends on which procedure is used.

If A is non-symmetric, the correction to x is given by A_tr_k. The problem that is really being solved is $A_tAx = A_ty$. If R_k is defined as $A_tr_k = A_tA_x - A_ty$, then R_N contains the same null eigenvectors as $-A_ty$ and will not always be zero. It is to be understood that if the y or A_ty is deficient in one or more of these null eigenvectors, the procedure will terminate prematurely since r_o or R_o contain less than N independent eigenvectors. Also, r_N can equal zero if the y (or A_ty) is completely deficient in these eigenvectors. This last will always happen if there are fewer equations than unknowns.

To discover what happens to the null eigenvectors in x_o by the time it becomes x_N, one starts out as follows:

r_o, $r_1 = r_o - m_o A r_o$ (read $A_t A$ for A if A is not symmetric, and R for r)

or $r_1 = (I - m_o A) r_o$, $r_2 = r_1 - m_1 A p_1 = r_1 - m_1 A r_1 - m_1 \mathcal{E}_o A r_o$

or $r_2 = (I - m_1 A) r_1 - m_1 \mathcal{E}_o A r_o = \left[I - (m_1 + m_o + m_1 \mathcal{E}_o) A + m_1 m_o A^2 \right] r_o$

and so on until

$$r_N = \Big[I - (m_o + m_1 + \dots + m_{N-1} + m_1 \quad {}_o + m_2 \quad {}_1 + \dots . \; m_{N-1} \quad {}_{N-2} \text{ plus}$$

$$\text{more terms)}A, \text{ plus terms in } A^2 \text{ up to } A^N \Big] r_o .$$

Now in the following, the small x_k and r_k will not be the entire approximation x_k or the entire residual r_k as previously used, but <u>only</u> the components of these in the null eigenvectors of A. That is, let $x_o = c_1 e_1 + c_2 e_2$ for a two-dimensional example, the c's being constants and e's being eigenvectors of A. If e_2 is a null eigenvector of A, i.e., $Ae_2 = 0$, and e_1 is not a null eigenvector, then the new x_o about which the next paragraph will deal is simply $x_o = c_2 e_2$. One desires at this point to trace only these components through the procedure, hence the artifice. Since all residuals have the same components of these null vectors, then under our new definition, $r_o = r_1 = \dots = r_N$.

$$x_1 = x_o - m_o r_o$$

$$x_2 = x_1 - m_1 p_1 = x_o - m_o r_o - m_1 r_1 - m_1 \mathcal{E}_o r_o$$

$$= x_o - (m_1 + m_o + m_1 \mathcal{E}_o) r_o .$$

Should one continue in this manner, one would discover that

$x_N = x_o - C r_o$ where C is the coefficient of A in the polynomial in A above.

One now sees exactly what happens to the null components of the initial guess. They are merely changed in magnitude by a constant amount, depending only on A and y. Thus the answer obtained depends entirely on the initial guess. The last residual will depend on the initial guess. Thus, the residuals are not minimized in the mean square sense when A is symmetric.

An example of this might be interesting.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} .$$

Choosing $x_o = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ it is evident that x_o can have no component of the null eigenvector in it. If $\lambda_1 = 0,\ v_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$, the null vector

$$\lambda_2 = 1,\ v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \lambda_3 = 2,\ \ v_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} .$$

One can verify by substitution that $y = -\frac{1}{2}v_1 + v_2 + \frac{1}{2}v_3$, and so the null component of y (or $-r_k$) is $-\frac{1}{2}v_1$. Then the final approximation x_2 (since A is of rank 2) should contain $C(-\frac{1}{2}v_1)$. If the work is carried out, $x_{2_t} = \begin{bmatrix} 2 & -2 & 2 \end{bmatrix}$, which is a poor answer, and $x_2 = 2v_2 - 2v_1$, and C is seen to be 4.

On the other hand, if $A_tA = A_ty = A^2x = Ay$ were solved using the procedure for non-symmetric equations, the answer would be $x_{2_t} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix}$, which is the mean square solution. $A_ty = Ay = v_2 + v_3$, and so with the initial guess of zero for x_o, the answer should have no v_1 in it and it does not: $x_2 = \frac{1}{4}v_3 + v_2$.

It is now seen that the simpler procedure is almost valueless, but that the longer one is practical. How does the author's procedure behave under these circumstances?

The author's procedure is based on minimizing the error. Since there is no answer, this naturally leads to havoc. When the author's procedure was used on the above example, the result was $x_{2_t} = \begin{bmatrix} 1/6 & 1/6 & 10/6 \end{bmatrix}$. While this is nearer the mean square answer than that obtained by the method of Stiefel and Hestenes, a surprising thing occurs when this answer is used to get a better approximation. The procedure simply retraces its steps and gives a new x_{2_t} of $= \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$. In other words, it oscillates. This does not occur with Stiefel and Hestenes' simpler procedure for if $x_{o_t} = \begin{bmatrix} 2 & -2 & 2 \end{bmatrix}$ is used, the new x_{2_t} is $\begin{bmatrix} 4 & -4 & 0 \end{bmatrix}$. This is seen to be $-4v_1$, and the answer is increased by another $-2v_1$. Thus the answer grows arithmetically.

If A_tA is singular, then in three dimensions the map $e_{k_t}A_tAe_k$ is a set of tubes or pipes which are elliptical in cross section. Using the author's method no additional components of the null eigenvectors are introduced in the new approximations (x_2 above $= 1/6\ v_3 + 10/6\ v_2$), which means that all motion from approximation to approximation is in a plane normal to v_1.

This is so similar to a wave guide at cutoff that the author cannot resist making the analogy. It will be recalled that an electromagnetic wave of wave length λ can pass through a rectangular wave guide if, and only if, $\lambda/2$ is less than d, where d is the length of the longer dimension of the guide's cross section. Moreover, the smaller the wave length, the more rapid the transmission along the guide.

If the relation

$$\frac{\lambda}{2d} \sim 1 - \frac{\lambda_{min}}{\lambda_{max}}$$ is written, one can see that if λ_{min} is zero, the iteration procedure will oscillate, and also the wave guide will oscillate as a cavity resonator. In both instances no information will come out the other end. If, however, the ratio of characteristic numbers is large, i.e., near unity, the iteration procedure will converge rapidly, and the wave guide will transmit the wave at almost top speed.

It was stated that the null vector components of x_o appeared in the same magnitude in all successive x_k. This occurs for a very interesting reason. The null eigenvectors of AA_t (which is the matrix used in the author's procedure) are also the null eigenvectors of A_t.

Proof: Let v be a null eigenvector of AA_t.

$$AA_t v = 0, \text{ so}$$

$$v_t AA_t v = 0, \text{ or}$$

$$\left|A_t v\right|^2 = 0.$$ Since the magnitude of a vector is zero only if all its components are, $A_t v = 0$ Q.E.D.

All corrections in the author's method to the x_k are by vectors $A_t p_k$, and hence all null eigenvectors of AA_t in p disappear.

In the case where y contains no null eigenvectors of A, then the simpler Stiefel-Hestenes procedure is of much value in these cases.

In the event there are more equations than unknowns, i.e., if A is MxN and $M > N$, the longer Stiefel-Hestenes procedure is still the best. It gives a mean square solution which is unique if at least one set of N equations out of the M form an independent set. This is usually probable.

CHAPTER VII

NON-LINEAR EQUATIONS

It is the purpose of this chapter to discuss ways of using the author's procedure for the solution of non-linear equations. Some examples are worked out, and an engineering problem concerned with the field of control is solved.

1.0 Approximate Minimized Error Technique for Non-Linear Simultaneous Equation

A. One desires that group of numbers $x_{11}, x_{21}, \ldots\ldots, x_{N1}$ for which the N equations below are simultaneously satisfied.

$$\begin{gathered} f_1(x_{11}, x_{21}, \ldots\ldots, x_{N1}) = 0 \\ f_2(x_{11}, x_{21}, \ldots\ldots, x_{N1}) = 0 \\ \cdot \\ \cdot \\ f_N(x_{11}, x_{21}, \ldots\ldots, x_{N1}) = 0 \end{gathered}$$

Let $x_t = \left[x_{11} \quad x_{21} \quad x_{31} \quad \cdot \quad \cdot \quad \cdot \quad x_{N1} \right]$ represent the solution, and x_k the k^{th} approximation to the solution. Let p_k be a vector which will be used as before for obtaining x_{k+1} from x_k.

As in the Newton-Raphson procedure of Appendix II:

$$A_k = \begin{bmatrix} \partial f_1/\partial x_{11} & \partial f_1/\partial x_{21} & \cdot & \cdot & \partial f_1/\partial x_{N1} \\ \partial f_2/\partial x_{11} & \partial f_2/\partial x_{21} & \cdot & \cdot & \partial f_2/\partial x_{N1} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \partial f_N/\partial x_{11} & \partial f_N/\partial x_{21} & \cdot & \cdot & \partial f_N/\partial x_{N1} \end{bmatrix}$$

It is understood that all the partial derivatives are evaluated at x_k.

One now sets

$$\begin{aligned} r_{11_k} &= f_1(x_k) \\ r_{21_k} &= f_2(x_k) \\ &\cdot \\ &\cdot \\ r_{N1_k} &= f_N(x_k) \end{aligned}$$

and proceeds to write

$$\begin{aligned} x_{11_k} &= g_1(r_k) \\ x_{21_k} &= g_2(r_k) \end{aligned}$$

and so on up to

$$x_{N1_k} = g_N(r_k), \quad \text{where}$$

$$r_{k_t} = \begin{bmatrix} r_{11_k} & r_{21_k} & \cdot & \cdot & r_{N1_k} \end{bmatrix} .$$

While it may not be possible in most cases to write the x_k in terms of the r_k explicitly, the first set of equations implies the second. One now defines

$$G_k = \begin{bmatrix} \partial g_1/\partial r_{11} & \partial g_1/\partial r_{21} & \cdot & \cdot & \partial g_1/\partial r_{N1} \\ \partial g_2/\partial r_{11} & \partial g_2/\partial r_{21} & \cdot & \cdot & \partial g_2/\partial r_{N1} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \partial g_N/\partial r_{11} & \partial g_N/\partial r_{21} & \cdot & \cdot & \partial g_N/\partial r_{N1} \end{bmatrix}$$

evaluated at r_k.

It is approximately true that

$$x_k - x \approx G_k r_k .$$

Proof: For any component x_{jl_k} of x_k

$$x_{jl_k} - x_{jl} = g_j(r_k) - g_j(0) \;.$$ A Taylor expansion of $g_j(0)$

about r_k yields

$$g_j(0) = g_j(r_k) + \left[\partial g_j/\partial r_{11} x(0-r_{11}) + \ldots\ldots + \partial g_j/\partial r_{N1} x(0-r_{N1})\right]$$

(plus terms of higher order.)

If the higher order partial derivatives are small, or if the residuals are small, then

$$g_j(r_k) - g_j(0) \approx r_{11}\partial g_j/\partial r_{11} \;\ldots\ldots\; r_{N1}\partial g_j/\partial r_{N1} .$$

The right side is evidently the j^{th} row of $G_k r_k$, hence

$$x_k - x \approx G_k r_k \qquad \text{Q.E.D.}$$

It is also useful to note that

$G_{k_t} A_{k_t} = I$, the unit matrix. (See pages 112 and 113, Osgood[28].)

Proof: The ij^{th} element of the product $G_{k_t} A_{k_t}$, call it i-j, will be the scalar product of the i^{th} column of G_k and the j^{th} row of A_k or

$$i\text{-}j = \begin{bmatrix} \partial g_1/\partial r_{i1} & \partial g_2/\partial r_{i1} & \cdot & \cdot & \partial g_N/\partial r_{i1} \end{bmatrix} x \begin{bmatrix} \partial f_j/\partial x_{11} \\ \partial f_j/\partial x_{21} \\ \cdot \\ \cdot \\ \partial f_j/\partial x_{N1} \end{bmatrix}$$

$$= \frac{\partial f_j}{\partial x_{11}} \cdot \frac{\partial g_1}{\partial r_{i1}} + \frac{\partial f_j}{\partial x_{21}} \cdot \frac{\partial g_2}{\partial r_{i1}} + \ldots\ldots + \frac{\partial f_j}{\partial x_{N1}} \cdot \frac{\partial g_N}{\partial r_{i1}} .$$

But $g_1 = x_{11}$, $g_2 = x_{21}$, etc., and $f_j = r_{j1}$, hence the right hand side is

$$\sum_{m=1}^{m=N} \cdot \frac{\partial r_{j1}}{\partial x_{m1}} \cdot \frac{\partial x_{m1}}{\partial r_{i1}} = \partial r_{j1}/\partial r_{i1} .$$

That is, all the terms on the right combine to give simply the partial derivatives of the j^{th} residual with respect to the i^{th} which is unity when $i = j$, and zero otherwise. Hence the product matrix is the unit matrix, which was to be proved.

B. The procedure

Let

$$x_{k+1} = x_k - m_k A_{k_t} p_k .$$

Choose m_k so that $|e_{k+1}|$ is a minimum:

$$|e_{k+1}|^2 = e_{k+1_t} e_{k+1} = e_{k_t} e_k - 2m_k e_{k_t} A_{k_t} p_k + m_k^2 p_{k_t} A_k A_{k_t} p_k$$

Differentiating and setting the derivative with respect to m_k equal to zero one obtains for m_k

$$m_k = \frac{e_{k_t} A_{k_t} p_k}{(A_{k_t} p_k)_t (A_{k_t} p_k)}$$

Since $e_k = x_k - x \approx G_k r_k$:

$$m_k \approx \frac{r_{k_t} G_{k_t} A_{k_t} p_k}{(A_{k_t} p_k)_t (A_{k_t} p_k)} = \frac{r_{k_t} p_k}{(A_{k_t} p_k)_t (A_{k_t} p_k)}$$

Now one sets $p_k = r_k + \varepsilon_{k-1} p_{k-1}$. Postmultiplying the transpose of this by $A_k A_{k-1_t} p_{k-1}$ and choosing ε_{k-1} so that $(A_{k_t} p_k)_t (A_{k-1_t} p_{k-1}) = 0$,

$$\varepsilon_{k-1} = \frac{r_{k_t} A_k A_{k-1_t} p_{k-1}}{p_{k-1_t} A_k A_{k-1_t} p_{k-1}} .$$

Unfortunately the expression for ε_{k-1} cannot be readily simplified.

C. Newton-Raphson procedure

If a high-speed digital computer is available, it seems expedient to use the Newton-Raphson procedure with the author's iterative scheme, using the scheme to solve the simultaneous equations which arise at each step. If the process is to be done on a hand calculator, it is possible to do this also, but the additional work involved in using the procedure of part B on the preceding page does seem simpler since convergence is much more rapid. Simple examples are given below.

2.0 Two Equations in Two Unknowns

A.

$$x_{11}e^{-x_{21}} - .368 = 0$$

$$x_{11}e^{-2x_{21}} - .135 = 0$$

(It is realized that this problem is simply solved by dividing the two equations, but the author wishes to ignore this.)

The answer is

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Solution

$$A_t = \begin{bmatrix} e^{-x_{21}} & e^{-2x_{21}} \\ -x_{11}e^{-x_{21}} & -2x_{11}e^{-2x_{21}} \end{bmatrix}$$

The first guess will be

$$x_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$$A_{o_t} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

$$r_o = \begin{bmatrix} -.368 \\ -.135 \end{bmatrix}$$

$$A_{o_t} r_o = \begin{bmatrix} -.503 \\ 0 \end{bmatrix}$$

$$m_o = .605$$

$$x_1 = -.605 \begin{bmatrix} -.503 \\ 0 \end{bmatrix} = \begin{bmatrix} .305 \\ 0 \end{bmatrix}$$

$$A_{1_t} = \begin{bmatrix} 1 & 1 \\ -.305 & -.610 \end{bmatrix} \qquad r_1 = \begin{bmatrix} -.063 \\ .170 \end{bmatrix}$$

$$A_{1_t} r_1 = \begin{bmatrix} .107 \\ .086 \end{bmatrix} \qquad A_{t_1} r_o = \begin{bmatrix} -.503 \\ .194 \end{bmatrix}$$

$$\varepsilon_o = .213$$

$$p_1 = r_1 + \varepsilon_o p_o = \begin{bmatrix} -.141 \\ .141 \end{bmatrix}$$

$$A_{1_t} p_1 = \begin{bmatrix} 0 \\ -.043 \end{bmatrix}$$

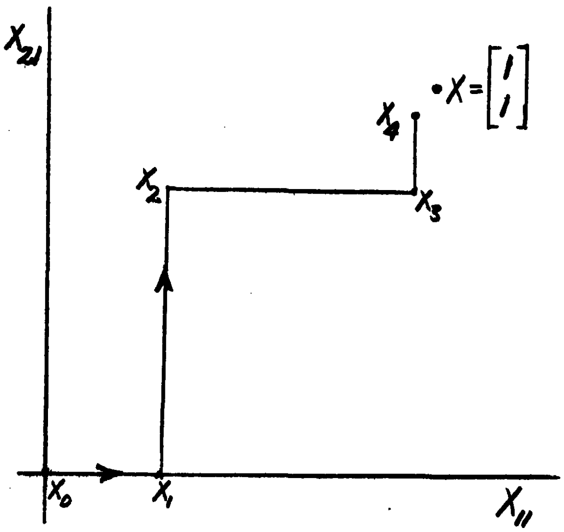
$$m_1 = 17.8$$

$$x_2 = \begin{bmatrix} .305 \\ .730 \end{bmatrix}$$

Continuing in this manner:

$$x_3 = \begin{bmatrix} .954 \\ .730 \end{bmatrix}$$

$$x_4 = \begin{bmatrix} .954 \\ .921 \end{bmatrix}$$



Evidently convergence is quite good, as indicated by the figure at the upper right.

B. To compare this procedure with the Newton-Raphson procedure, the example worked out in Appendix II will be worked for two steps. It will be recalled that the solution was $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and $x_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The first two steps (one step in the procedure, but two equations in two unknowns had to be solved) yielded as an approximation $x_2 = \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}$. Below is shown the results of the first two steps of the author's procedure.

$$(x_{11} - 1)^2 + (x_{21} + 1)^2 - 1 = 0$$

$$(x_{11} - 2)^2 + (x_{21}/2)^2 - 1 = 0$$

$$A_{k_t} = \begin{bmatrix} 2(x_{11}-1) & 2(x_{11}-2) \\ 2(x_{21}+1) & x_{21} \end{bmatrix}$$

$$A_{o_t} = \begin{bmatrix} -2 & -4 \\ 2 & 0 \end{bmatrix} \qquad r_o = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$A_{o_t} r_o = \begin{bmatrix} -14 \\ 2 \end{bmatrix} \qquad m_o = .05$$

$$x_1 = \begin{bmatrix} .7 \\ -.1 \end{bmatrix}$$

$$A_{1_t} = \begin{bmatrix} -.6 & -2.6 \\ 1.8 & -.1 \end{bmatrix}$$

$$r_1 = \begin{bmatrix} -.1 \\ .69 \end{bmatrix} \qquad A_{1_t} r_1 = \begin{bmatrix} -.173 \\ -.249 \end{bmatrix} \qquad A_{1_t} r_o = \begin{bmatrix} -8.4 \\ 1.5 \end{bmatrix}$$

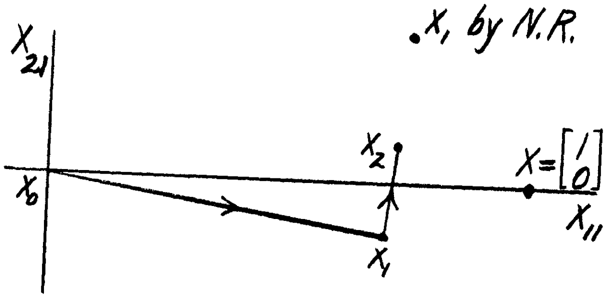
$$\mathcal{E}_o = -.197$$

$$p_1 = \begin{bmatrix} -.297 \\ .099 \end{bmatrix}$$

$$A_{1_t} p_1 = \begin{bmatrix} -.078 \\ -.545 \end{bmatrix}$$

$$m_1 = .327$$

$$x_2 = \begin{bmatrix} .725 \\ .078 \end{bmatrix}$$



If the two solutions are compared, one sees that the second is better than that obtained by the Newton-Raphson procedure whose error was .354, and the error by the latter method .286. Though this is an improvement, it is questionable just how much and whether the improved accuracy is worth the additional work.

3.0 The Impulse Response

Suppose the engineer is confronted with the problem of determining the impulse response of a black box. It will be assumed that the contents of the box are unknown to him, but that the box behaves in a linear fashion in the range desired. (If it is non-linear, the impulse response is of little value.) The linearity may be checked by experiment by exciting the box with similar signals of varying amplitudes and observing the response of the box to each of these. If the responses are similar and vary in amplitude with the input, then it can be assumed that the system is reasonably linear.

The impulse response of a linear system is a sum of exponentials and will be called h(t). Thus

$$h(t) = x_{11}e^{-x_{21}\ t} + x_{31}e^{-x_{41}\ t} + x_{51}e^{-x_{61}\ t} + \ . \ . \ .$$

where the unknowns x_{j1} are constants, real or complex. In physical systems the presence of a complex coefficient or exponent is always accompanied by another term whose coefficient and exponent are conjugate to the first. Thus h(t) is a real function of time. In general, however, the x_{j1} are otherwise unrelated. The real parts of the exponents x_{2n1} are zero or positive.

In most engineering work it is not necessary to obtain the entire impulse response. Indeed, this is not possible with the knowledge of present day servomechanisms. Usually the location of three terms is sufficient, but this number is, of course, open to conjecture. Some consideration needs to be made concerning the type of signal that is to be fed into the box. If, for example, it is to be run by a 440-cycle alternator, evidently the response is needed in this frequency range. If, on the other hand, it is a control mechanism, then the lower frequency range is the more important. It will be assumed that the engineer answers these questions as adequately as he can before attempting a solution to the problem.

The Laplace Transform of the impulse response above will be called H(s), thus

$$H(s) = \frac{x_{11}}{s+x_{21}} + \frac{x_{31}}{s+x_{41}} + \frac{x_{51}}{s+x_{61}} + \cdot \quad \cdot \quad \cdot$$

It is now apparent that the constants sought are residues and poles of the transform which is in turn the transfer function of the black box.

It is possible to obtain a graphic picture of the impulse response by other means. These means are not accurate, and moreover, give the answer to the problem in just the form that is almost valueless. The real desire of the engineer is to have a transform on which he can rely for use in analysis and control work. He might wish to draw a simplified equivalent circuit. All these require the exponential form. In addition, the exponential form can be manipulated easily in convolution, etc.

The question now arises, what if there are really five important poles and the engineer assumes but three for his solution? It is clear that none of the poles the engineer finds will be the correct ones since his three must approximate five. How much does this matter? This is an interesting problem and not part of this thesis. It is the author's feeling, however, that the location of the poles is not generally critical, and not much variation in response would be observed if these were moved around a little. Some appreciation for this can be observed by taking a simple example.

Let $h(t) = e^{-t}$, then $H(s) = 1/(s+1)$. $H(s)$ has a pole at $s = -1$. At $t = T$, where $T \gg 1$, one cuts the above impulse response off and obtains a new impulse response $h_1(t)$ which, for all practical purposes is the same as $h(t)$, thus

$h_1(t) = e^{-t}\left[u(t) - u(t-T)\right]$. $u(t-a)$ is the unit step function at $t = a$. If T is very large, little difference can be detected between the two impulse responses. Indeed, if T is greater than 4 seconds, the error is only about two per cent. But look what happens to the transform!

$H_1(s) = \dfrac{1 - e^{-T(s+1)}}{s+1}$ has no poles at all in the entire s-plane but has an essential singularity at infinity. While nothing very important happened to $h(t)$, its transform was really pushed around. For this reason and others the author feels that pushing the poles around is legitimate sport and not something to be frowned upon. The reader should not infer from this that the author feels the transform is not of much value. If this were the case, less effort would be spent trying to obtain it. It might be an interesting thesis to try to determine how accurate the location of poles need to be to give the engineer enough information to proceed.

A. Test signals

D-C. If the black box is excited by a d-c source of one volt (or one ampere depending on the manner of excitation), the steady-state response

$\Theta_{o_{d-c}}$ is given by using the final value theorem on the output transform:

$$\Theta_{o_{d-c}} = \lim_{s \to 0} \quad s(\frac{x_{11}}{s+x_{21}}) + \; . \; . \;)1/s$$

$$= \frac{x_{11}}{x_{21}} + \frac{x_{31}}{x_{41}} + \frac{x_{51}}{x_{61}} + \; . \; . \; . \qquad (1)$$

Since this response should be simple to measure, one has obtained an equation relating the variables.

Sinusoid. If the box is now excited by a voltage cos ωt, the steady-state response can be obtained from the residues of the poles at $s = \pm j\omega$, where $j = \sqrt{-1}$.

$$\lim_{s \to j\omega} \quad (s-j\omega)(\frac{x_{11}}{s + x_{21}} + \; . \; . \; . \;) \frac{s}{s^2 + \omega^2}$$

$$= \quad 1/2(\frac{x_{11}}{x_{21}+j\omega} + \frac{x_{31}}{x_{41}+j\omega} \quad . \; . \;).$$

The conjugate expression yields the residue at $s = -j\omega$. If the observed response is A sin ωt + B cos ωt, then its transform may be written

$$L(\Theta_{o_{a-c}}) = \frac{A\omega}{s^2 + \omega^2} + \frac{Bs}{s^2 + \omega^2} = 1/2 \frac{B - j\omega A}{s - j\omega} + 1/2 \frac{B = j\omega A}{s + j\omega} .$$

Equating residues gives

$$B - j\omega A = \frac{x_{11}}{x_{21} + j\omega} + \frac{x_{31}}{x_{41} + j\omega} + \frac{x_{51}}{x_{61} + j\omega} + \; . \; . \; . \qquad (2)$$

$$B + j\omega A = \frac{x^*_{11}}{x^*_{21}-j\omega} + \frac{x^*_{31}}{x^*_{41}-j\omega} + \frac{x^*_{51}}{x^*_{61}-j\omega} + \; . \; . \; . \qquad (3)$$

Again the asterisk is to be read "conjugate."

There are two cases to consider.

Case one: All x are real. Thus the asterisks may be dropped in Equation (3) and addition and subtraction of the equations gives

$$B = \frac{x_{11}x_{21}}{x_{21}^2+\omega^2} + \frac{x_{31}x_{41}}{x_{41}^2+\omega^2} + \frac{x_{51}x_{61}}{x_{61}^2+\omega^2} + \;.\;.\;. \qquad (4)$$

$$\omega A = \frac{x_{11}}{x_{21}^2+\omega^2} + \frac{x_{31}}{x_{41}^2+\omega^2} + \frac{x_{51}}{x_{61}^2+\omega^2} \quad .\;.\;. \qquad (5)$$

Case two: There are one or more conjugate pairs. If x_{31} is complex, then x_{51} must be its conjugate. Also if x_{41} is complex, $x_{61} = x^*_{41}$, and so Equations (2) and (3) on the preceding page may be written:

$$B - j\omega A = \frac{x_{11}}{x_{21}+j\omega} + \frac{x_{31}}{x_{41}+j\omega} + \frac{x_{51}}{x_{61}+j\omega} + \;.\;.\;.$$

$$B + j\omega A = \frac{x_{11}}{x_{21}-j\omega} + \frac{x_{51}}{x_{61}-j\omega} + \frac{x_{31}}{x_{41}-j\omega} + \;.\;.\;.$$

where x_{11} and x_{21} are assumed to be real, and any other conjugate pairs are treated as these were. These equations are now seen to be the same as (2) and (3) but with the conjugates dropped, and hence solving for B and ωA will again yield Equations (4) and (5). Equations (1), (4), and (5) can now be used to obtain as many relations as are necessary to solve the problem. Other relations can be obtained from the transient response, if necessary, but these yield poor results in general. The transient response is usually helpful in determining a reasonable first guess to the solution. In the example that follows, excitations of 1 - cos ωt and 1 - cos 2ωt will be used, and it can be determined from the transform that the entire response to the first is given by

$$\theta_o(t) = \theta_{o_{d-c}} - A \sin \omega t - B \cos \omega t - C e^{-x_{21} t} - D e^{-x_{41} t} - E e^{-x_{61} t} \dots$$

where Equations (1), (4), and (5) yield $\theta_{o_{d-c}}$, B, and A respectively, and C, D, E, etc., are given by

$$C = \frac{\omega^2 x_{11}}{x_{21}(x_{21}{}^2+\omega^2)}, \quad D = \frac{\omega^2 x_{31}}{x_{41}(x_{41}{}^2+\omega^2)}, \text{ etc.}$$

Setting t = 0 yields no valuable information, but setting t at some positive values will yield equations. These are more difficult in form and harder to work with. They are not very accurate for large t, and hence constitute an ill-conditioned set. One such equation is more than necessary.

B. <u>A specific problem</u>

Suppose, for this example, the exact answer is known, viz.

$$h(t) = e^{-t}+e^{-2t} - (1+j)e^{-3/2(1+j)t} - (1-j)e^{-3/2(1-j)t}.$$

Thus one has an impulse response with four exponentials, and the transform has four poles. What happens when one assumes h(t) is composed of only three exponentials?

The equations to be solved are

1. $$\frac{x_{11}}{x_{21}} + \frac{x_{31}}{x_{41}} + \frac{x_{51}}{x_{61}} = .167$$

2. $$\frac{x_{11}}{1+x_{21}{}^2} + \frac{x_{31}}{1+x_{41}{}^2} + \frac{x_{51}}{1+x_{61}{}^2} = .182$$

3. $$\frac{x_{11}x_{21}}{1+x_{21}{}^2} + \frac{x_{31}x_{41}}{1+x_{41}{}^2} + \frac{x_{51}x_{61}}{1+x_{61}{}^2} = -.227$$

4. $\dfrac{x_{11}}{4+x_{21}^{2}} + \dfrac{x_{31}}{4+x_{41}^{2}} + \dfrac{x_{51}}{4+x_{61}^{2}} = -.144$

5. $\dfrac{x_{11}x_{21}}{4+x_{21}^{2}} + \dfrac{x_{31}x_{41}}{4+x_{41}^{2}} + \dfrac{x_{51}x_{61}}{4+x_{61}^{2}} = -.294$

6. $x_{11} + x_{31} + x_{51} = 0.$

These equations are obtained simply by exciting the box with functions 1 - cos t ahd 1 - cos 2t. The response is a d-c plus a sinusoid whose phase can be determined. Since the output is usually recorded on graph paper, the constants on the right side of the above equations cannot be obtained any more accurately than to three significant figures.

Equation 6 will not always be true. This relationship among the residuals means that the response to an impulse starts at zero. This means the box is very sluggish at the outset. If the device is complicated, as is a guided missile, this assumption is not bad.

The first approximation should be a good one. It is reasonable to suppose that two of the poles are complex conjugates, the third being real. This implies that two of the unknowns are real and the other four complex. If just the first two equations are taken, and only the first terms of each, one finds that $x_{11} = .07 \pm j1.37$, and $x_{21} = .46 \pm j8.22$. Since these are complex, one should take half of x_{11} since the equations are linear in x_{11}. Thus $x_{11} = j.7$ and $x_{21} = .5 + j8$ are an indication of the magnitude of two of the unknowns. These assume that the other residual is zero, which it probably is not. If these values are substituted in equation 3, one obtains

$$\frac{x_{51}x_{61}}{1 + x_{61}^{2}} = -.402 .$$

If the same substitution is made in equation 4 one obtains

$$\frac{x_{51}}{4+x_{61}^{2}} = -.141.$$

Solving these two equations for x_{51} and x_{61} gives respectively -.87 and 1.5. From equation 6 the real parts of x_{11} and x_{31} should be about .44. Going back to the experimental results, the transient seems to have a frequency of about one or two radians per second, so choose as an initial guess:

$$x_{o_t} = \left[.5 + j.7 \quad .5 + j1.5 \quad .5 - j.7 \quad .5 - j1.5 \quad -1 \quad 1.5 \right] .$$

One now evaluates the residuals:

$$r_{o_t} = \left[.206 \quad -.050 \quad 1.27 \quad .64 \quad .062 \quad 0 \right] .$$

$$A_{k_t} = \begin{bmatrix} \frac{1}{x_{21}} & \frac{1}{1+x_{21}^{2}} & \frac{x_{21}}{1+x_{21}^{2}} & \frac{1}{4+x_{21}^{2}} & \frac{x_{21}}{4+x_{21}^{2}} & 1 \\ \frac{-x_{11}}{x_{21}^{2}} & \frac{-2x_{11}x_{21}}{(1+x_{21}^{2})^{2}} & \frac{x_{11}(1-x_{21}^{2})}{(1+x_{21}^{2})^{2}} & \frac{-2x_{11}x_{21}}{(4+x_{21}^{2})^{2}} & \frac{x_{11}(4-x_{21}^{2})}{(4+x_{21}^{2})^{2}} & 0 \\ \frac{1}{x_{41}} & \frac{1}{1+x_{41}^{2}} & \frac{x_{41}}{1+x_{41}^{2}} & \frac{1}{4+x_{41}^{2}} & \frac{x_{41}}{4+x_{41}^{2}} & 1 \\ \frac{-x_{31}}{x_{41}^{2}} & \frac{-2x_{31}x_{41}}{(1+x_{41}^{2})^{2}} & \frac{x_{31}(1-x_{41}^{2})}{(1+x_{41}^{2})^{2}} & \frac{-2x_{31}x_{41}}{(4+x_{41}^{2})^{2}} & \frac{x_{31}(4-x_{41}^{2})}{(4+x_{41}^{2})^{2}} & 0 \\ \frac{1}{x_{61}} & \frac{1}{1+x_{61}^{2}} & \frac{x_{61}}{1+x_{61}^{2}} & \frac{1}{4+x_{61}^{2}} & \frac{x_{61}}{4+x_{61}^{2}} & 1 \\ \frac{-x_{51}}{x_{61}^{2}} & \frac{-2x_{51}x_{61}}{(1+x_{61}^{2})^{2}} & \frac{x_{51}(1-x_{61}^{2})}{(1+x_{61}^{2})^{2}} & \frac{-2x_{51}x_{61}}{(4+x_{61}^{2})^{2}} & \frac{x_{51}(4-x_{61}^{2})}{(4+x_{61}^{2})^{2}} & 0 \end{bmatrix}$$

One look at the residuals tells the engineer that these could be made smaller by just choosing $x_{11} = x_{31} = x_{51} = 0$. Actually this is a more intelligent guess since the signs of the residues can be positive or negative, while the real parts of the poles have to have negative values. Then a more intelligent guess is

$$x_{o_t} = \begin{bmatrix} 0 & .5 + j1.5 & 0 & .5 - j1.5 & 0 & 1.5 \end{bmatrix}$$

$$r_{o_t} = \begin{bmatrix} -.167 & -.182 & .227 & .144 & .294 & 0 \end{bmatrix}$$ which is evidently better than before.

$$A^*_{o_t} = \begin{bmatrix} .2+j.6 & -.308+j.462 & .538+j.693 & .32+j.24 & .52-j.36 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ .2-j.6 & -.308-j.462 & .538-j.693 & .320j.24 & .52+j.36 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ .667 & .307 & .461 & .16 & .24 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Note that this is singular and that the Newton-Raphson method will not work. This does not affect the author's procedure for at the next approximation the matrix will change.

After thirty steps the last approximation (using slide rule, hence three significant figures) one obtains:

$$x_{30_t} = \begin{bmatrix} -.853-j.533 & 1.227+j1.52 & -.853+j.533 & 1.227-j1.52 & 1.706 & 1.51 \end{bmatrix}$$

with residuals:

$$r_{30_t} = \begin{bmatrix} -.013 & .027 & -.037 & .024 & .039 & 0 \end{bmatrix} .$$

In the course of the solution several short cuts were taken, such as using $\mathcal{E}_{k-1} = |r_k|^2 / |r_{k-1}|^2$ instead of the more complicated formula, and the matrix $A^*_{k_t}$ was used about four or five times before a new one was computed. Total computing time by hand about forty hours. It turns out that slide rule accuracy will not improve the answer appreciably. This is not too important since the example is merely for demonstration purposes.

If the two impulse responses are plotted on the same graph, one would notice that the original is about one-third larger, but that they are essentially the same shape. One need not be too disappointed at this for there was no reason to anticipate that the three-pole approximation would have a graph identical to the four-pole original.

Another plan of attack is to obtain the step response from the measured data by subtracting out the steady-state sinusoid. One can fit two or three exponentials to this curve in a rather simple manner. Again there is no reason to suppose that the response obtained in this fashion would satisfy the original equations unless the correct number of poles are assumed.

In any event the solution of a problem of this sort is hardly an exact business, and much judgment should go into the use to which the results are put.

APPENDIX I

MATRIX ALGEBRA

It is the author's belief that the study of simultaneous equations is most easily done from the viewpoint of geometry. Much use, however, is made of Matrix Algebra, and it is felt that a review of the more important theorems of this study will assist the reader in comprehending the thesis. Even if the reader is conversant with the algebra, it is suggested that this section be given at least a cursory glance.

1. Definition: A matrix is a rectuangular array of numbers. It will be denoted simply by a single letter as

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix} .$$

The letters with double subscripts will be called the elements of the matrix, the first subscript representing the row in which this element appears, and the second, the column. Thus a_{ij} is the element in the i^{th} row and j^{th} column. The matrix A above is called a four by three matrix (abbreviated 4x3) indicating that it contains four rows and three columns.

The matrix consisting of one column is called a column matrix or vector. The notion of a vector is borrowed from geometry. Thus the vector

$$x = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix}$$

is the directed line joining the origin (0,0,0) and (x_{11},x_{21},x_{31}). Though the human mind is confined to visualizing only three dimensions, it is a trivial matter to write a vector consisting of forty elements. This is imagined as being a vector in forty-dimensional space.

2. Addition (Subtraction) of Matrices. Only matrices which have the same number of rows and columns may be added. If a_{ij} is the ij^th^ element of A, and b_{ij} the ij^th^ element of B, then the ij^th^ element of A+B is simply $a_{ij} + b_{ij}$.

3. Multiplication of a Matrix by a scalar. A matrix is multiplied by a scalar if all its elements are multiplied by the scalar.

4. The Transpose of a matrix will be the matrix with the rows and columns interchanged. Hence, if the subscript "t" indicates the transpose then if

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \quad \text{it follows that} \quad A_t = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}.$$

5. The scalar product of two vectors is defined as the sum of the products of corresponding elements. For example

$$x_t y = \begin{bmatrix} x_{11} & x_{21} & x_{31} \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \end{bmatrix} = (x_{11}y_{11}+x_{21}y_{21}+x_{31}y_{31}).$$

The transposed column matrix or vector, x_t, above is called a row matrix. It will always be understood that a product involving a transposed vector, and a vector indicates the scalar product.

6. A non-column or non-row matrix will be thought of as a double vector set. If A is a NxM matrix it may be imagined as a set of N M-dimensional vectors transposed (the rows) or a set of M N-dimensional vectors (the columns). Let v_{1_t} be the first row of the matrix A, and v_{i_t} the i^{th} row. Let u_i be the i^{th} column of a matrix B. If these vectors have the same number of elements, then the scalar product of $v_{i_t} u_j$ will, by definition, be the i-j^{th} element of the product AB. Then, to obtain the product of two matrices, the number of columns of the first must equal the number of rows of the second, and the elements of the product are obtained by taking the scalar products of the rows of the first with the columns of the second. The matrix representing the product will have the same number of rows as the first, and the same number of columns as the second. Thus, if

$$A = \begin{bmatrix} v_{1_t} \\ v_{2_t} \\ v_{3_t} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix} \quad \text{then}$$

$$AB = \begin{bmatrix} v_{1_t}u_1 & v_{1_t}u_2 & v_{1_t}u_3 & v_{1_t}u_4 \\ v_{2_t}u_1 & v_{2_t}u_2 & v_{2_t}u_3 & v_{2_t}u_4 \\ v_{3_t}u_1 & v_{3_t}u_2 & v_{2_t}u_3 & v_{3_t}u_4 \end{bmatrix} .$$

It is now evident that an NxM matrix times an MxL matrix yields an NxL matrix. It must be pointed out that the multiplication depends on the order in which the matrices appear, and that AB $\neq$ BA in general.

Example:

$$\begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1x1-2x2 \end{bmatrix} = \begin{bmatrix} -3 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \end{bmatrix} = \begin{bmatrix} 1x1 & 1x-2 \\ 2x1 & -2x2 \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 2 & -4 \end{bmatrix}.$$

7. The transpose of a product is the product of the transposes in the reverse order, i.e.,

$$(ABCD)_t = D_t C_t B_t A_t.$$

8. A Symmetric Matrix is a matrix which is identically equal to the transpose, hence

$A = A_t$ implies that A is symmetric.

9. A matrix is said to be singular if there is some vector v such that $A_v = 0$. Since the matrix A consists of N vectors transposed (rows), this implies that each of these rows is a vector perpendicular to v. If A is not square, it is always singular. If it is square, A = 0 implies that the rows of A do not span N-dimensional space, and hence at least one of these rows is a linear combination of all the others. Determinant theory indicates that in this event, the determinant of A is zero, and hence if the determinant of a matrix is nonzero, the matrix is nonsingular.

10. The main diagonal of a matrix consists of the elements on the diagonal beginning in the upper left and proceeding to the lower right of a square matrix. All elements whose column number and row number are equal lie on the main diagonal or trace.*

11. A diagonal matrix is a square matrix all elements of which are zero except those on the trace. The unit matrix is a diagonal matrix whose trace consists solely of ones. The property of the unit matrix is that for

* This is a generalization it appears. Hildebrand[26] calls the sum of the diagonal elements the "trace."

any matrix C

$$IC = CI = C, \text{ I being the unit matrix.}$$

12. The order of a square matrix is the number of rows or columns.

13. A square non-singular matrix A always possesses a matrix A^{-1} called its inverse such that

$A^{-1}A = A \ A^{-1} = I$, where I is the unit matrix of the same order as A. It can also be shown that the inverse of a product of matrices is the product of the inverses in the reverse order, viz.

$$(ABCD)^{-1} = D^{-1}C^{-1}B^{-1}A^{-1}.$$

14. A square matrix is skew symmetric if it is equal to the negative of its transpose. The trace of a skew-symmetric matrix consists of zeros. In equation form $S = -S_t$ implies S is skew symmetric.

15. A square matrix with complex elements is called Hermitian if it is equal to the conjugate of its transpose. That is,

$H = H_t^*$ implies that H is Hermitian. The superscript * is to be read "conjugate."

16. Latent roots or characteristic values. A square matrix A possesses the property of transforming certain vectors into themselves, i.e., if v is a vector and λ a number and

$$Av = \lambda v,$$

then the vector v is called a characteristic (or eigen) vector of A, and λ a latent root (or characteristic value) of A. Evidently if

$$Av = \lambda v, \text{ then } (A - \lambda I)v = 0.$$

The matrix $(A - \lambda I)$ is therefore singular, and hence the determinant is zero. The determinant $|A - \lambda I|$ when evaluated is a polynomial in λ of the N^{th} degree. Hence there are N λ for which $A - \lambda I$ is singular, and generally, N vectors for which $(A - \lambda I)v = 0$.

It is a simple matter to show that the latent roots of a symmetric of Hermitian matrix are real, and that the roots of a skew-symmetric matrix are imaginary.

17. The Cayley-Hamilton Theorem. The characteristic equation of a square matrix defined by $|A - \lambda I| = 0$, as stated before, is a polynomial of N^{th} degree in λ. The Cayley-Hamilton Theorem states that the matrix A satisfies this equation.

Example: Find the characteristic polynomial of

$$A = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}$$

$$|A - \lambda I| = \begin{vmatrix} 1-\lambda & -1 \\ 2 & -\lambda \end{vmatrix} = (1-\lambda)(-\lambda) + 2 = \lambda^2 - \lambda + 2.$$

According to the Cayley-Hamilton Theorem then

$A^2 - A + 2I$ should vanish.

$$A^2 = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 2 & -2 \end{bmatrix}$$

$$\text{or} \quad A^2 - A + 2I = \begin{bmatrix} -1 & -1 \\ 2 & -2 \end{bmatrix} - \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

18. Quadratic forms. If x is an unknown column vector and A any square matrix, then the product $x_t Ax$ is known as a quadratic form. Since this is a multiplication of a 1xN by an NXN by a Nx1 matrix, the product is a scalar. Therefore the transpose of a quadratic form equals the quadratic form or

$$x_t AX = x_t A_t x.$$

This quadratic form admits of a geometric interpretation. The equation $x_t Ax = c$, where c is a constant, is a quadric surface composed of N-dimensional ellipses or hyperbolae. Any matrix (square) can be expressed as the sum of a symmetric matrix B and a skew-symmetric matrix S, i.e.,

$$B = 1/2(A+A_t) \quad \text{and} \quad S = 1/2(A-A_t),$$

$$\text{and } A = B+S.$$

The quadratic form of a skew-symmetric matrix and a vector x is zero since $x_t Sx = x_t S_t x = -x_t Sx$. Therefore $x_t Ax = x_t Bx$.

It can be shown that the quadric surface given by the equation $x_t Ax = 1$ is one whose axes are given by the eigenvectors of B, and the reciprocal of the squares of the lengths of the semiaxes of this surface are the characteristic values of B. The surface, by the way, is centered at the origin, and hence there always exists a linear change in variables which will rotate the axes x so that the quadratic form is reduced to a sum of squares.

The change in variables may be expressed in matrix notation as

$x = L\,x'$, where L is a matrix whose columns are the normalized eigenvectors of B (by normalized is meant the sum of the squares of the components is unity). This matrix L is usually called the normalized <u>modal</u> matrix of B, and is such that

$$x_t Ax = x_t' L_t AL\,x' = \lambda_1 x_{11}'^2 + \lambda_2 x_{21}'^2 + \ldots\ldots + \lambda_N x_{NL}'^2 .$$

$$= x_t' \Lambda x' .$$

This places the λ_k of B in evidence. It may also be noted that Λ is the diagonal matrix of the characteristic numbers of B and not A, unless A is symmetric.

If all the characteristic numbers of B are positive, then the quadratic form is always positive for any non-zero x. In this case x_tAx is a positive definite quadratic form. In this event, the matrix A will be considered to be positive definite if and only if the characteristic values of its symmetric part are all positive. For other definitions see Schönhardt.[25]

19. Rayleigh Quotient. The quadratic form $x_tIx = x_tx$ will be referred to as the squared magnitude of the length of the vector x. The ratio

$$\frac{x_tAx}{x_tx} = \frac{x_tBx}{x_tx}$$

is called the Rayleight Quotient of A and is bounded above by the largest characteristic value of B and below by the smallest characteristic value of B.

20. The characteristic values of A and A_t are identical. This follows from a knowledge of the fact that the value of a determinant is unchanged if its rows and columns are interchanged.

21. If A is a non-singular matrix, then the characteristic values of A_tA are the same as those of AA_t. (Note: A_tA and AA_t, with this one exception, have nothing in common. They are neither equal nor transposes of one another.) The proof of this amounts to a recognition of the fact that the determinant of a product of matrices is equal to the product of the determinants, hence

$$\left|A(A_tA - \quad I)A^{-1}\right| = \left|AA_t - \quad I\right| .$$

The theorem is probably true even if A is singular, but fortunately this need not be shown for the work which follows deals with non-singular matrices.

APPENDIX II

HISTORY OF ITERATION PROCEDURES

An iteration procedure is understood to be a rule or set of rules the repeated application of which will yield an answer or improved approximations to the answer of a mathematical problem.

In general, problems are solved by first guessing an answer. In the linear system no particular rules are stated for this first guess, but in non-linear systems some shrewdness is necessary. After the first guess, an automatic application of the rules of the procedure should produce the answer or an "improved" approximation to the answer.

Evidently the procedures must converge, that is, the engineer or mathematician must be certain that his work is fruitful. This would mean not only assured convergence, but rapid convergence. It is not enough to tell a man he will arrive at his destination unless he knows how long it will take. If it takes too long, he may not have the time, desire, or energy to make the trip.

1.0 History of the Problem

From the time of Newton until 1929, iteration procedures were the plaything of mathematicians. With little practical incentive, not much was done in this field. In 1929 Hardy Cross[3] brought forth a moment distribution scheme for the solution of trusses which excited wide interest. Since then much work has been done.

A. The Newton-Raphson method is familiar to all college sophomores in calculus. It is used to determine zeros of functions, or, in its extended form, to solve non-linear simultaneous equations. Though it may well be classed as the first iteration procedure of note chronologically, it has been the first in order of importance also. The reason is its rapid convergence.

In addition to the simplicity of the method, it has a geometric interpretation. Essentially it linearizes a non-linear problem in the vicinity of the approximation and obtains a new approximation by solving the linear set obtained.

For example, let $y = f(x)$. Suppose it is desired to find a zero of $f(x)$. With some restrictions it is possible to guess an x_o such that $f(x_o)$ is small. Assuming that the curve represented by $y = f(x)$ is nearly linear in the vicinity of the zero, one passes a straight line through the point $(x_o, f(x_o))$ tangent to the curve at this point. The intersection of this straight line and the x-axis is assumed to be a better approximation to the zero of $f(x)$ than x_o. Usually this is the case, and when near the answer this procedure converges with increasing speed at each step.

This procedure has been so important to the work of the author that he wises to emphasize it further. An example will make it clearer.

$$\text{Let } y = x^2 - 2.$$

It is desired to find the zero of x^2-2 which the reader will note is the square root of two. The slope of the tangent to this curve at any point x is given by the derivative, namely 2x. Thus, for any point x_k, $y_k = x_k^2 - 2$, and the slope of the tangent is $2x_k$. From the figure at the right it can be seen that the tangent line erected at $(x_k y_k)$ intersects the x-axis a distance of $\Delta x = y_k/2x_k$ from x_k. Hence a new approximation x_{k+1} can be obtained from x_k by subtracting Δx, i.e.,

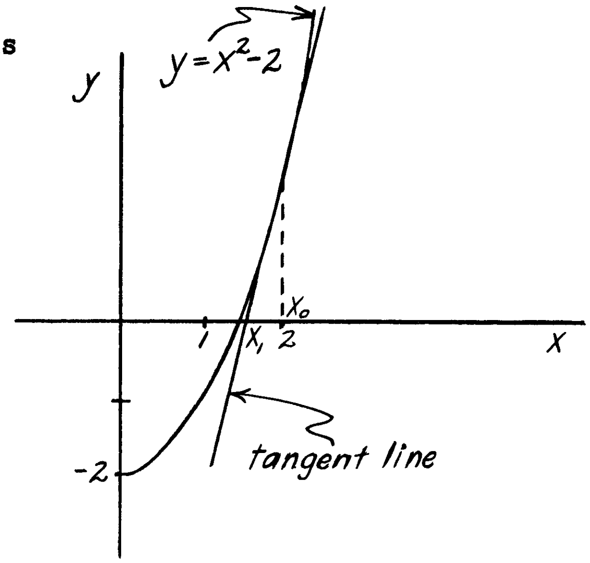
$$x_{k+1} = x_k - (x_k^2-2)/2x_k.$$

Suppose we let $x_0 = 2$, then

$$x_1 = 2 - 2/4 = 1.5$$

$$x_2 = 1.5 - .25/3 = 1.4167$$

$$x_3 = 1.4167 - .007/2.833 = 1.4142$$



This answer is already correct to five significant figures since to seven significant figures $\sqrt{2} = 1.414214$.

The speed of this method is amazing! Note that the new approximation is obtained from the old by subtraction of a correction.

Suppose now, one is confronted with two functions in two unknowns and wishes to find the values of these unknowns which will reduce both functions to zero. That is, let $z_1 = f_1(x,y)$ and $z_2 = f_2(x,y)$. As in the case of the one unknown, each of these functions can be represented as surfaces in three dimensions. The intersections of these surfaces and the x-y plane will be curves, and the point of intersection of these curves is the point desired. The Newton-Raphson method suggests that one guess at the answer as before.

One then proceeds to erect at the point corresponding to this guess a tangent plane to surface number one and a tangent plane to surface number two. These tangent planes intersect the x-y plane in straight lines, and the intersection of these straight lines is to be the new approximation to the answer.

Without going into the details the mechanics of this procedure are as follows: If h is the correction to be applied to x_k, and q the correction to be applied to y_k, then h and q are found by solving the pair of linear simultaneous equations,

$$\partial f_1/\partial x \circ h + \partial f_1/\partial y \circ q = f_1(x_k,y_k)$$
$$\partial f_2/\partial x \circ h + \partial f_2/\partial y \circ q = f_2(x_k,y_k)$$

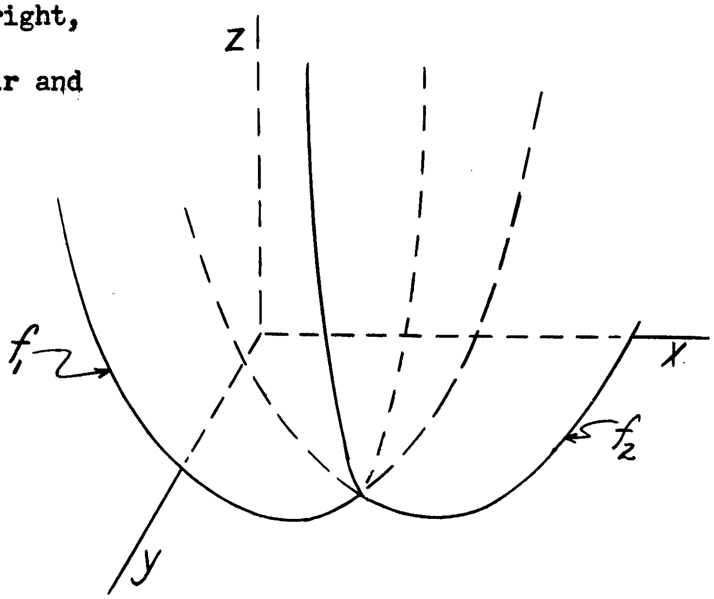
where it is to be understood that the partial derivatives are evaluated at the point (x_k,y_k). The new approximation then is

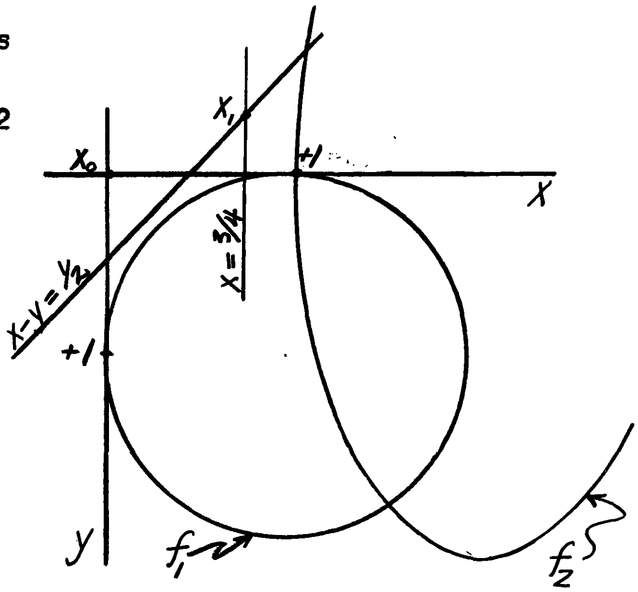
$$x_{k+1} = x_k - h \quad \text{and} \quad y_{k+1} = y_k - q.$$

Again it is felt that an example with pictures will prove helpful.

$$\text{Let } z_1 = f_1(x,y) = (x-1)^2 + (y+1)^2 -1$$
$$z_2 = f_2(x,y) = (x-2)^2 + (y/2)^2 -1.$$

If these surfaces are plotted on three-dimensional Cartesian coordinates as at the right, they are paraboloids with circular and elliptical cross sections.





The second figure, on the right, shows the intersections of the surfaces f_1 and f_2 and the x-y plane, where it is evident at a glance that one solution is (1,0). However, the first guess will be chosen as (0,0).

Tangent planes are now constructed to each of these paraboloids which intersect the x-y plane in the two lines shown, the point of intersection of these lines being the new approximation (1/4, 3/4). The reader will note that this new approximation is a great improvement over the first guess. The computations have been omitted purposely since it is the geometry that the author wished to emphasize.

The generalization of this method to N unknowns is simple. If the N unknowns to be found are x_{11}, x_{12} ,...., x_{N1}, and the N functions of these N variables are f_1, f_2, ..., f_N, one begins by defining the matrix of the Jacobian of these functions evaluated at the k^{th} approximation as

$$A_k = \begin{bmatrix} \partial f_1/\partial x_{11} & \partial f_1/\partial x_{21} & \cdot \ \cdot \ \cdot & \partial f_1/\partial x_{N1} \\ \partial f_2/\partial x_{11} & \partial f_2/\partial x_{21} & \cdot \ \cdot \ \cdot & \partial f_2/\partial x_{N1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \partial f_N/\partial x_{11} & \partial f_N/\partial x_{21} & \cdot \ \cdot \ \cdot & \partial f_N/\partial x_{N1} \end{bmatrix}$$

and r_k as the vector given by

$$r_{k_t} = \begin{bmatrix} f_1(x_k) & f_2(x_k) & \cdot \ \cdot \ \cdot & f_N(x_k) \end{bmatrix}.$$

It is to be understood here that x_k is a matrix or vector composed of all the components of the k^{th} approximation.. $f_j(x_k)$ is intended to mean $f_j(x_{1l}, \ldots, x_{Nl})$ evaluated at the k^{th} approximation.

Then the simultaneous equations representing the corrections may be written in matrix form as

$$A_k \Delta x_k = r_k, \text{ or}$$
$$\Delta x_k = A_k^{-1} r_k.$$

Thus the iteration procedure may be summed up by writing

$$x_{k+1} = x_k - A_k^{-1} r_k.$$

B. The Spark of Life. Surprisingly little was done before 1929. In 1847 Cauchy[27] presented an iteration procedure as did Seydel in 1874. In 1929, however, Hardy Cross[3] presented a paper to the A.S.C.E. entitled, "Analysis of Continuous Frames by Distributing Fixed-end Moments," His is really an application of the Seydel[1] (Seidel) procedure to structural frames, but it has the advantage that it appealed to engineers of the day in a psuedo-physical way. To give some indication of its appeal, the paper by Cross is eleven pages long and following this are 127 pages of discussion. Evidently it was thought important then, even as it is now, since almost no paper on iteration omits mention of the work.

In 1939 Temple[4] wrote a masterful paper discussing Southwell[5] and iteration in general. The Method of Descent is discussed, and amplification of Temple's work can be found in Ham.[6]

The first attempt at a geometric visualization was made by Synge[7] in 1944, and this is certainly well written in Ham.

C. N-Step Procedures. An N-step procedure is a method which, starting with a wild guess at the answer to a set of N simultaneous equations, this guess can be reduced to the exact answer in N iterations.

In 1948 Fox, Huskey, and Wilkinson[8] presented an N-step procedure for linear simultaneous equations. Though this is probably not the first such procedure of its type, it is the first reference of recent years. Lanczos[9] in 1950 presented an N-step procedure for the solution of the eigenvalue problem. Hestenes[20] in 1951 and Stiefel[10] in 1952 presented a method based on Descent.

APPENDIX III

THE AUTHOR'S PROCEDURE EXTENDED TO COMPLEX MATRICES

This section deals with the author's procedure entirely. The procedure is extended to include linear complex simultaneous equations, and a procedure for evaluating the characteristic equation of any real or complex matrix. A proof is given for the general case which is long and difficult. It is felt that only those readers who are particularly interested need to go through this. It is included simply because the work would be incomplete without it.

1.0 Complex Equations

Let the problem to be solved be

$$Ax = y$$

where A is a non-singular, complex, square matrix, x an unknown complex column matrix to be found, and y a known column matrix.

Then a procedure which will yield x in N steps is given by the formulae:

$$x_{k+1} = x_k - m_k A^*{}_t p_k \text{ .}$$

(The asterisk is to be read "conjugate.")

$$p_k = r_k + \varepsilon_{k-1} p_{k-1}$$

$$r_k = Ax_k - y$$

$$m_k = \frac{r^*{}_{k_t} r_k}{(A^*{}_t p_k)_t (A^*{}_t p_k)^*} \qquad \text{which is real and}$$

$$\varepsilon_{k-1} = \frac{r_{k_t} r^*{}_k}{r_{k-1_t} r^*{}_{k-1}} \qquad \text{which is also real.}$$

2.0 The Characteristic Equation of A

If A is a non-singular complex matrix, the characteristic equation of A can be obtained by the following double iteration procedure:

Choose e_o arbitrarily.

Let $e_o{}' = e_o$ and

$$e_{k+1} = e_k - m_k p_k \qquad e'_{k+1} = e'_k - m_k{}^* p'_k$$

$$p_k = Ae_k + \mathcal{E}_{k-1} p_{k-1} \qquad p'_k = A^*{}_t e'_k + \mathcal{E}^*{}_{k-1} p'_{k-1}$$

$$p_o = Ae_o \qquad p'_o = A^*{}_t e_o{}'$$

$$m_k = \frac{e_{k_t} p^*{}_k}{p_{k_t} p^*{}_k{}'} = \frac{p_{k_t} e^*{}_k{}'}{p_{k_t} p^*{}_k{}'}$$

$$\mathcal{E}_{k-1} = \frac{e_{k_t} A_t e^*{}_k{}'}{e_{k-1_t} A_t e^*{}_{k-1}} .$$

Once the m_k and $\mathcal{E}_{k-1}$ are all found, the iterative scheme

$$P_o = 1 \quad , Q_o = 1$$

$$P_{k+1} = P_k - m_k \lambda Q_k$$

$$Q_{k+1} = P_{k+1} + \mathcal{E}_k Q_k ,$$

where P_k and Q_k are polynomials in λ of the k^{th} degree, yields $P_N(\lambda)$ as the characteristic equation of A. If, however, the process terminates before the N^{th} step, P_M is a factor of the characteristic equation of A.

It is readily admitted that this is a somewhat complex procedure. It is practically valueless if A is not Hermitian since some of the m_k may become infinite otherwise.

If the m_k and ε_k are obtained by using the procedure of Stiefel and Hestenes and used with the two polynomial difference equations on the preceding page, one does obtain the characteristic equation of A as a by-product of the solution of the set of equations. If the m_k and ε_k are obtained from a solution of a set of equations using the author's procedure, the equation obtained is that of the characteristic equation of AA_t if A is real, and AA^*_t if A is complex.

It is the author's opinion that Lanczos' procedure is superior to these. His procedure does not even require A to be nonsingular, and hence should be easier.

3.0 General Proof

Let A be a non-singular complex square matrix of order N,

x a column matrix, and

y,y',p,p',r,r', etc. be column matrices.

$$Ax = y \tag{1}$$

$$x = A^{-1}y$$

$$A^*_t x = A^*_t A^{-1} y = y'$$

Define x_k as the k^{th} approximation to the answer x of the problem $Ax = y$, and x_k' as the k^{th} approximation to the answer x of the problem $A^*_t x = y'$, where x is the same in both instances. If $e_k = x_k - x$ and $e_k' = x_k' - x$ then

$$r_k = Ae_k$$

$$\text{and} \qquad r_k' = A^*_t e_k'. \tag{2}$$

Let $x_o = x_o'$ so that $e_o = e_o'$. (3)

$$x_{k+1} = x_k - m_k p_k \qquad \text{and} \qquad x'_{k+1} = x'_k - m_k^* p_k' \tag{4}$$

$$p_k = Ae_k + \mathcal{E}_{k-1} p_{k-1} \qquad \text{and} \qquad p'_k = A^*_t e'_k + \mathcal{E}^*_{k-1} p'_{k-1} \tag{5}$$

$$e_{k+1} = e_k - m_k p_k \qquad \text{and} \qquad e'_{k+1} = e'_k - m^*_k p'_k \tag{6}$$

m_k is now chosen so that $e_{k+1_t} e^{*\prime}_{k+1}$ is stationary, and $\mathcal{E}_{k-1}$ so that

$$p_{k_t} p^{*\prime}_{k-1} = 0.$$

One sets $p_o = Ae_o$ and $p_o' = A^*_t e_o$. . , . (7)

(a) <u>m_k</u>

$$e_{k+1_t} e^{*\prime}_{k+1} = (e_{k_t} - m_k p_{k_t})(e^{*\prime}_k - m_k p^{*\prime}_k)$$

When this is multiplied out and the first derivative is taken with respect to m_k and set equal to zero, one obtains

$$m_k = \frac{p_{k_t} e^{*\prime}_k + e_{k_t} p^{*\prime}_k}{2\, p_{k_t} p^{*\prime}_k} .$$

Assume here that the denominator does not vanish, and if it does, one starts over with a new x_o and hopes for the best.

It is now desired to demonstrate that

$$p_{k_t} e^{*\prime}_k = e_{k_t} p^{*\prime}_k \quad \text{for all } k.$$

<u>Proof</u>: (by induction)

Setting $k = 0$, and noting that $e_o = e'_o$

$$p_{o_t} e^{*\prime}_o = (Ae_o)_t e_o^{*\prime} = e_{o_t} A_t e^*_o$$

and $$e_{o_t} p^{*\prime}_o = e_{o_t} (A^*_t e_o)^* = e_{o_t} A_t e^*_o .$$

So it is true for k = 0.

Assume that it is true for k = k.

$$p_{k_t} e^{*'}{}_k = e_{k_t} p^{*'}{}_k .$$

From equations (5) and (6)

$$p_{k+1} e^{*'}{}_{k+1} = (Ae_{k+1} + \varepsilon_k p_k)_t (e^{*'}{}_k - m_k p^{*'}{}_k)$$

$$p_{k+1_t} e^{*'}{}_{k+1} = e_{k+1_t} A_t e^{*'}{}_k + \varepsilon_k p_{k_t} e^{*'}{}_k - m_k e_{k+1_t} A_t p^{*'}{}_k$$
$$- \varepsilon_k m_k p_{k_y} p^{*'}{}_k .$$

and $$e_{k+1_t} p^{*'}{}_{k+1} = (e_k - m_k p_k)_t (A_t e^{*'}{}_{k+1} + \varepsilon_k p^{*'}{}_k)$$
$$= e_{k_t} A_t e^{*'}{}_{k+1} - m_k p_{k_t} A_t e^{*'}{}_{k+1} + \varepsilon_k e_{k_t} p^{*'}{}_k$$
$$- \varepsilon_k m_k p_{k_t} p^{*'}{}_k .$$

Subtracting the two:

$$e_k{}^{+1}{}_t p^{*'}{}_{k+1} - p_{k+1_t} e^{*'}{}_{k+1} = e_{k_t} A_t e^{*'}{}_{k+1} - e_{k+1_t} A_t e^{*'}{}_k$$
$$+ m_k (e_{k+1_t} A_t p^{*'}{}_k - p_{k_t} A_t e^{*'}{}_{k+1})$$
$$+ \varepsilon_k (e_{k_t} p^{*'}{}_k - p_{k_t} e^{*'}{}_k).$$

The coefficient of ε_k is zero by assumption. Taking the first and last terms on the right together, and the second and third:

$$e_{k+1_t} p^{*'}{}_{k+1} - p_{k+1_t} e^{*'}{}_{k+1} = (e_{k_t} - m_k p_{k_t}) A_t e^{*'}{}_{k+1} - e_{k+1_t} A_t (e^{*'}{}_k - m_k p^{*'}{}_k)$$
$$= e_{k+1_t} A_t e^{*'}{}_{k+1} - e_{k+1_t} A_t e^{*'}{}_{k+1} = 0.$$

so $e_{k+1_t} p^{*'}{}_{k+1} = p_{k+1_t} e^{*'}{}_{k+1}$ Q.E.D.

and

$$m_k = \frac{e_{k_t} p^{*'}{}_{k}}{p_{k_t} p^{*'}{}_{k}} = \frac{p_{k_t} e^{*'}{}_{k}}{p_{k_t} p^{*'}{}_{k}} \qquad (8)$$

(b) $\mathcal{E}_{k-1}$

One chooses $\mathcal{E}_{k-1}$ so that

$$p_{k_t} p^{*'}{}_{k-1} = p^{*'}{}_{k_t} p_{k-1} = 0.$$

Postmultiplying (5) by $p^{*'}{}_{k-1}$ and p_{k-1}

$$p_{k_t} p^{*'}{}_{k-1} = e_{k_t} A_t p^{*'}{}_{k-1} + \mathcal{E}_{k-1} p_{k-1_t} p^{*'}{}_{k-1} = 0$$

and $$p^{*'}{}_{k_t} p_{k-1} = e^{*'}{}_{k_t} A p_{k-1} + \mathcal{E}_{k-1} p^{*'}{}_{k-1_t} p_{k-1} = 0.$$

$$\text{so} \quad \mathcal{E}_{k-1} = - \frac{e_{k_t} A_t p^{*'}{}_{k-1}}{p_{k-1_t} p^{*'}{}_{k-1}} \quad \text{or} \quad - \frac{e^{*'}{}_{k_t} A p_{k-1}}{p^{*'}{}_{k-1_t} p_{k-1}} \qquad (9)$$

It is not apparent that the above expressions for $\mathcal{E}_{k-1}$ are equivalent. To show this it must be demonstrated that

$$e_{k_t} A_t p^{*'}{}_{k-1} = e_{k_t}^{*'} A p_{k-1}$$

<u>Proof</u>: (by induction)

Setting $k = 1$

$$e_{1_t} A_t p^{*'}{}_{o} = e_{1_t} A_t A_t e^{*}{}_{o}$$

$$e^{*'}{}_{1_t} A p_o = e^{*'}{}_{1_t} A A e_o \ .$$

But $e_1 = e_o - m_o p_o$ and $e'_1 = e_o - m^*_o p'_o$

so $e^{*\prime}{}_{1_t} A^2 e_o = e^{*\prime}{}_{o_t} A^2 e_o - m_o p^{*\prime}{}_{o_t} A^2 e_o = e^*{}_{o_t} A^2 e_o - m_o e^*{}_{o_t} A^3 e_o$

and $e_{1_t} A_t{}^2 e_o = e_{o_t} A_t{}^2 e^*{}_o - m_o p_{o_t} A_t{}^2 e^*{}_o = e_{o_t} A_t{}^2 e^*{}_o - m_o e_{o_t} A_t{}^3 e^*{}_o$

$$\text{so} \qquad e_{1_t} A_t p^{*\prime}{}_o = e^{*\prime}{}_{1_t} A p_o \, .$$

It is true for k = 1.

Assume it is true for k = k, i.e., that

$$e_{k_t} A p^{*\prime}{}_{k-1} = e^{*\prime}{}_{k_t} A p_{k-1} \, .$$

From equations (5) and (6)

$$e_{k+1_t} A p^{*\prime}{}_k = (e_{k_t} - m_k p_{k_t}) A_t (A_t e^{*\prime}{}_k + \mathcal{E}_{k-1} p^{*\prime}{}_{k-1})$$

and
$$e^{*\prime}{}_{k+1_t} A p_k = (e^{*\prime}{}_{k_t} - m_k p^{*\prime}{}_t) A (A e_k + \mathcal{E}_{k-1} p_{k-1})$$

Multiplying these out and subtracting and cancelling like terms, one obtains

$$e_{k+1_t} A_t p^{*\prime}{}_k - e^{*\prime}{}_{k+1_t} A p_k = m_k (p^{*\prime}{}_{k_t} A^2 e_k + \mathcal{E}_{k-1} p^{*\prime}{}_{k_t} A p_{k-1} - p_{k_t} A_t{}^2 e^{*\prime}{}_k$$
$$- \mathcal{E}_{k-1} p_{k_t} A_t p^{*\prime}{}_{k-1}) + \mathcal{E}_{k-1} (e_{k_t} A_t p^{*\prime}{}_{k-1} - e^{*\prime}{}_{k_t} A p^{*\prime}{}_{k-1}) .$$

The last term on the right is zero by assumption, so

$$e_{k+1_t} A_t p^{*\prime}{}_k - e^{*\prime}{}_{k+1} A p_k = m_k (p^{*\prime}{}_{k_t} A \left[A e_k + \mathcal{E}_{k-1} p_{k-1}\right] - p_{k_t} A_t \left[A_t e^{*\prime}{}_k + \mathcal{E}_{k-1} p^{*\prime}{}_{k-1}\right])$$
$$= m_k (p^{*\prime}{}_{k_t} A p_k - p_{k_t} A_t p^{*\prime}{}_k) = 0 .$$

$$\text{or} \quad e_{k+1_t} A_t p^{*\prime}{}_k = e^{*\prime}{}_{k+1_t} A p_k \, . \qquad \text{Q.E.D.} \qquad (10)$$

(c) It can now be shown that

$$p_{k_t}p^{*\prime}_j = p^{*\prime}_{k_t}p_j = 0 \quad \text{for} \quad j \neq k \quad \text{and}$$

$$e^{*\prime}_{k_t}Ae_j = e_{k_t}A_te^{*\prime}_j = 0 \quad \text{for} \quad j \neq k.$$

To prove this, one first shows that

$$p_{k+1_t}p^{*\prime}_{k-1} = p^{*\prime}_{k+1_t}p_{k-1} = 0.$$

<u>Proof:</u> (by induction)

Postmultiplying (6) by $p^{*\prime}_k$, and using (8)

$$e_{k+1_t}p^{*\prime}_k = e_{k_t}p^{*\prime}_k - \frac{e_{k_t}p^{*\prime}_k}{p_{k_t}p^{*\prime}_k}p_{k_t}p^{*\prime}_k = 0,$$

and similarly
$$e^{*\prime}_{k+1_t}p_k = 0. \tag{12}$$

Postumltiplying (6) by $p^{*\prime}_{k-1}$:

$$e_{k+1_t}p^{*\prime}_{k-1} = e_{k_t}p^{*\prime}_{k-1} - m_kp_{k_t}p^{*\prime}_{k-1}$$

and

$$e^{*\prime}_{k+1_t}p_{k-1} = e^{*\prime}_{k_t}p_{k-1} - m_kp^{*\prime}_{k_t}p_{k-1} .$$

The right-hand members of both equations are all zero by equations (11) and (12) so

$$e_{k+1_t}p^{*\prime}_{k-1} = e^{*\prime}_{k+1_t}p_{k-1} = 0. \tag{13}$$

Premultiplying (5) by $e^{*\prime}_{k+1_t}$:

$$e^{*\prime}_{k+1_t}p_k = e^{*\prime}_{k+1_t}Ae_k + \varepsilon_{k-1}e^{*\prime}_{k+1_t}p_{k-1}$$

and

$$e_{k+1_t}p^{*\prime}_t = e_{k+1_t}A_te^{*\prime}_k + \varepsilon_{k-1}e_{k+1_t}p^{*\prime}_{k-1}.$$

The left-hand quantities are zero by (12), and the coefficients of $\mathcal{E}_{k-1}$ are zero by (13) so

$$e^{*\prime}_{k+1_t} Ae_k = e_{k+1_t} A_t e^{*\prime}_k = 0 \qquad (14)$$

Postmultiplying (6) by Ae_{k-1}:

$$e^{*\prime}_{k+1_t} A_t e^{*\prime}_{k-1} = e^{*\prime}_{k_t} Ae_{k-1} - m_k p^{*\prime}_{k_t} Ae_{k-1}$$

and

$$e_{k+1_t} A_t e^{*\prime}_{k-1} = e_{k_t} A_t e^{*\prime}_{k-1} - m_k p_{k_t} A_t e^{*\prime}_{k-1}.$$

The first terms on the right are zero by (14) so

$$\left.\begin{aligned} e_{k+1_t} A_t e^{*\prime}_{k-1} &= -m_k p_{k_t} A_t e^{*\prime}_{k-1} \\ \text{and} \quad e^{*\prime}_{k+1_t} Ae_{k-1} &= -m_k p^{*\prime}_{k_t} Ae_{k-1} \end{aligned}\right\} \qquad (15)$$

Setting k = 1 one obtains

$$e^{*\prime}_{2_t} Ae_o = -m_k p^{*\prime}_{1_t} Ae_o$$

$$\text{and} \quad e_{2_t} A_t e^{*\prime}_o = -m_1 p_1 t A_t e^{*}_o.$$

But $Ae_o = p_o$ and $A_t e^{*}_o = p^{*\prime}_o$ so

$$\left.\begin{aligned} e^{*\prime}_{2_t} Ae_o &= -m_1 p^{*\prime}_{1_t} p_o = 0 \\ \text{and} \quad e_{2_t} A_t e^{*\prime}_o &= -m_1 p^{*\prime}_{o_t} p_1 = 0 \text{ by equation (11).} \end{aligned}\right\} \qquad (16)$$

From equations (6):

$$p_k = \frac{1}{m_k}(e_k - e_{k+1})$$

$$p^{*\prime}_k = \frac{1}{m_k}(e^{*\prime}_k - e^{*\prime}_{k+1}).$$

From equations (5)

$$p_{k+2} = Ae_{k+2} + \varepsilon_{k+1}p_{k+1}$$

and
$$p^{*'}_{k+2} = A_t e^{*'}_{k+2} + \varepsilon_{k+1}p_{k+1}.$$

Combining these two equations:

$$p_{k+2_t} p_k^{*'} = \frac{e^{*'}_{k+2_t} A_t}{m_k} (e^{*'}_k - e^{*'}_{k+1})$$

and
$$p^{*'}_{k+2_t} p_k = \frac{e^{*'}_{k+2_t} A}{m_k} (e_k - e_{k+1})$$

When multiplied out, the last term on the right is zero by (14) so

$$\left.\begin{aligned} p_{k+2_t} p^{*'}_k &= \frac{1}{m_k} e_{k+2_t} A_t e^{*'}_k \\ \text{and} \quad p^{*'}_{k+2_t} p_k &= \frac{1}{m_k} e^{*'}_{k+2_t} A e_k \end{aligned}\right\} \qquad (17)$$

Setting k = 0, equation (16) indicates that

$$p_{k+2_t} p^{*'}_k = p^{*'}_{k+2_t} p_k = 0 \quad \text{for } k = 0.$$

Assume it is true for k = k, i.e.,

$$p_{k+2_t} p^{*'}_k = p^{*'}_{k+2_t} p_k = 0.$$

From equation (5)

$$Ae_{k+1} = p_{k+1} - \varepsilon_k p_k \quad \text{and} \quad A_t e^{*'}_{k+1} = p^{*'}_{k+1} - \varepsilon_k p^{*'}_k$$

From equation (15):

$$\begin{aligned} e^{*'}_{k+3_t} Ae_{k+1} &= -m_{k+2} p^{*'}_{k+2_t} Ae_{k+1} = -m_{k+2} p^{*'}_{k+2_t} (p_{k+1} - \varepsilon_k p_k) \\ &= -m_{k+2} p^{*'}_{k+2_t} p_{k+1} + m_{k+2}\, \varepsilon_k p^{*'}_{k+2_t} p_k \end{aligned}$$

and similarly

$$e_{k+3_t}A_t e^{*\prime}{}_{k+1} = -m_{k+2}p_{k+2_t}p^{*\prime}{}_{k+1} + m_{k+2}\;\varepsilon_k p_{k+2_t}p^{*\prime}{}_k.$$

But the first terms on the right of each equation is zero by (11) so

$$\left.\begin{aligned} e^{*\prime}{}_{k+3_t}Ae_{k+1} &= m_{k+2}\;\varepsilon_k p^{*\prime}{}_{k+2_t}p_k \\ \text{and}\qquad e_{k+3_t}A_t e^{*\prime}{}_{k+1} &= m_{k+2}\;\varepsilon_k p_{k+2_t}p_k{}^{*\prime}{}_k \end{aligned}\right\} \tag{18}$$

Equation (17) is now substituted in (18), increasing k by 1:

$$p_{k+3_t}p^{*\prime}{}_{k+1} = \frac{1}{m_{k+1}}\,e_{k+3_t}A_t e^{*\prime}{}_{k+1} = \frac{m_{k+2}\;\varepsilon_k}{m_{k+1}}\,p_{k+2_t}p^{*\prime}{}_k$$

Since the right-most term is zero by assumption,

$$p_{k+3_t}p^{*\prime}{}_{k+1} = 0.$$

In a similar way one shows that $p^{*\prime}{}_{k+3_t}p_{k+1} = 0$ so

$$p_{k+2_t}p^{*\prime}{}_k = p^{*\prime}{}_{k+2_t}p_k = 0 \quad \text{for all } k \qquad \text{Q.E.D.} \tag{19}$$

From (18) one sees that

$$e^{*\prime}{}_{k+2_t}Ae_k = e_{k+2_t}A_t e^{*\prime}{}_k = 0 \text{ also.} \tag{20}$$

This equation is similar to (14), hence one may proceed with the proof of

$$p_{k+3_t}p^{*\prime}{}_k = p^{*\prime}{}_{k+3_t}p_k = 0$$

in exactly the same way, and so forth. Hence it is true then that

$$p_{k_t}p^{*\prime}{}_j = p^{*\prime}{}_{k_t}p_j = e_{k_t}A_t e^{*\prime}{}_j = e^{*\prime}{}_{k_t}Ae_j = 0$$

for $j \neq k$.

(d) The p_n and p'_n form an independent set if for <u>no</u> k

$$p_{k_t} p^{*\prime}_k = 0.$$

This can be shown by assuming that this is not true, that is, by assuming one p_j is a linear combination of the others. Premultiplying this by $p^{*\prime}_{j_t}$ yields a contradiction.

(e) Thus $p_N = p'_N = 0$, and e_N and $e'_N = 0$.

(f) If A is Hermitian (or symmetric if real), convergence is assured, and double iteration unnecessary for the m_k and $\mathcal{E}_k$ are real, and all the primed quantities are therefore equal to the unprimed quantities. That the m_k and $\mathcal{E}_k$ are real can be most easily seen by noting that if they are real, the unprimed quantities equal the primed quantities, and that if this is true, the constants must be real. A proof of this is a simple matter and can be done by induction.

If $A = A^*_t$ and since $e_o = e'_o$ then $p^{*\prime}_o = A_t e^*_o = p^*_o$. Surely then, m_o is real and $e_1 = e'_1$ so $\mathcal{E}_o$ is real. It follows therefore that $p_1 = p'_1$ and so forth.

(g) $e_{k_t} p^{*\prime}_k = e_{k_t} A_t e^{*\prime}_k$ no matter what A is.

Since $p^{*\prime}_k = A_t e^{*\prime}_k + \mathcal{E}_{k-1} p^{*\prime}_{k-1}$

then $$e_{k_t} p^{*\prime}_k = e_{k_t} A_t e^{*\prime}_k + \mathcal{E}_{k-1} e_{k_t} p^{*\prime}_{k-1}.$$

The last term is zero by equation (12), so

$$e_{k_t} p^{*\prime}_k = e_{k_t} A_t e^{*\prime}_k \text{ as asserted.}$$

Thus m_k can be rewritten in light of this and equation (8)

$$m_k = \frac{e_{k_t}A_te^{*'}{}_k}{p_{k_t}p^{*'}{}_k} \tag{21}$$

(h) Using $e^{*'}{}_{j-1} = e^{*'}{}_j + m_{j-1}p^{*'}{}_{j-1}$

one shows that $e_{k_t}A_te^{*'}{}_{j-1} = 0$ implies that

$$m_{j-1} = \frac{-e_{k_t}A_te^{*'}{}_j}{e_{k_t}A_tp^{*'}{}_{j-1}} \qquad \text{for} \quad k \neq j-1.$$

Multiplying this by equation (9) yields

$$\varepsilon_{k-1}m_{k-1} = \frac{e_{k_t}A_te^{*'}{}_k}{p_{k-1_t}p^{*'}{}_{k-1}} \quad \text{by setting } j = k.$$

Using equation (8) for m_{k-1}, and the result of (g) above

$$\varepsilon_{k-1} = \frac{e_{k_t}A_te^{*'}{}_k}{e_{k-1_t}A_te^{*'}{}_{k-1}} \tag{22}$$

4.0 Résumé

The procedure then can be written alternatively as

$$e_o$$

$$e'_o = e_o$$

$$p_o = Ae_o;\ p'_o = A^*_te_o$$

$$e_{k+1} = e_k - m_kp_k \qquad\qquad e'_{k+1} = e'_k - m^*_kp_k{}'$$

$$p_k = Ae_k + \varepsilon_{k-1}p_{k-1} \qquad\qquad p'_k = A^*_te'_k + \varepsilon^*_{k-1}p'_{k-1}$$

$$m_k = \frac{e_{k_t} A_t e^{*\prime}{}_k}{p_{k_t} p^{*\prime}{}_k}$$

$$\mathcal{E}_{k-1} = \frac{e_{k_t} A_t e^{*\prime}{}_k}{e_{k-1_t} A_t e^{*\prime}{}_{k-1}}$$

If, in this procedure, A is replaced by $A^*{}_t A$, and p_k by $A^*{}_t p_k$, the procedure reduces to that of Section 1 in this chapter. Since $A_t A$ is Hermitian, then only one iteration is necessary, as asserted.

5.0 Further Examples of these Procedures

A. Non-symmetric, hon-hermitian complex matrix.

$$Ax = y \text{ where}$$

$$A = \begin{bmatrix} j & 1 & 1 \\ -1 & 1+j & 1-j \\ 0 & -j & 2 \end{bmatrix} \quad y = \begin{bmatrix} 1+j \\ 3 \\ 2+j \end{bmatrix} \quad \text{where } j = \sqrt{-1}. \text{ Let } x_o = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad r_o = \begin{bmatrix} -1-j \\ -3 \\ -2-j \end{bmatrix}$$

$$A^*{}_t r_o = \begin{bmatrix} 2+j \\ -3 \\ -8-6j \end{bmatrix} \quad m_o = \frac{8}{57} \quad x_1 = \frac{8}{57} \begin{bmatrix} -2-j \\ 3 \\ 8+6j \end{bmatrix} \quad r_1 = \frac{1}{57} \begin{bmatrix} 39-25j \\ -19+16j \\ 14+15j \end{bmatrix}$$

. Note that the residuals are not orthogonal, but the conjugate of one is orthogonal to the other.

$$\mathcal{E}_o = \frac{199}{57^2} \quad p_1 = \frac{8}{57^2} \begin{bmatrix} 253-203j \\ -210+114j \\ 50+82j \end{bmatrix} \quad A^*{}_t p_1 = \frac{8}{57^2} \begin{bmatrix} 7-367j \\ 75+171j \\ 29-135j \end{bmatrix} \quad m_1 = \frac{199 \text{x} 57}{4 \text{x} 3,310}$$

$$x_2 = \frac{1}{5x331}\begin{bmatrix} -489+1,049j \\ 435 - 597j \\ 1,757+1,865j \end{bmatrix} \quad r_2 = \frac{1}{5x331}\begin{bmatrix} -512-876j \\ 178-1,103j \\ -393+1,640j \end{bmatrix} \quad \mathcal{E}_1 = \frac{57^2x12,869}{5^2x331^2x8}$$

$$p_2 = \frac{199}{5^2x331^2}\begin{bmatrix} 12,103-20,413j \\ -12,100-1,801j \\ -35+18,942j \end{bmatrix} \qquad A^*{}_t p_2 = \frac{17x199}{5^2x331^2}\begin{bmatrix} -489-606j \\ -1,220-597j \\ 102+210j \end{bmatrix}$$

$$m_2 = \frac{5x331}{17x199} \qquad x_3 = \begin{bmatrix} j \\ 1 \\ 1+j \end{bmatrix} \quad \text{Ans.}$$

Using the constants in the polynomial iteration scheme, one obtains the characteristic equation of $AA^*{}_t$:

$P_o = 1$, $Q_o = 1$, $P_1 = 1 - \frac{8}{57}\lambda$, $Q_1 = \frac{3448}{57^2} - \frac{8}{57}\lambda$, $P_2 = 1 - \frac{99,009}{57x1655}\lambda + \frac{199}{1655}\lambda^2$, $Q_2 = \frac{5^2x331^2+12,869x431}{5^2x331^2} - \frac{99,009x1655+57^2x12,869}{57x5^2x331^2}\lambda + \frac{199}{1655}\lambda^2$,

$P_3 = 1 - \frac{43}{17}\lambda + \frac{13}{17}\lambda^2 - \frac{1}{17}\lambda^3$, or multiplying by -17, $\lambda^3 - 13\lambda^2 + 43\lambda - 17$.

B. The characteristic equation of A, where A is nonsymmetric and nondefinite.

For purposes of comparison, the same matrix will be used as in Examples 1 and 2 of Chapter V.

$A = \begin{bmatrix} 1 & 1 & -1 \\ 2 & -2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. This is a double iteration procedure, as previously indicated, and curiously enough a choice of $e_o = \begin{bmatrix}1\\0\\0\end{bmatrix}$ causes $p_t p' = 0$, i.e., the procedure does not work for this initial guess. If e_o is chosen as $\begin{bmatrix}0\\1\\0\end{bmatrix}$ it works fine.

$$e_o = \begin{bmatrix}0\\1\\0\end{bmatrix} \quad p_o = \begin{bmatrix}1\\-2\\0\end{bmatrix} \quad p'_o = \begin{bmatrix}2\\-2\\0\end{bmatrix} \quad m_o = -\frac{1}{3} \quad e_1 = \frac{1}{3}\begin{bmatrix}1\\1\\0\end{bmatrix} \quad e'_1 = \frac{1}{3}\begin{bmatrix}2\\1\\0\end{bmatrix}$$

$$Ae_1 = \frac{1}{3}\begin{bmatrix}2\\0\\1\end{bmatrix} \quad A_t e'_1 = \frac{1}{3}\begin{bmatrix}4\\0\\-2\end{bmatrix} \quad \varepsilon_o = -\frac{2}{9} \quad p_1 = \frac{1}{9}\begin{bmatrix}4\\4\\3\end{bmatrix} \quad p'_1 = \frac{2}{9}\begin{bmatrix}4\\2\\-3\end{bmatrix}$$

$$m_1 = \frac{6}{5} \quad e_2 = \frac{1}{5}\begin{bmatrix}-1\\-1\\-2\end{bmatrix} \quad e'_2 = \frac{1}{5}\begin{bmatrix}-2\\-1\\4\end{bmatrix} \quad Ae_2 = \frac{1}{5}\begin{bmatrix}0\\0\\-3\end{bmatrix} \quad A_t e'_2 = \frac{1}{5}\begin{bmatrix}0\\0\\6\end{bmatrix}$$

$$\varepsilon_1 = -\frac{27}{25} \quad p_2 = \frac{12}{25}\begin{bmatrix}-1\\-1\\-2\end{bmatrix} \quad p'_2 = \frac{12}{25}\begin{bmatrix}-2\\-1\\4\end{bmatrix} \quad m_2 = \frac{5}{12} \quad e_3 = e'_3 = \begin{bmatrix}0\\0\\0\end{bmatrix}.$$

Using the constants m_k and ε_k in the polynomial iteration scheme as before, one obtains

$P_o = 1$, $Q_o = 1$, $P_1 = 1 + \frac{1}{3}\lambda$, $Q_1 = \frac{7}{9} + \frac{1}{3}\lambda$, $P_2 = 1 - \frac{3}{5}\lambda - \frac{2}{5}\lambda^2$, $Q_2 = \frac{4}{25} - \frac{24}{25}\lambda - \frac{2}{5}\lambda^2$, $P_3 = 1 - \frac{2}{3}\lambda + \frac{1}{6}\lambda^3$. Multiplying P_3 by 6 gives the characteristic equation of A, viz. $\lambda^3 - 4\lambda + 6$.

BIOGRAPHICAL NOTE

Edward J. Craig was born in Springfield, Massachusetts, on July 17, 1924, the son of an uncommonly good portrait photographer with an uncommonly good sense of humor, and the son of a brilliant and lovable mother.

Perhaps the simplest way to convey a picture of Frank and Lillian Craig is to remark that if their son has done anything well in this world it was no accident.

The author was graduated from the Holy Name Grammar School in Chicopee, Mass., in 1937. He attended Cathedral High School in Springfield, Mass., for three years, transferring to the Albany High School, Albany, New York, in 1940. He was graduated from the latter in 1941 and entered Union College in September of that year.

The Army beckoned in June 1943, and over three years were spent in the Army and Army Air Corps. In November 1944 he was commissioned as a navigator on a bomber and was finally released in September 1946 as a first lieutenant.

Studies were resumed at Union College, and graduation came in 1948. He then spent a year in the Mathematics Department of Union College as an Instructor and in 1949 entered the Massachusetts Institute of Technology as a graduate student and teaching assistant. He is now serving as Assistant Professor at Northeastern University.

He has one older brother, Frank, Jr., and a younger sister, Lillian. He was married June 14, 1947 to Jeanne M. McDonald and is the father of Theresa, 5, and David, 3.

BIBLIOGRAPHY

1. Seydel, Ph., Münchener Akademische Abhandlungen, 1874, 2, Abh. p 81-108.

   or Whittaker and Robinson, *The Calculus of Observations*, p 255, sec. 130 London, Glasgow, Blackie and Son Ltd., 3rd Edition 1940.

2. Southwell, R. V., *Relaxation Methods in Engineering Science*, Oxford Univ. Press, 1940.

3. Cross, Hardy, Transactions A.S.C.E. 96, pp 1-10, discussion pp 11-138, 1932.

4. Temple, G., Proc. Roy. Soc. London (A) 169, pp 476-500, 1939.

5. Southwell, R. V., Proc. Roy. Soc. London (A) 151, pp 56-95, 1935.

6. Ham, J. M., "The Solution of a Class of Linear Operational Equations by Methods of Successive Approximations", Sc.D. Thesis, M.I.T., Chap. 3, July 1952.

7. Synge, J. L., "A Geometric Interpretation of the Relaxation Method", Quart. Journal App. Math 2, p 87, 1944.

8. Fox, Huskey, Wilkinson, Quart. Journal of Mech. and App. Math 1, p 149, 1948.

9. Lanczos, Cornelius, Journal Research, Nat. Bureau of Standards 45, pp 255-282, #4 October 1950.

10. Stiefel, Eduard, Zeitschrift für Angewandte Math. und Physik (Basel), Vol. 3, p 1-33, 1952.

11. Guillemin, E. A., *The Mathematics of Circuit Analysis*, Technology Press, John Wiley and Sons, Inc., 1949.

12. Frazer, Duncan, and Collar, *Elementary Matrices*, Cambridge Univ. Press; The MacMillan Co., New York, 1946.

13. Booth, A. D., Quart. Journal of Mech. and App. Math 2, pp 460-468, 1949.

14. Scarborough, J. B., *Numerical Math Analysis*, 2d Edition, The Johns Hopkins Press, Baltimore; Geoffrey Cumberlege, Oxford Univ. Press, London, 1950.

15. Turing, A. M., Quart. Journal of Mech. and App. Math 1, pp 287-308, 1948.

16. Turetsky, Quart. App. Math 9, pp 108-110, 1951.

17. Geiringer, H., On the Sol. of Lin. Eq. by Certain Iteration Methods, Reissner Anniversary Volume, pp 365-393, Edwards, 1949.

18. Black, A. N., Quart. Journal of Mech. and App. Math 2, pp 321-324, 1949.

19. Morris, J. Phil. Mag. Series 7, 37, p 106, 1946.

20. Hestenes, M. R., NAML Rept. 52-9, Nat. Bureau of Standards, August 1951.

21. Hestenes, M. R., and Karush, W., Jour. of Res., Nat. Bureau of Standards, 47, p 471, #6 December 1951.

22. Aitken, Proc. Roy. Soc. Ed. Sec. A 63, pp 52-60, 1950.

23. Lanczos, C., Jour. of Res., Nat. Bureau of Standards, 49, pp 33-53, #1 July 1952.

24. Stiefel, Eduard and Hestenes, M. R., Jour. of Res., Nat. Bureau of Standards, 49, pp 409-436, December 1952.

25. Schönhardt, Erich, "Über positiv definite Matrizen". Zeit. Angew. Math. und Mech. 32, p 157-158, 1952.

26. Hildebrand, F. B., Methods of Applied Math, N. Y. Prentice-Hall Inc., 1952.

27. Cauchy, A. L., "Méthode générale pour la résolution des systèms d'équations simultanées", Academie des Sciences, Paris, Comptes-Renus, 25, p 536-538, 1847.

28. Osgood, W. F., Advanced Calculus, The MacMillan Co., N. Y., 1925.

29. Tassky, O., Bibliography on bounds of characteristic roots of Matrices, NBSR #1162, 1951.

30. Paige, L. J. and Taussky, O. (Ed.), Simultaneous Linear Equations and the Determination of Eigenvalues. Nat. Bur. Stan. Applied Math. Series 29, Aug. 31, 1953. See especially sections by Rosser, J. B. and Stiefel, E.