NASA Technical Memorandum 88377

Eigensystem Analysis of Classical Relaxation Techniques With Applications to Multigrid Analysis

Harvard Lomax and Catherine Maksymiuk

March 1987

National Aeronautics and
Space Administration

NASA Technical Memorandum 88377

Eigensystem Analysis of Classical Relaxation Techniques With Applications to Multigrid Analysis

Harvard Lomax,
Catherine Maksymiuk, Ames Research Center, Moffett Field, California

March 1987

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

Table of Contents

EIGENSYSTEM ANALYSIS OF CLASSICAL RELAXATION TECHNIQUES WITH APPLICATIONS TO MULTIGRID ANALYSIS

Harvard Lomax, Catherine M. Maksymiuk
Ames Research Center

SUMMARY

Classical relaxation techniques are related to numerical methods for solution of ordinary differential equations. Eigensystems for Point–Jacobi, Gauss–Seidel, and SOR methods are presented. Solution techniques such as eigenvector annihilation, eigensystem mixing, and multigrid methods are examined with regard to the eigenstructure.

1. MATRIX FORMS OF FINITE DIFFERENCE SCHEMES

1.1 Banded Matrices

The symbol $B(M : arguments)$ is used to represent a matrix of order M, the elements of which are all zero except for those along diagonals close to the central diagonal. The number of arguments is always odd and the central argument refers to entries in the central diagonal. The M is often omitted from the argument list. Thus

$$B(b_{-2}, b_{-1}, b_0, b_1, b_2) \tag{1.1.1}$$

represents a matrix with scalar entries that are constant along the central diagonal and the two diagonals just above and just below it. This is a pentadiagonal matrix. A form of banded matrix that is very common in numerical analysis is the tridiagonal system illustrated by

$$B\left(M : \vec{a},\vec{b},\vec{c}\right) \equiv \begin{bmatrix} b_1 & c_2 & & & & \\ a_1 & b_2 & c_3 & & & \\ & a_2 & b_3 & & & \\ & & & \ddots & & \\ & & & & & c_M \\ & & & & a_{M-1} & b_M \end{bmatrix} \tag{1.1.2}$$

In this case the elements vary along the diagonals and are represented by vectors. The particular indexing of the elements, constant along columns, is used in so–called conservation–law formulations.

Banded matrices can be used to represent finite difference schemes, in which case the nature of the boundary conditions is signified by the matrix formulation. Eq (1.1.2) would be used if the boundary conditions were Dirichlet. If periodic boundary conditions were desired, the appropriate form would be

$$B_p\left(M : \vec{a},\vec{b},\vec{c}\right) \equiv \begin{bmatrix} b_1 & c_2 & & & & a_M \\ a_1 & b_2 & c_3 & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & \cdot \\ & & & & \cdot & c_M \\ c_1 & & & & & b_M \end{bmatrix} \tag{1.1.3}$$

An important set of banded matrices with constant diagonal entries is represented by the following expression

$$B_p(a,b,c) \equiv \begin{bmatrix} b & c & & & & & a \\ a & b & c & & & & \\ & a & b & & & & \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & & \cdot & \\ & & & & & \cdot & c \\ c & & & & & a & b \end{bmatrix} \tag{1.1.4}$$

This particular matrix is referred to as a tridiagonal circulant matrix. Circulant matrices need not be tridiagonal; they can be completely dense. Notice that a circulant matrix is a special form of a periodic matrix.

The notation I is used for the identity matrix and the notation $D(b)$, is used for a diagonal matrix with constant elements. Notice that

$$B(b) = D(b) \qquad \text{and} \qquad B(1) = D(1) = I \tag{1.1.5}$$

1.2 Difference Schemes as Banded Matrices

A difference scheme is usually written as a point operator. The three-point central-difference scheme for a second derivative is given by

$$(\delta_{xx}u)_j = \frac{1}{\Delta x^2}(u_{j+1} - 2u_j + u_{j-1}) \tag{1.2.1}$$

where the index j refers to the location of the dependent variable in the equispaced x direction. This difference scheme can be expressed as a matrix operator if the boundary conditions are specified. If u is specified at a boundary, the $(\vec{bc})$ (boundary condition) is referred to as Dirichlet, and if $\frac{\partial u}{\partial n}$ is specified at a boundary, the $(\vec{bc})$ is referred to as Neumann. Figure (1.2.1) shows three possibilities. In Fig. (1.2.1a) a mesh with 4 interior points is shown. The value of u is specified at the two end points a and b, so the $(\vec{bc})$ are said to be Dirichlet. Figure (1.2.1b) illustrates a situation representing a Neumann $(\vec{bc})$ for the right side and a Dirichlet $(\vec{bc})$ for the left side. Periodic $(\vec{bc})$ are illustrated in Fig. (1.2.1c). Notice that the normalized length of the mesh is π in the top case, $\pi/2$ in the middle case, and 2π in the bottom case. These conventions fix a condition between the space step size and the number of points in a mesh which is convenient in later developments.

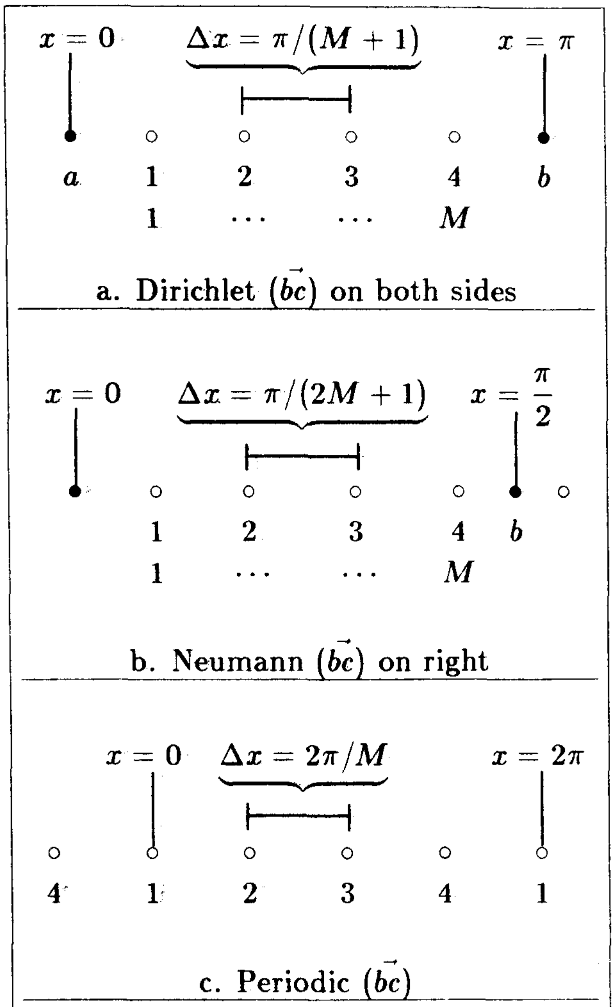


Figure 1.2.1 - Boundary conditions.

The matrix representation of eq (1.2.1) for the three kinds of $(\vec{bc})$ illustrated can now be written. For the complete Dirichlet problem illustrated in Fig. (1.2.1a)

$$\begin{aligned}\delta_{xx}\vec{u} &= \frac{1}{\Delta x^2}\left[B(1,-2,1)\vec{u} + \left(\vec{bc}\right)\right] \\ \left(\vec{bc}\right) &= [u_a, 0, \cdots, 0, u_b]^T\end{aligned} \qquad (1.2.2)$$

For the mixed Neumann–Dirichlet problem in Fig. (1.2.1b)

$$\left.\begin{aligned}\delta_{xx}\vec{u} &= \frac{1}{\Delta x^2}\left[B\left(1,-\vec{b},1\right)\vec{u}+\left(\vec{bc}\right)\right]\\ \vec{b} &= [2,2,2,1]^T\\ \left(\vec{bc}\right) &= \left[u_a,0,0,\Delta x\left(\frac{\partial u}{\partial x}\right)_b\right]^T\end{aligned}\right\} \tag{1.2.3}$$

It should be noted that for the Neumann $(\vec{bc})$ the value of $\frac{\partial u}{\partial x}$ is assumed to given at the point $M+\frac{1}{2}\Delta x$. Futhermore this particular scheme is only first order accurate, although the multiplying constant is small. Finally, for periodic $(\vec{bc})$ shown in Fig. (1.2.1c) we have the relation

$$\delta_{xx}\vec{u} = \frac{1}{\Delta x^2}B_p(1,-2,1)\vec{u} \quad . \tag{1.2.4}$$

Notice that the periodic case has no $(\vec{bc})$ and the expression is homogeneous.

2. FORMULATION OF THE MODEL PROBLEM

2.1 The Basic Equation and its Solution

Is is assumed that some time dependent partial differential equation represents some valid fluid flow. It is further assumed that this equation has proper boundary conditions and that appropriate differencing schemes have been applied to approximate the space derivatives and the boundary conditions. The result is representable in matrix form as

$$\frac{d\vec{u}}{dt} + A_b\vec{u} - \vec{f}_b = 0 \tag{2.1.1}$$

where $\vec{u}$ represents the dependent variables, $\vec{f}_b$ holds the boundary conditions and the forcing function (if there is one), and A_b is usually nonlinear (i.e., depends on $\vec{u}$). Now it is assumed that a solution to this equation exists for which $\vec{u}$ is time invariant. We refer to this as a steady state solution of eq (2.1.1). Such a solution satisfies the equations

$$A_b\vec{u} - \vec{f}_b = 0 \tag{2.1.2a}$$

$$\text{or} \qquad \vec{u} = A_b^{-1}\vec{f}_b \tag{2.1.2b}$$

where A_b^{-1} is assumed to exist. None of the questions of existence and well-posed boundary conditions are considered in this paper. Finally, it is assumed that the solution given by eq (2.1.2) is going to be found by an iterative process, and we wish to study methods for carrying out these iterations.

The above was written to treat the general case. It is instructive in formulating the concepts to consider the special case given by the linear diffusion equation in one dimension

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - p(x) \tag{2.1.3}$$

This has the steady state solution

$$\frac{\partial^2 u}{\partial x^2} = p(x) \tag{2.1.4}$$

which is the one dimensional form of the Poisson equation. Introducing the three-point central differencing scheme for the second derivative, eq (1.2.1), we find

$$\frac{d\vec{u}}{dt} = \frac{1}{\Delta x^2}B(1,-2,1)\vec{u} + (\vec{bc}) - \vec{p} \tag{2.1.5}$$

where $(\vec{bc})$ contains the boundary conditions. In this case

$$\begin{aligned} A_b &= \frac{1}{\Delta x^2}B(1,-2,1) \\ \vec{f}_b &= \vec{p} - (\vec{bc}) \end{aligned} \tag{2.1.6}$$

We work extensively with this model form.

2.2 Preconditioning the Basic Matrix

The iteration process mentioned above is referred to as a relaxation procedure. It is standard practice in applying relaxation procedures to precondition the basic equation. This preconditioning has the effect of multiplying eq (2.1.2a) from the left by some nonsingular matrix. In the simplest possible case the conditioning matrix is a diagonal matrix composed of a constant $D(b)$, see eq (1.1.5). If we designate the conditioning matrix by C, the problem becomes one of solving for $\vec{u}$ in

$$CA_b\vec{u} - C\vec{f}_b = 0 \tag{2.2.1}$$

Notice that the solution of eq (2.2.1) is

$$\vec{u} = [CA_b]^{-1}C\vec{f}_b = A_b^{-1}C^{-1}C\vec{f}_b = A_b^{-1}\vec{f}_b \tag{2.2.2}$$

which is identical to the steady-state solution of eq (2.1.1), provided C^{-1} exists.

In the following we will see that our approach to the iterative solution of eq (2.2.1) depends crucially on the eigenvalue and eigenvector structure of the matrix $[CA_b]$, and, equally important, does not depend at all on the eigensystem of the basic matrix A_b. A simple example illustrates the point. Consider the first-order backward difference scheme (If properly applied this can also be referred to as an "upwind" scheme).

$$(\delta_x u) = \frac{1}{\Delta x}(u_j - u_{j-1}) \tag{2.2.3}$$

If $\vec{u}$ is specified on the boundary at the left (u_a given in Fig. 1.2.1a), the difference scheme forms the matrix operation

$$\delta_x \vec{u} = \frac{1}{\Delta x}\left[B(-1,1,0)\vec{u} + \left(\vec{bc}\right)\right] \tag{2.2.4}$$

With this operator, the simple relation

$$\frac{\partial u}{\partial x} = g(x) \qquad \text{or} \qquad u = \int_0^x g(x_1)dx_1 \tag{2.2.5}$$

can be approximated by

$$\frac{1}{\Delta x}B(-1,1,0)\vec{u} + \frac{1}{\Delta x}\left(\vec{bc}\right) = \vec{g} \tag{2.2.6}$$

$$\text{where} \qquad \left(\vec{bc}\right) = [-u_a, 0, \cdots, 0]^T$$

Eq (2.2.6) takes the form of eq (2.1.2a) if

$$A_b = \frac{1}{\Delta x}B(-1,1,0) \qquad \text{and} \qquad \vec{f}_b = \vec{g} - \frac{1}{\Delta x}\left(\vec{bc}\right) \tag{2.2.7}$$

The eigensystem of A_b is fully defective.[1] However, let $C = -\Delta x \cdot B(0,1,-1)$. The resulting product matrix $[CA_b]$, see eq (2.2.9), has a complete set of eigenvectors and the eigenvalues are all real and negative. The relaxation of eq (2.2.1) with this choice of C and A_b can proceed along well-defined classical lines in spite of the fact that the original basic matrix had no resemblance to the classical form.

Finding the solution to eq (2.2.6) is a trivial matter, but that is not our object. Our object is to use eq (2.1.2) and (2.2.7) to illustrate a general iterative solution process. This process consists of first, choosing C so that $[CA_b]$ is "close" to a model matrix given in the next section, and second, making a thorough analysis of the various iteration techniques available for the model matrices. There are well-known techniques for accelerating relaxation schemes if the eigenvalues of $[CA_b]$ are all real and of the same sign. This paper is limited to a study of only these techniques. Such a limitation imposes the requirement:

> The conditioning matrix C is chosen such that the eigenvalues of $[CA_b]$ are all real and negative, that is, such that $[CA_b]$ is negative definite. (2.2.8)

[1] See Section 4.7

It is also well known that a choice of C which ensures this condition is the negative transpose of A_b. In the above example given by eq (2.2.7), $B^T(-1,1,0) = B(0,1,-1)$ and

$$-B(0,1,-1)B(-1,1,0) = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix} \qquad (2.2.9)$$

which is the same matrix that arises for a second-derivative approximation with mixed Dirichlet–Neumann boundary condition, see eq (1.2.3) and (5.6.2). It is easy to verify that if, instead of $C = -B(0,1,-1)$, C is chosen to be

$$C = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & 1 & -1 & \\ & & & 1 & -1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix} \qquad (2.2.10)$$

the product $C \cdot B(-1,1,0)$ is equal to $B(1,-2,1)$, a finite-difference matrix for a second derivative with Dirichlet conditions on both sides, see eq (5.6.1).

Another interesting example arises from the study of first-order partial differential equations when a central differencing (rather than upwind, as in eq (2.2.3)) scheme is used for the approximation of the derivative. However, the physics of this problem permits a Dirichlet $\left(\vec{bc}\right)$ on one end but allows no constraint on the other; so on one end (and we assume it is the right side) we use a first order upwind scheme. This leads to the matrix difference equation

$$\delta_x \vec{u} = \frac{1}{2\Delta x} \left\{ \begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & -1 & 0 & 1 & \\ & & -1 & 0 & 1 \\ & & & -2 & 2 \end{bmatrix} \vec{u} + \begin{bmatrix} u_a \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\} \qquad (2.2.11)$$

The matrix in eq (2.2.11) can first be conditioned so that the modulus of each element is 1, and then further conditioned with multiplication by the negative transpose. The result is

$$A_2 = -A_1^T A_1 =$$

$$\begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & -1 & 0 & 1 & \\ & & -1 & 0 & 1 \\ & & & -1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & -1 & 0 & 1 & \\ & & -1 & 0 & 1 \\ & & & -1 & 1 \end{bmatrix}$$

$$- \begin{bmatrix} -1 & 0 & 1 & & \\ 0 & -2 & 0 & 1 & \\ 1 & 0 & -2 & 0 & 1 \\ & 1 & 0 & -2 & 1 \\ & & 1 & 1 & -2 \end{bmatrix} \qquad (2.2.12)$$

The matrix on the right side of eq (2.2.12) does not look familiar, but if we define a permutation matrix P and carry out the process $P^T[-A_1^T A_1]P$ we find

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 & & \\ 0 & -2 & 0 & 1 & \\ 1 & 0 & -2 & 0 & 1 \\ & 1 & 0 & -2 & 1 \\ & & 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix} \qquad (2.2.13)$$

which leads to exactly the same matrix as that derived from upwind differencing presented in eq (2.2.9).

The importance of these concepts is much more evident when they are used to precondition the Cauchy-Riemann and Euler equations. In these cases even when the basic matrix A_b has nearly imaginary eigenvalues, as it does if central difference schemes are used for the first derivatives, the conditioned matrix $[-A_b^T A_b]$ is nevertheless negative definite and the study of the model equations in the next section is pertinent to its solution.

2.3 The Model Equations

Preconditioning processes such as those described in the last section allow us to prepare our algebraic equations in advance so that certain eigenstructures are guaranteed. In the remaining part of this paper we wish to thoroughly investigate some simple equations that model these structures. There is absolutely nothing new in these model equations; they (especially the two-dimensional case) are used in all the standard texts treating the subject of relaxation. However, the approach to their solution, i.e., as a subset of the theory of ordinary differential equations, is novel in some places, and the picture of the subject from this point of view is helpful in unifying a large number of relaxation methods that have been proposed.

Consider the preconditioned iterative equation having the form

$$\frac{d\vec{\phi}}{dt} + A\vec{\phi} - \vec{f} = 0 \tag{2.3.1}$$

where A is negative definite and the symbol for the dependent variable has now been changed to ϕ signifying that the physics being modeled is no longer time accurate. However, notice that a steady state of eq (2.3.1) is guaranteed to exist (because it is negative definite) and that steady state solution is $A^{-1}\vec{f}$,which is identical to eq (2.1.2). In the notation of eq (2.1.2) and (2.2.1)

$$A = CA_b \qquad \text{and} \qquad \vec{f} = C\vec{f}_b \tag{2.3.2}$$

For the model equations, in the one-dimensional case A has the form

$$\begin{gathered} A = B\left(1, \vec{b}, 1\right) \\ \vec{b} = \left[-2, -2, \cdots, -2, s\right]^T \\ s = -2 \qquad \text{or} \qquad -1 \end{gathered} \tag{2.3.3}$$

For the two-dimensional case, A has the form

$$\begin{gathered} A = B[\alpha_y I, B(\alpha_x, -4, \alpha_x), \alpha_y I]; \\ \alpha_x + \alpha_y = 2 \quad ; \quad \alpha_x \leq 2 \end{gathered} \tag{2.3.4}$$

A tremendous amount of insight to the basic features of relaxation is gained by an appropriate study of the one-dimensional case, and much of the remaining material is devoted to this case. We attempt to do this in such a way, however, that it is directly applicable to two- and three-dimensional problems.

3. THE DELTA FORM OF AN ITERATION SCHEME

3.1 Basic Theory

We assume that our basic difference matrix has been preconditioned to form

$$A\vec{\phi} - \vec{f} = 0 \qquad (3.1.1)$$

and thatA is negative definite. We choose some iterative scheme to find the solution, $A^{-1}\vec{f}$, and designate the iteration count by the subscript n or the superscript (n). The converged solution is designated $\vec{\phi}_\infty$ so that

$$\vec{\phi}_\infty = A^{-1}\vec{f} \qquad (3.1.2)$$

The manner in which convergence is measured in actual practice is not considered here. (It is generally related to the magnitude of a residual, e.g. $\sum |A\vec{\phi} - \vec{f}|$, summed at each point in the mesh.) Instead we assume that after some step N, the solutions $\phi_N, \phi_{N+1}, \cdots, \phi_{N+k}$ are all "close enough" that each could be considered the final answer. That is to say, for our purposes, one can write $\phi_N = \phi_{N+1} = \cdots = \phi_{N+k}$. Suppose $F(z)$ is some function having the property $F(0) = 0$. Then the general expression for the "delta form" used to relax eq (3.1.1) can be written

$$F\left(\vec{\phi}_{n+k} - \vec{\phi}_{n+l}\right) = A\vec{\phi}_n - \vec{f}; \begin{cases} k = 0, 1, 2, \cdots \\ l = 0, 1, 2, \cdots \end{cases} \qquad (3.1.3)$$

If F is a linear operator, one can easily prove by the theory of finite difference equations that the particular solution of eq (3.1.3) is our desired solution $A^{-1}\vec{f}$, if $\vec{f}$ is independent of n. It is only necessary then to formulate F in such a way that the complementary solution of the difference equation (3.1.3) goes to zero as $n \to \infty$.

3.2 Examples

Consider the model equation

$$B(1, -2, 1)\vec{\phi} - \vec{f} = 0 \qquad (3.2.1)$$

One example of a delta form, written as a point operator, is

$$a\left(\phi_j^{(n+1)} - \phi_j^{(n-1)}\right) + b\left(\phi_j^{(n+1)} - 2\phi_j^{(n)} + \phi_j^{(n-1)}\right) = \phi_{j-1}^{(n)} - 2\phi_j^{(n)} + \phi_{j+1}^{(n)} - f_j \tag{3.2.2}$$

Notice that the left side sums to zero if $\phi_j^{(n)}$ converges. If we set $b = 1$ and $a = 1/(2h)$, eq (3.2.2) reduces to

$$\phi_j^{(n+1)} - \phi_j^{(n-1)} = 2h\left(\phi_{j-1}^{(n)} - \phi_j^{(n+1)} - \phi_j^{(n-1)} + \phi_{j+1}^{(n)} - f_j\right) \tag{3.2.3}$$

which is a representation of the DuFort–Frankel method used to solve the diffusion equation, see e.g. ref. 1, p. 60. Notice also that eq (3.2.2) can be interpreted in the form of an ordinary differential equation

$$a\frac{d\vec{\phi}}{dt} + b\frac{d^2\vec{\phi}}{dt^2} = B(1, -2, 1)\vec{\phi} - \vec{f}(x) \tag{3.2.4}$$

if the iteration index is thought of as a "time" displacement.

A second example of a delta form written as a point operator for eq (3.2.1) is

$$a\left(\phi_j^{(n+1)} - \phi_j^{(n)} - \phi_{j-1}^{(n+1)} + \phi_{j-1}^{(n)}\right) + b\left(\phi_j^{(n+1)} - \phi_j^{(n)}\right) = \phi_{j-1}^{(n)} - 2\phi_j^{(n)} + \phi_{j+1}^{(n)} - f_j \tag{3.2.5}$$

Again we see that the left side sums to zero at convergence. This time, however, the expression can be interpreted in the form of a partial differential equation

$$a\frac{\partial^2\phi}{\partial x\partial t} + b\frac{\partial\phi}{\partial t} = \frac{\partial^2\phi}{\partial x^2} - f(x) \tag{3.2.6}$$

This approach to relaxation has been used by Garabedian, see e.g. ref. 1, p. 125.

3.3 The Ordinary Differential Equation Formulation

The particular type of delta form considered in the remaining part of this paper is one that leads to a differential interpretation composed of a set of coupled, first-order, ordinary, differential equations. In difference notation it is expressed as

$$H\left[\vec{\phi}_{n+1} - \vec{\phi}_n\right] = A\vec{\phi}_n - \vec{f} \tag{3.3.1}$$

where H is some nonsingular matrix which is independent of n for "stationary" methods and is a function of n for "nonstationary" ones. In differential notation it can be written

$$H\frac{d\vec{\phi}}{dt} = A\vec{\phi}_n - \vec{f}_n \tag{3.3.2}$$

Since the basic matrix A_b has already been conditioned by the matrix C to produce $A = CA_b$ and $\vec{f} = C\vec{f}_b$, the secondary conditioning matrix H may seem superfluous. We find it convenient, however, to separate the relaxation procedure into two distinct parts. First, we precondition the basic equation to put it in a model form. Second, we further condition the model form to generate optimum algorithms. For this reason both C and H play a useful role.

4. THE ANALYTICAL AND NUMERICAL SOLUTION OF FIRST-ORDER ORDINARY DIFFERENTIAL EQUATIONS WITH CONSTANT COEFFICIENTS AND A CONSTANT FORCING FUNCTION

4.1 Introduction

One purpose of this report is to present the subject of relaxation as a subset of the theory of the numerical solution of ODE. That concept in itself is certainly not new. However, a systematic treatment of such an approach has not, to the author's knowledge, been published, and it leads to some interesting and useful interpretations. In order to make the discussion clear, a review of the theory for the analytical and numerical solution of ODE is given.

4.2 The Analytical Solution of ODE

We consider next the analytical solution of a set of coupled first-order, ordinary differential equations given by

$$\frac{d\vec{u}}{dt} = [\mathrm{A}]\vec{u} - \vec{f} \tag{4.2.1}$$

where both $[\mathrm{A}]$ and $\vec{f}$ are independent of $\vec{u}$ and t. We first consider only those cases for which the eigenvectors of $[\mathrm{A}]$ are linearly independent. Introduce the left and right eigenvector matrices X^{-1} and X such that

$$X^{-1}AX = D\left(\vec{\lambda}\right) = \Lambda \tag{4.2.2}$$

where Λ is a diagonal matrix having the eigenvalues of $[\mathrm{A}]$ as the entries. When we speak of a single eigenvector of $[\mathrm{A}]$, say $\vec{x}_m$, we are referring to a column in X corresponding to a λ_m in Λ. Linear independence of the eigenvectors means that $a \cdot \vec{x}_m + b \cdot \vec{x}_n \neq \vec{x}_k$ for any values of (complex) a and b, and where $m \neq n \neq k$. To solve eq (4.2.1) we multiply each term from the left by X^{-1} and insert the identity matrix $X^{-1}X = I$ between $[\mathrm{A}]$ and $\vec{u}$. There results

$$X^{-1}\frac{d\vec{u}}{dt} = X^{-1}AXX^{-1}\vec{u} - X^{-1}\vec{f} \tag{4.2.3}$$

which reduces to

$$\frac{d}{dt}\left(X^{-1}\vec{u}\right) = \Lambda\left(X^{-1}\vec{u}\right) - \left(X^{-1}\vec{f}\right) \tag{4.2.4}$$

Define

$$\vec{w} = X^{-1}\vec{u} \quad ; \quad \vec{g} = X^{-1}\vec{f} \tag{4.2.5}$$

and the solution can be written

$$\frac{d\vec{w}}{dt} = D(\vec{\lambda})\vec{w} - \vec{g} \tag{4.2.6}$$

Eqs (4.2.1) and (4.2.6) are expressing the same equality but in different algebraic forms. The key difference is that eq (4.2.6) is completely uncoupled. It can be written line by line as a set of first-order differential equations (defining w' as $\frac{dw}{dt}$)

$$\left.\begin{aligned} w_1' &= \lambda_1 w_1 - g_1 \\ w_2' &= \lambda_2 w_2 - g_2 \\ &\vdots \\ w_m' &= \lambda_m w_m - g_m \\ &\vdots \end{aligned}\right\} \tag{4.2.7}$$

each of which can be solved separately. The resulting solutions can then be recoupled, using the inverse of eqs (4.2.5), to form the solution to eq (4.2.1). When the forcing function, $\vec{g} = X^{-1}\vec{f}$, is not a function of t, the solution of the mth equation in (4.2.7) is

$$w_m = c_m e^{\lambda_m t} + \frac{1}{\lambda_m} g_m \tag{4.2.8}$$

Tracing the algebra backward, one can easily show that the solution to eq (4.2.1) can be written

$$\vec{u} = X\overrightarrow{\left(ce^{\lambda t}\right)} + X\Lambda^{-1}X^{-1}\vec{f} \tag{4.2.9}$$

where $\overrightarrow{(\quad)}$ denotes a vector of the enclosed terms. Thus

$$\overrightarrow{\left(ce^{\lambda t}\right)} \equiv \left[c_1 e^{\lambda_1 t}, c_2 e^{\lambda_2 t}, \cdots, c_M e^{\lambda_M t}\right]^T \tag{4.2.10}$$

Another way to express the solution given by eq (4.2.9) is

$$\vec{u} = \underbrace{c_1 e^{\lambda_1 t}[\vec{x}_1] + c_2 e^{\lambda_2 t}[\vec{x}_2] + \cdots + c_M e^{\lambda_M t}[\vec{x}_M]}_{\text{Transient solution}} + \underbrace{[\mathbf{A}]^{-1}\vec{f}}_{\text{Steady state}} \tag{4.2.11}$$

Here in classical terminology $[\mathbf{A}]^{-1}\vec{f}$ designates the particular solution to the complete equation and the remaining terms are the complementary solution to the homogeneous equation, $d\vec{u}/dt = [\mathbf{A}]\vec{u}$. In fluid flow problems $[\mathbf{A}]^{-1}\vec{f}$ is often referred to as the steady state solution and the complementary solution is often referred to as the transient solution.

Given a linear system, the two possible formulations of the same problem discussed above and given by eqs (4.2.1) and (4.2.7) play an important role in our following development. For that reason we give them special names so we can quickly extract the concept in a given situation. This is summarized in the following:

$$\frac{d\vec{u}}{dt} = [\mathbf{A}]\vec{u} - \vec{f}(t) \tag{4.2.12a}$$

is referred to as an equation in real space,

$$\frac{d\vec{w}}{dt} = [\Lambda]\vec{w} - \vec{g}(t) \tag{4.2.12b}$$

is referred to as an equation in eigen space (also referred to as wave space)

4.3 The Isolation Theorem and the Representative Equation

Let us summarize the results of the previous section in order to simplify the discussion in the next section which treats the numerical solution of ODE. The fact that eqs (4.2.1) and (4.2.7) express the same equality is the basis for the Isolation Theorem which is stated next.

Given a set of first order ordinary differential equations such that

- The coupled equations are linear with time-constant coefficients.
- The eigensystem of the $[\mathbf{A}]$ matrix is not defective.

(4.3.1)

one can develop:

The Isolation Theorem[1]

> Applying any standard time-march method to each equation in a coupled set of equations having the constraints specified in (4.3.1) is mathematically identical to:
>
> - Uncoupling the set including the forcing terms,
> - Individually integrating each equation in the uncoupled set,
> - Recoupling the group to form the final result.
>
> (4.3.2)

In the terminology of (4.2.12) this amounts to the observation that (under the conditions stated in (4.3.1)) using any time-march method on the coupled equations in real space is identical to using the same method on the single equation in eigen space.

The theorem uses the terminology "mathematically identical to" which is rigorously correct. Unfortunately it is not strictly correct for the statement "numerically identical to" because of roundoff error. Numerical experiments with simple eigensystems are easy to construct and quite informative in verifying the substance of the theorem.

The "single equation in eigen space", mentioned above, has the form shown by eq (4.2.7). We simplify this to

$$\frac{dw}{dt} = \lambda w + a \tag{4.3.3}$$

where λ and a are (complex) constants. Notice that we only consider the case for a forcing function that is independent of time, since that is sufficient when our application is in the field of relaxation. We refer to eq (4.3.3) as the representative equation.

[1] Proof of this theorem is a simple exercise using the concepts outlined in Section 4.2.

4.4 The Numerical Solution of ODE

Now we must choose some numerical procedure to integrate the representative eq (4.3.3), knowing from the isolation theorem that that method will advance each eigenvector in eq (4.2.11) according to its associated eigenvalue. These methods are referred to as time–marching methods and they convert the representative ODE into an ordinary difference equation (OΔE), or a set of coupled OΔE, depending upon the choice of method. These OΔE can be solved analytically and their solution provides the result required to evaluate the method and compare it with other time-marching methods.

First let us review the process for finding the solution of coupled first–order OΔE that are linear and have constant (with respect to n) coefficients. These can be expressed in the form

$$\vec{u}_{n+1} = [\mathrm{C}]\vec{u}_n - \vec{\mathcal{f}} \tag{4.4.1}$$

following a procedure exactly the same as that used to express eq (4.2.1) in the form of eq (4.2.7), we can re–express eq (4.4.1) as

$$\begin{aligned} (w_{n+1})_1 &= \sigma_1(w_n)_1 - \mathcal{G}_1 \\ (w_{n+1})_2 &= \sigma_2(w_n)_2 - \mathcal{G}_2 \\ &\vdots \\ (w_{n+1})_m &= \sigma_m(w_n)_m - \mathcal{G}_m \\ &\vdots \end{aligned} \tag{4.4.2}$$

where the σ are the eigenvalues of $[\mathrm{C}]$ (which is assumed to be nondegenerate). This represents a set of uncoupled first–order OΔE each of which has the form

$$w_{n+1} = \sigma w_n - \mathcal{G} \tag{4.4.3}$$

This simple first-order difference equation has the solution

$$w_n = c(\sigma)^n - \frac{\mathcal{G}}{1-\sigma} \tag{4.4.4}$$

which can be verified by substitution.

Let us consider an example of how these results can be used to analyze a numerical time–marching method. We introduce some notation for the discrete variables, thus

$$
\begin{aligned}
t = t_n &= nh \quad ; \quad h \equiv \Delta t \\
u_{n+k} &= u(t_n + kh) \\
u'_{n+\ell} &= \frac{d}{dt} u(t_n + \ell h)
\end{aligned}
\tag{4.4.5}
$$

The numerical time–march method is described by a linear combination of the terms u_{n+k} and $u'_{n+\ell}$ where $k, \ell = \pm 0, 1, 2, \ldots$

The simplest example of a numerical time–marching method is the explicit Euler scheme which can be written

$$
\vec{u}_{n+1} = \vec{u}_n + h\vec{u}'_n \tag{4.4.6}
$$

Applying this to the coupled set of ODE represented by eq (4.2.1) gives

$$
\begin{aligned}
\vec{u}_{n+1} &= \vec{u}_n + h[\mathrm{A}]\vec{u}_n - h\vec{f} \\
&= [I + h[\mathrm{A}]]\vec{u}_n - h\vec{f}
\end{aligned}
\tag{4.4.7}
$$

Comparing this with eq (4.4.1) we see that for the explicit Euler method

$$
\begin{aligned}
[\mathrm{C}] &= [I + h\mathrm{A}] \\
\vec{\mathcal{f}} &= h\vec{f}
\end{aligned}
\tag{4.4.8}
$$

First of all we notice that, since the identity matrix $[I]$ commutes with any matrix and h is a scalar, the eigenvectors of $[\mathrm{C}]$ and $[\mathrm{A}]$ are identical. From this it follows that

$$
\sigma_m = 1 + \lambda_m h \tag{4.4.9}
$$

Further, it can be shown (combining eqs (4.4.8) with eq (4.4.3) and recoupling the systems) that the particular solution of the OΔE produced by the explicit Euler scheme is

$$
\vec{u}_n = c_1(\sigma_1)^n \vec{x}_1 + c_2(\sigma_2)^n \vec{x}_2 + \cdots + c_M(\sigma_M)^n \vec{x}_M + [\mathrm{A}]^{-1}\vec{f} \tag{4.4.10}
$$

where $[\mathrm{A}]^{-1}\vec{f}$ is the <u>exact</u> <u>solution</u> of the ODE, and the eigenvectors $\vec{x}_m$ are the eigenvectors of $[\mathrm{A}]$.

4.5 The Analytic Solution of O∆E

For completeness we discuss briefly the analytic solution of linear difference equations with constant coefficients. For background, we recall that one classical approach to the study of linear ODE is carried out in terms of the operator D where

$$D^n \equiv \frac{d^n}{dt^n} \tag{4.5.1}$$

The basic part of that solution process consists of replacing the time derivatives with the appropriate power of D and thereby deriving a characteristic polynomial, denoted $P(D)$. This is followed by finding the roots, λ_m, of $P(\lambda) = 0$. The homogeneous part of the solution is then characterized by the expression

$$u(t) = \sum_m c_m e^{\lambda_m t} \tag{4.5.2}$$

where the c_m are determined from the initial conditions. If $u(t)$ is a vector, the terms on the right would be multiplied by the associated eigenvectors $\vec{x}_m$ as in eq (4.2.1).

A similar process can be carried out for linear O∆E , except that the D operator is replaced by the operator E, referred to as the displacement or shift operator, and defined in terms of the D operator by

$$E \equiv e^{hD} = 1 + hD + \frac{1}{2}h^2D^2 + \cdots \tag{4.5.3}$$

From this definition it should be clear that

$$u_{n+k} = E^k u_n \quad b^{n+\alpha} = E^\alpha b_n \quad u_{n+\frac{1}{2}} = E^{\frac{1}{2}} u_n \tag{4.5.4}$$

where k and α can be integer, fractional, or even irrational numbers. Again the basic part of the solution process consists of finding a characteristic polynomial, this time designated $P(E)$, and then finding the roots of $P(\sigma) = 0$. In this case for O∆E the homogeneous part of the solution is characterized by

$$u_n = \sum_m c_m(\sigma_m)^n \tag{4.5.5}$$

where σ_m are the roots of the characteristic polynomial and the c_m are determined by the initial conditions. If u_n is a vector, the terms on the right would be multiplied by the associated eignevectors as in eq (4.4.10).

As a simple example, consider the process for finding the solution to the equation

$$u_{n+2} + a_1 u_{n+1} + a_0 u_n = b^{n+\alpha} + b^n \tag{4.5.6}$$

In operational form this becomes

$$P(E)u_n = Q(E)b^n \tag{4.5.7}$$

where

$$\begin{aligned} P(E) &= E^2 + a_1 E + a_0 \\ Q(E) &= E^\alpha + 1 \end{aligned} \tag{4.5.8}$$

We refer to $Q(E)$ as the particular polynomial since it serves to determine the particular solution. The general solution of eq (4.5.7), which can be derived using Boole's first rule (see ref. 2), is

$$u_n = \sum_m c_m (\sigma_m)^n + b^n \frac{Q(b)}{P(b)} \tag{4.5.9a}$$

which reduces to

$$u_n = \sum_m c_m (\sigma_m)^n + \frac{Q(1)}{P(1)} \tag{4.5.9b}$$

if the right side is independent of n (i.e. if $b = 1$). The general solution for the polynomials given by eq (4.5.8) is

$$u_n = c_1 \left(\frac{-a_1 + \sqrt{a_1^2 - 4a_0}}{2} \right)^n + c_2 \left(\frac{-a_1 - \sqrt{a_1^2 - 4a_0}}{2} \right)^n + \frac{b^\alpha + 1}{b^2 + a_1 b + a_0} \cdot b^n \tag{4.5.10}$$

4.6 The Analysis of Time–Marching Methods

One can apply the analysis in Section 4.5 to investigate all manner of time–marching methods (e.g. Runge–Kutta, predictor corrector, and multistep methods) as they apply to the representative equation (4.3.3). Consider the first–order explicit Euler method for a scalar equation

$$u_{n+1} = u_n + hu'_n \tag{4.6.1}$$

Applied to the representative equation it gives

$$u_{n+1} = (1 + \lambda h)u_n + ha \tag{4.6.2}$$

and this produces eq (4.5.7) where

$$\begin{aligned} P(E) &= E - 1 - \lambda h \\ Q(E) &= ah \end{aligned} \tag{4.6.3}$$

Using the results of eq (4.5.9b), we find

$$u_n = c(1 + \lambda h)^n - \frac{a}{\lambda} \tag{4.6.4}$$

Notice that the steady–state solution is the exact steady–state solution of the ODE and the σ root approximates the Taylor series expansion of $\left(e^{\lambda h}\right)$ through the order of the method.

This σ, λ relation can be established for any linear time–march scheme such as Adams–Moulton and Runge–Kutta, see e.g., ref. 3. It is fundamental to the approach we present later because the convergence of the relaxation methods depends on $|\sigma|$, and the value of σ is some function of the product λh. The exact nature of the functional dependence of σ on λh is determined by the choice of the differencing scheme which fixes the characteristic polynomial $P(E)$. If the time–march method produces only one σ for each λ, it is referred to as a one–root method. Many methods generate more than one σ for each λ, in which case one of the σ approximates $e^{\lambda h}$ and the others are referred to as spurious. The subject of spurious roots is outside the scope of this report. Some examples of the σ, λ relation are given in Table 4.6.1.

TABLE 4.6.1 - THE σ, λ RELATION FOR SOME WELL–KNOWN NUMERICAL METHODS

Method	σ, λ relation
Runge–Kutta of Order:	
First	$\sigma = 1 + \lambda h$
Second	$\sigma = 1 + \lambda h + \frac{1}{2}\lambda^2 h^2$
N^{th}	$\sigma = \sum_0^N \frac{1}{n!}(\lambda h)^n$
Leapfrog	$\sigma_{1,2} = \lambda h \pm \sqrt{1 + \lambda^2 h^2}$
3-pt Adams–Bashforth	$\sigma_{1,2} = \frac{1}{2}\left[1 + \frac{3}{2}\lambda h \pm \sqrt{1 + \lambda h + \frac{9}{4}\lambda^2 h^2}\right]$
Implicit Euler	$\sigma = 1/(1 - \lambda h)$
Trapezoidal method	$\sigma = \left(1 + \frac{1}{2}\lambda h\right)/\left(1 - \frac{1}{2}\lambda h\right)$

The importance of the σ, λ relation is brought out very clearly by comparing eqs (4.2.11) and (4.4.10), which are repeated here (using $t = nh$) for emphasis.

Analytical solution of ODE

$$\vec{u} = \underbrace{c_1(e^{\lambda_1 h})^n \vec{x}_1 + c_2(e^{\lambda_2 h})^n \vec{x}_2 + \cdots + c_M(e^{\lambda_M h})^n \vec{x}_M}_{\text{Transient solution}} + \underbrace{[\mathrm{A}]^{-1}\vec{f}}_{\text{Steady state}} \tag{4.6.5a}$$

Numerical solution of ODE (one-root method)

$$\vec{u}_n = c_1(\sigma_1)^n \vec{x}_1 + c_2(\sigma_2)^n \vec{x}_2 + \cdots + c_M(\sigma_M)^n \vec{x}_M + [\mathrm{A}]^{-1}\vec{f} \tag{4.6.5b}$$

All of the $c_m, \vec{x}_m, \vec{f}$, and elements of $[\mathrm{A}]$ in these eqs are identical. All of the one-root methods represented in Table 4.6.1 produce a result identical to that given by eq (4.6.5b) in which one inserts the appropriate σ "signature" instead of eq (4.4.9). If a spurious root is generated, it adds another row to the expression for the transient with new constants, c_m (which are again fixed by the initial conditions), but with the same eigenvectors and the same steady-state solution.

4.7 Defective and/or Derogatory Matrices

In general, the eigenvalues of a matrix may not be distinct, in which case the possibility exists that it cannot be diagonalized. If the eigenvalues of a matrix are not distinct, but all of the eigenvectors are linearly independent, the matrix is said to be derogatory, but it can still be diagonalized. A set of eigenvectors is linearly independent if $a \cdot \vec{x}_m + b \cdot \vec{x}_n \neq \vec{x}_k$, where $m \neq n \neq k$ for any complex a and b and for all possible combinations of vectors in the set. However, if a matrix does not have a complete set of linearly independent eigenvectors, the matrix cannot be diagonalized, and it is said to be defective. A repeated root which causes a matrix to be defective will be referred to as a defective eigenvalue. Notice that, by this definition, a matrix can have an eigensystem with repeated eigenvalues that cause it to be both defective and derogatory. An example is given at the end of this section.

Matrices which are not strictly diagonalizable can still be put into a compact form by a similarity transform, S, such that

$$S^{-1}AS = [J] \tag{4.7.1}$$

and $[J]$ is a Jordan matrix composed of blocks of submatrices spread along the diagonal, each submatrix having the form

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \ddots & \vdots \\ 0 & 0 & \lambda_i & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 0 & \lambda_i \end{bmatrix} \tag{4.7.2}$$

Use of this transformation is known as putting A into its Jordan canonical form. If a matrix A has two or more Jordan submatrices that have the same eigenvalue, the matrix is said to be derogatory. For each Jordan submatrix with an eigenvalue λ_i of multiplicity r, there exists one eigenvector. The other $r - 1$ vectors associated with this eigenvalue are referred to as principal vectors.

In general, J has the form

$$J = \begin{bmatrix} J_1 & [0] & \cdots & [0] \\ [0] & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & [0] \\ [0] & \cdots & [0] & J_k \end{bmatrix} \tag{4.7.3}$$

where there are at most k distinct eigenvalues. We use the term Jordan subblock, or simply Jordan block, to represent a matrix having the form given by eq (4.7.3) or to represent λ_i itself. For example, the matrix

$$\begin{bmatrix} \begin{bmatrix} \lambda_1 & 1 & \\ & \lambda_1 & 1 \\ & & \lambda_1 \end{bmatrix} & & & & & \\ & \begin{bmatrix} \lambda_1 \end{bmatrix} & & & & \\ & & \begin{bmatrix} \lambda_1 \end{bmatrix} & & & \\ & & & \begin{bmatrix} \lambda_2 & 1 \\ & \lambda_2 \end{bmatrix} & & \\ & & & & \begin{bmatrix} \lambda_3 \end{bmatrix} & \\ & & & & & \begin{bmatrix} \lambda_4 \end{bmatrix} \end{bmatrix} \tag{4.7.4}$$

is both defective and derogatory, having:

9 eigenvalues,

4 distinct eigenvalues,

6 Jordan blocks,

6 linearly independent eigenvectors,

2 principal vectors with λ_1,

1 principal vector with λ_2,

2 defective eigenvalues,

3 derogatory eigenvalues.

Examples of defective eigenvalues occur in our subsequent development, see, for example, the discussion of the Gauss–Seidel method given later.

5. SOME PROPERTIES OF TRIDIAGONAL MATRICES

5.1 Standard Eigensystem for Simple Tridiagonals

In the following sections, tridiagonal banded matrices are prevalent. It is useful to list some of their properties. Many of these can be derived by solving the simple linear difference equations that arise in deriving recursion relations.

Let us consider a simple tridiagonal matrix, i.e. a tridiagonal with constant scalar elements a,b, and c. If we examine the conditions under which the determinant of this matrix is zero, we find

$$\left.\begin{array}{rl} \det[B(M:a,b,c)] = 0 & \text{if} \\ b + 2\sqrt{ac}\cos\left(\dfrac{m\pi}{M+1}\right) = 0 \quad , & m = 1,2,\cdots,M \end{array}\right\} \tag{5.1.1}$$

From this it follows at once that the eigenvalues of $B(a,b,c)$ are

$$\lambda_m = b + 2\sqrt{ac}\cos\left(\frac{m\pi}{M+1}\right) \quad , \quad m = 1,2,\cdots,M \tag{5.1.2}$$

The right-hand eigenvector of $B(a,b,c)$ that is associated with the eigenvalue λ_m satisfies the equation

$$B(a,b,c)\vec{x}_m = \lambda_m \vec{x}_m \tag{5.1.3}$$

It is given by

$$\vec{x}_m = (x_j)_m = \left(\frac{a}{c}\right)^{\frac{j-1}{2}} \sin\left[j\left(\frac{m\pi}{M+1}\right)\right] \quad , \quad m = 1,2,\cdots,M \tag{5.1.4}$$

These vectors are the columns of the right-hand eigenvector matrix, the elements of which are given by

$$X = (x_{jm}) = \left(\frac{a}{c}\right)^{\frac{j-1}{2}} \sin\left[\frac{jm\pi}{M+1}\right] \quad , \quad \begin{array}{l} j = 1,2,\cdots,M \\ m = 1,2,\cdots,M \end{array} \tag{5.1.5}$$

Notice that if $a = -1$ and $c = 1$,

$$\left(\frac{a}{c}\right)^{\frac{j-1}{2}} = e^{i(j-1)\frac{\pi}{2}} \tag{5.1.6}$$

The left-hand eigenvector matrix of $B(a,b,c)$ can be written

$$X^{-1} = \frac{2}{M+1}\left(\frac{c}{a}\right)^{\frac{m-1}{2}} \sin\left[\frac{mj\pi}{M+1}\right] \quad , \quad \begin{array}{l} m = 1,2,\cdots,M \\ j = 1,2,\cdots,M \end{array} \tag{5.1.7}$$

Notice that if $a = -1$ and $c = 1$

$$\left(\frac{c}{a}\right)^{\frac{m-1}{2}} = e^{-i(m-1)\frac{\pi}{2}} \tag{5.1.8}$$

5.2 Generalized Eigensystem for Simple Tridiagonals

This system is defined as follows

$$\begin{bmatrix} b & c & & & \\ a & b & c & & \\ & a & b & & \\ & & & \ddots & c \\ & & & a & b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_M \end{bmatrix} = \lambda \begin{bmatrix} e & f & & & \\ d & e & f & & \\ & d & e & & \\ & & & \ddots & f \\ & & & d & e \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_M \end{bmatrix} \tag{5.2.1}$$

In this case one can show after some algebra that

$$\left.\begin{array}{r} \det[B(a-\lambda d, b-\lambda e, c-\lambda f)] = 0 \quad , \quad \text{if} \qquad\qquad\qquad \\ b - \lambda_m e + 2\sqrt{(a-\lambda_m d)(c-\lambda_m f)}\cos\left(\dfrac{m\pi}{M+1}\right) = 0 \quad , \quad m = 1,2,\cdots,M \end{array}\right\} \tag{5.2.2}$$

If we define

$$\theta_m = \frac{m\pi}{M+1}, \quad p_m = \cos\theta_m \tag{5.2.3}$$

$$\lambda_m = \frac{eb - 2(cd + af)p_m^2 \pm 2p_m\sqrt{(ec - fb)(ea - bd) + [(cd - af)p_m]^2}}{e^2 - 4fdp_m^2} \tag{5.2.4}$$

The right-hand eigenvectors are

$$\vec{x}_m = \left[\frac{a - \lambda_m d}{c - \lambda_m f}\right]^{\frac{j-1}{2}} \sin[j\theta_m] \quad , \quad \begin{array}{c} m = 1, 2, \cdots, M \\ j = 1, 2, \cdots, M \end{array} \tag{5.2.5}$$

These relations are useful in studying SOR methods later in this report.

5.3 The Inverse of a Simple Tridiagonal

The inverse of $B(a, b, c)$ can also be written in analytic form. Let D_M represent the determinant of $B(M : a, b, c)$

$$D_M \equiv \det[B(M : a, b, c)] \tag{5.3.1}$$

Defining D_0 to be 1, it is simple to derive the first few determinants, thus

$$\left.\begin{aligned} D_0 &= 1 \\ D_1 &= b \\ D_2 &= b^2 - ac \\ D_3 &= b^3 - 2abc \end{aligned}\right\} \tag{5.3.2}$$

One can also find the recursion relation

$$D_M = bD_{M-1} - acD_{M-2} \tag{5.3.3}$$

Eq (5.3.3) is a linear OΔE the solution of which was discussed in Section 4. Its characteristic polynomial $P(E)$ is $P(E^2 - bE + ac)$ and the two roots to $P(\sigma) = 0$ result in the solution

$$D_M = \frac{1}{\sqrt{b^2 - 4ac}}\left\{\left[\frac{b + \sqrt{b^2 - 4ac}}{2}\right]^{M+1} - \left[\frac{b - \sqrt{b^2 - 4ac}}{2}\right]^{M+1}\right\} \tag{5.3.4}$$

$$M = 0, 1, 2, \cdots$$

where we have made use of the initial conditions $D_0 = 1$ and $D_1 = b$. In the limiting case when $b^2 - 4ac = 0$, one can show that

$$D_M = (M+1)\left(\frac{b}{2}\right)^M \quad ; \quad b^2 = 4ac \tag{5.3.5}$$

Then for $M = 4$

$$B^{-1} = \frac{1}{D_4}\begin{bmatrix} D_3 & -cD_2 & c^2D_1 & -c^3D_0 \\ -aD_2 & D_1D_2 & -cD_1D_1 & c^2D_1 \\ a^2D_1 & -aD_1D_1 & D_2D_1 & -cD_2 \\ -a^3D_0 & a^2D_1 & -aD_2 & D_3 \end{bmatrix} \tag{5.3.6}$$

and for $M = 5$

$$B^{-1} = \frac{1}{D_5}\begin{bmatrix} D_4 & -cD_3 & c^2D_2 & -c^3D_1 & c^4D_0 \\ -aD_3 & D_1D_3 & -cD_1D_2 & c^2D_1D_1 & -c^3D_1 \\ a^2D_2 & -aD_1D_2 & D_2D_2 & -cD_2D_1 & c^2D_2 \\ -a^3D_1 & a^2D_1D_1 & -aD_2D_1 & D_3D_1 & -cD_3 \\ a^4D_0 & -a^3D_1 & a^2D_2 & -aD_3 & D_4 \end{bmatrix} \tag{5.3.7}$$

The general element d_{mn} is

Upper triangle:

$$m = 1, 2, \cdots, M-1 \quad ; \quad n = m+1, m+2, \cdots, M$$

$$d_{mn} = D_{m-1}D_{M-n}(-c)^{n-m}/D_M$$

Diagonal:

$$n = m = 1, 2, \cdots, M$$

$$d_{mm} = D_{M-1}D_{M-m}/D_M \tag{5.3.8}$$

Lower triangle:

$$m = n+1, n+2, \cdots, M \quad ; \quad n = 1, 2, \cdots, M-1$$

$$d_{mn} = D_{M-m}D_{n-1}(-a)^{m-n}/D_M$$

5.4 Periodic or Circulant Tridiagonal Matrices

Next consider the periodic tridiagonal shown in eq (1.1.4)

$$B_p(M : a, b, c) \tag{5.4.1}$$

The eigenvalues are

$$\lambda_m = b + (a + c)\cos\left(\frac{2\pi m}{M}\right) - i(a - c)\sin\left(\frac{2\pi m}{M}\right) \quad , \quad m = 0, 1, 2, \cdots, M - 1 \tag{5.4.2}$$

Notice the slight shift in the index which makes the notation for the periodic analysis more convenient. The right-hand eigenvector that satisfies $B_p(a, b, c)\vec{x}_m = \lambda_m \vec{x}_m$ is

$$\vec{x}_m = (x_j)_m = e^{ij\,(2\pi m/M)} \quad , \quad j = 0, 1, \cdots, M - 1 \tag{5.4.3a}$$

where $i \equiv \sqrt{-1}$. This can also be written

$$\vec{x}_m = \cos\left[j\left(\frac{2\pi m}{M}\right)\right] + i \quad \sin\left[j\left(\frac{2\pi m}{M}\right)\right] \quad , \quad j = 0, 1, \cdots, M - 1 \tag{5.4.3b}$$

The left-hand eigenvector matrix is

$$X^{-1} = \frac{1}{M} e^{im\left(\frac{2\pi j}{M}\right)} \quad , \quad \begin{matrix} m = 0, 1, \cdots, M - 1 \\ j = 0, 1, \cdots, M - 1 \end{matrix} \tag{5.4.4}$$

Notice the remarkable fact that the elements of the eigenvector matrices X and X^{-1} for the periodic matrix do not depend on the elements a, b, c in the original matrix. In fact, all periodic (or circulant) matrices of order M have the same set of linearly independent eigenvectors. Further examination shows that the elements in these eigenvectors correspond to the elements in a complex harmonic analysis or complex discrete Fourier series.

A full (or completely dense) circulant matrix of order $M = 5$ is shown in eq (5.4.5).

$$B_p = \begin{bmatrix} b_0 & b_1 & b_2 & b_3 & b_4 \\ b_4 & b_0 & b_1 & b_2 & b_3 \\ b_3 & b_4 & b_0 & b_1 & b_2 \\ b_2 & b_3 & b_4 & b_0 & b_1 \\ b_1 & b_2 & b_3 & b_4 & b_0 \end{bmatrix} \tag{5.4.5}$$

As we have just pointed out, the eigenvectors are given by eq (5.4.3) and are independent of the b_j. However, the eigenvalues do depend on the matrix elements and they have the general form

$$\lambda_m = \sum_{j=0}^{M-1} b_j e^{i(2\pi jm/M)} \tag{5.4.6}$$

of which eq (5.4.2) is a special case.

5.5 Special Cases Found From Symmetries

Consider a mesh with an even number of interior points such as that shown in Fig 5.5.1a. One can seek from the tridiagonal matrix $B(2M : a, b, a,)$ the eigenvector subset that has even symmetry. When spanning the interval $0 \leq x \leq \pi$. For example, we seek the set of eigenvectors $\vec{x}_m$ for which

$$\begin{bmatrix} b & a & & & & \\ a & b & a & & & \\ & a & \ddots & & & \\ & & & \ddots & a & \\ & & & a & b & a \\ & & & & a & b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_2 \\ x_1 \end{bmatrix} = \lambda_m \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_2 \\ x_1 \end{bmatrix} \tag{5.5.1}$$

This leads to the subsystem of order M which has the form

$$B(M : a, \vec{b}, a)\vec{x}_m = \lambda_m \vec{x}_m \tag{5.5.2}$$

where

$$B(M : a,\vec{b},a) = \begin{bmatrix} b & a & & & & \\ a & b & a & & & \\ & a & \ddots & & & \\ & & & \ddots & a & \\ & & & a & b & a \\ & & & & a & b+a \end{bmatrix} \tag{5.5.3}$$

By folding the known eigenvectors of $B(2M : a,b,a)$ about the center, one can show from previous results that the eigenvalues of (5.5.3) are

$$\lambda_m = b + 2a\cos\left(\frac{(2m-1)\pi}{2M+1}\right) \quad , \quad m = 1,2,\cdots,M \tag{5.5.4}$$

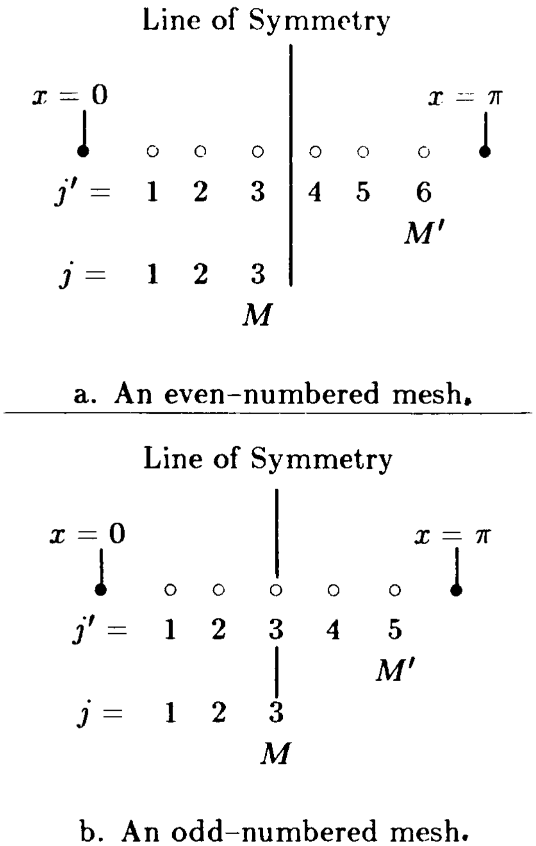
and the corresponding eigenvectors are

$$\vec{x}_m = \sin\left(\frac{j(2m-1)\pi}{2M+1}\right) \quad , \qquad j = 1,2,\cdots,M \tag{5.5.5}$$

Imposing symmetry about the same interval but for a mesh with an odd number of points, see Fig. 5.5.1b, leads to the matrix

$$B(M : \vec{a},b,a) = \begin{bmatrix} b & a & & & \\ a & b & a & & \\ & a & \ddots & & \\ & & \ddots & a & \\ & & a & b & a \\ & & & 2a & b \end{bmatrix} \tag{5.5.6}$$

By folding the known eigenvalues of $B(2M - 1 : a,b,a)$ about the center, one can show from previous results that the eigenvalues of (5.5.6) are



a. An even-numbered mesh.

b. An odd-numbered mesh.

Figure 5.5.1 – Symmetrical folds for special cases.

$$\lambda_m = b + 2a\cos\left(\frac{(2m-1)\pi}{2M}\right) \quad , \quad m = 1,2,\cdots,M \tag{5.5.7}$$

and the corresponding eigenvectors are

$$\vec{x}_m = \sin\left(\frac{j(2m-1)\pi}{2M}\right) \quad , \quad j = 1,2,\cdots,M \tag{5.5.8}$$

5.6 Special Cases Involving Boundary Conditions

We consider three special cases for the matrix operator representing the 3-point central difference approximation for the second derivative $\partial^2/\partial x^2$ at all points away from the boundaries, combined with special conditions imposed at the boundaries.

Note: In every case

$$\begin{aligned} m &= 1,2,\cdots,M \\ j &= 1,2,\cdots,M \\ -2 + 2\cos(\alpha) &= -4\sin^2(\alpha/2) \end{aligned}$$

When the boundary conditions are Dirichlet on both sides,

$$\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \quad \left.\begin{aligned} \lambda_m &= -2 + 2\cos\left(\frac{m\pi}{M+1}\right) \\ \vec{x}_m &= \sin\left[j\left(\frac{m\pi}{M+1}\right)\right] \end{aligned}\right\} \tag{5.6.1}$$

When one boundary condition is Dirichlet and one is (first–order) Neumann

$$\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix} \quad \left.\begin{aligned} \lambda_m &= -2 + 2\cos\left[\frac{(2m-1)\pi}{2M+1}\right] \\ \vec{x}_m &= \sin\left[j\left(\frac{(2m-1)\pi}{2M+1}\right)\right] \end{aligned}\right\} \tag{5.6.2}$$

When the boundary conditions are Neumann on both sides

$$\begin{bmatrix} 1 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{bmatrix} \qquad \left.\begin{aligned} \lambda_m &= -2 + 2\cos\left[\frac{(m-1)\pi}{M}\right] \\ \vec{x}_m &= \cos\left[(j - \frac{1}{2})\left(\frac{(m-1)\pi}{M}\right)\right] \end{aligned}\right\} \tag{5.6.3}$$

Notice that only the last matrix is singular.

6. CLASSICAL RELAXATION

6.1 The Converged Solution, the Residual, and the Error

Most standard texts, see refs. 4,5,6, approach the subject of relaxation by considering what we have defined as the preconditioned form, see eq (2.2.1) and (2.3.2),

$$A\vec{\phi} - \vec{f} = 0 \tag{6.1.1}$$

and applying to it the iterative process

$$\vec{\phi}_{n+1} = [I + \mathcal{M}A]\vec{\phi}_n - \mathcal{M}\vec{f} \tag{6.1.2}$$

In the terminology of Section 3, $\mathcal{M}$ is the secondary conditioning matrix equal to H^{-1} in eq (3.3.1). Equation (6.1.2) is usually rewritten in the form

$$\left.\begin{aligned} &\vec{\phi}_{n+1} = G\vec{\phi}_n - \mathcal{M}\vec{f} \\ \text{where} \qquad &G \equiv I + \mathcal{M}A \end{aligned}\right\} \tag{6.1.3}$$

In this notation, G is the basic iteration matrix and its eigenvalues, which we designate as σ_m, determine the convergence rate of a method. The converged solution to all three eqs (6.1.1), (6.1.2), and (6.1.3) is

$$\vec{\phi}_\infty = A^{-1}\vec{f} \tag{6.1.4}$$

The error at the nth iteration is defined as

$$\vec{e}_n \equiv \vec{\phi}_n - \vec{\phi}_\infty = \vec{\phi}_n - A^{-1}\vec{f} \tag{6.1.5}$$

The residual at the nth iteration is defined as

$$\vec{r}_n \equiv A\vec{\phi}_n - \vec{f} \tag{6.1.6}$$

Multiply eq (6.1.5) by A from the left, and use the definition in eq (6.1.6). There results the relation between the error and the residual

$$A\vec{e}_n - \vec{r}_n = 0 \tag{6.1.7}$$

It is not difficult to show that

$$\vec{e}_{n+1} = G\vec{e}_n \tag{6.1.8}$$

In all of the above, we have considered only what are usually referred to as stationary processes. We are in fact much more interested in nonstationary processes, but our approach to them is not standard so we defer the discussion of these to Section 7.

6.2 Point Operator Schemes in One Dimension

Let us consider three classical relaxation procedures for the one dimensional equation

$$\frac{\partial^2 \phi}{\partial x^2} = g(x) \tag{6.2.1}$$

which, with three-point central differencing and Dirichlet boundary conditions, reduces to the model equation

$$B(1,-2,1)\vec{\phi} = \vec{f} \tag{6.2.2}$$

where $\vec{f} = \Delta x^2 \vec{g}$. These methods are very well known but the terminology is not universal. For example, the Gauss-Seidel method is sometimes called the Liebman method and the Point-Jacobi method has been referred to as the Richardson method.

What we refer to as the Point-Jacobi method is expressed in point operator form for the one-dimensional case as

$$u_j^{(n+1)} = \frac{1}{2}\left[u_{j-1}^{(n)} + u_{j+1}^{(n)} - f_j\right] \tag{6.2.3}$$

The Gauss-Seidel method is

$$u_j^{(n+1)} = \frac{1}{2}\left[u_{j-1}^{(n+1)} + u_{j+1}^{(n)} - f_j\right] \tag{6.2.4}$$

The method of successive overrelaxation (SOR) is usually expressed in two steps as

$$\begin{aligned}\tilde{u}_j &= \frac{1}{2}\left[u_{j-1}^{(n+1)} + u_{j+1}^{(n)} - f_j\right] \\ u_j^{(n+1)} &= u_j^{(n)} + \omega\left[\tilde{u}_j - u_j^{(n)}\right]\end{aligned} \qquad (6.2.5a)$$

but it can also be written in the single line

$$u_j^{(n+1)} = \frac{\omega}{2}u_{j-1}^{(n+1)} + (1-\omega)u_j^{(n)} + \frac{\omega}{2}u_{j+1}^{(n)} - \frac{\omega}{2}f_j \qquad (6.2.5b)$$

6.3 The Convergence Rates

The usual measure of the convergence rate is the eigenvalue σ_m of G, see eq (6.1.3), having maximum absolute value. Thus

$$\text{Convergence} \quad \sim \quad |\sigma_m|_{max} \quad ; \quad m = 1, 2, \cdots, M \qquad (6.3.1)$$

These values are well known for Laplace's equation using the three methods just defined. They are

$$\text{Point–Jacobi} \qquad |\sigma_m|_{max} = \cos\left(\frac{\pi}{M+1}\right) \qquad (6.3.2a)$$

$$\text{Gauss–Seidel} \qquad |\sigma_m|_{max} = \left[\cos\left(\frac{\pi}{M+1}\right)\right]^2 \qquad (6.3.2b)$$

$$\text{SOR} \qquad \left.\begin{aligned}|\sigma_m|_{max} &= \omega_{opt} - 1 \\ \omega_{opt} &= 2/\left[1 + \sin\left(\frac{\pi}{M+1}\right)\right]\end{aligned}\right\} \qquad (6.3.2c)$$

7. ODE APPROACH TO CLASSICAL RELAXATION

7.1 ODE Form of the Classical Methods

The three iterative procedures defined by eqs (6.2.3), (6.2.4), and (6.2.5) obey no apparent pattern except that they are easy to implement in a computer code since all of the data required to update the value of one point are explicitly available at the time of the update. Now let us study these methods as subsets of ODE as formulated in Section 3. Insert the model equation (6.2.2) into the ODE form (3.3.2). Then

$$H\frac{d\vec{\phi}}{dt} = B(1,-2,1)\vec{\phi} - \vec{f} \tag{7.1.1}$$

As a start, let us use for the numerical integration the explicit Euler method

$$\phi_{n+1} = \phi_n + h\phi'_n \tag{7.1.2}$$

with a step size, h, equal to 1. We arrive at

$$H(\vec{\phi}_{n+1} - \vec{\phi}_n) = B(1,-2,1)\vec{\phi}_n - \vec{f} \tag{7.1.3}$$

It is clear that the best choice of H from the point of view of matrix algebra is $-B(1,-2,1)$ since then multiplication from the left by $-B^{-1}(1,-2,1)$ gives the correct answer in one step. However, this is not in the spirit of our study, since multiplication by the inverse amounts to solving the problem by direct methods without iteration. The constraint on H that is in keeping with the formulation of the three methods described in Section 6 is that all the elements above the diagonal (or below the diagonal if the sweeps are from right to left) are zero. If we impose this constraint and further restrict ourselves to banded tridiagonals with a single constant in each band, we are led to

$$B(-\beta,\frac{2}{\omega},0)(\vec{\phi}_{n+1} - \vec{\phi}_n) = B(1,-2,1)\vec{\phi}_n - \vec{f} \tag{7.1.4}$$

where β and ω are arbitrary. With this choice of notation the three methods presented in Section 6 can be identified using the entries in Table 7.1.1.

TABLE 7.1.1 – VALUES OF β and ω IN EQ (7.1.4) THAT LEAD TO CLASSICAL RELAXATION METHODS

β	ω	Method	Equation
0	1	Point–Jacobi	6.2.3
1	1	Gauss–Seidel	6.2.4
1	$2/\left[1+\sin\left(\frac{\pi}{M+1}\right)\right]$	Optimum SOR	6.2.5

The fact that the values in the tables lead to the methods indicated can be verified by simple algebraic manipulation. However, our purpose is to examine the whole procedure as a special subset of the theory of ordinary differential equations. In this light, the three methods are all contained in the set of ODE

$$\frac{d\vec{\phi}}{dt} = B^{-1}(-\beta, \frac{2}{\omega}, 0)\left[B(1,-2,1)\vec{\phi} - \vec{f}\right] \tag{7.1.5}$$

and appear from it in the special case when the explicit Euler method is used for its numerical integration. The point operator that results from the use of the Euler scheme, $\phi_{n+1} = \phi_n + h\phi'_n$, is

$$\phi_j^{(n+1)} = \left(\frac{\omega\beta}{2}\phi_{j-1}^{(n+1)} + \frac{\omega}{2}(h-\beta)\phi_{j-1}^{(n)}\right) - \left((\omega h - 1)\phi_j^{(n)}\right) + \left(\frac{\omega h}{2}\phi_{j+1}^{(n)}\right) - \frac{\omega h}{2}f_j \tag{7.1.6}$$

This represents a generalization of classical relaxation that results from the numerical solution of ODE.

7.2 The λ Eigenvalues and the Error

It is at this point that we start to deviate from the usual presentation of relaxation. In our approach it is essential that we first identify the λ eigenvalues of the iterative procedure, and then determine the σ eigenvalues as a function of λh. The λ eigenvalues are fixed by the basic matrix in eq (2.1.1), the preconditioning matrix in (2.2.1), and the secondary conditioning matrix in (3.3.2). The σ eigenvalues are fixed for a given λh by the choice of numerical method as illustrated in Table 4.6.1. Throughout the remaining discussion we will refer to the independent variable t as "time", even though no true time accuracy is necessarily involved.

The general form of the relaxation procedure, as we are reviewing it, is

$$\frac{d\vec{\phi}}{dt} = H^{-1}C\Big[A_b\vec{\phi} - \vec{f}_b\Big] = H^{-1}[A\vec{\phi} - \vec{f}] \tag{7.2.1}$$

From Section 4, we see that, if the eigenvectors of $[H^{-1}A]$ are linearly independent, the solution can be written as

$$\vec{\phi} = \underbrace{c_1 e^{\lambda_1 t}\vec{x}_1 + \cdots + c_M e^{\lambda_M t}\vec{x}_M}_{\text{error}} + \vec{\phi}_\infty \tag{7.2.2}$$

where what is referred to in time-accurate analysis as the transient solution (see eq (4.2.11)), is now referred to in relaxation analysis as the error (see Section 6). We see that the problem in the relaxation subset of ODE is to remove the transient solution from the general solution in the most efficient way possible.

7.3 Stationary Processes

If H and C in eq (7.2.1) are independent of t, that is, are not changed throughout the iteration process, the method is referred to, in relaxation terminology, as stationary. The generalization of this in our approach is to make h, the "time" step, a constant for the entire iteration.

Suppose the explicit Euler method is used for the time integration. Then, from Section 6 and eq (7.2.2), the numerical solution can be expressed, after n steps, as

$$\vec{\phi}_n = \underbrace{c_1\vec{x}_1(1+\lambda_1 h)^n + \cdots + c_m\vec{x}_m(1+\lambda_m h)^n + \cdots + c_M\vec{x}_M(1+\lambda_M h)^n}_{\text{error}} + \vec{\phi}_\infty \tag{7.3.1}$$

The initial amplitudes of the eigenvectors are given by the magnitudes of the c_m. These were fixed by the initial guess. In general it is assumed that any or all of the eigenvectors could have been given an equally "bad" excitation by the initial guess, so that we must devise a way to remove them all from the general solution on an equal basis. Assuming that $[H^{-1}A]$ has been chosen (that is, an iteration process has been decided upon), the only free choice remaining to accelerate the removal of the error terms is the choice of h. The three methods represented in Table 7.1.1 have all been conditioned by the choice of H to have an optimum h equal to 1 for a stationary iteration process.

7.4 Nonstationary Processes

In classical terminology a method is said to be nonstationary if the conditioning matrices, $[H^{-1}C]$, are varied at each time step. This would not change the steady-state solution $A_b^{-1}\vec{f}_b$, but it can greatly affect the convergence rate. In our ODE approach this could also be considered and would lead to a study of equations with nonconstant coefficients. It is much simpler, however, to study the case of fixed $[H^{-1}C]$ but variable step size, h. This process changes the Point-Jacobi method to Richardson's method in standard terminology, see ref. 4. For the Gauss-Seidel and SOR methods it leads to processes that are superior to the stationary methods but, to the author's knowledge, unpublished.

The nonstationary form of eq (7.3.1) is

$$\begin{aligned}\vec{\phi}_n = c_1\vec{x}_1 \prod_{n=1}^{N}(1+\lambda_1 h_n) + \cdots + c_m\vec{x}_m \prod_{n=1}^{N}(1+\lambda_m h_n) \\ + \cdots + c_M\vec{x}_M \prod_{n=1}^{N}(1+\lambda_M h_n) + \vec{\phi}_\infty\end{aligned} \tag{7.4.1}$$

where the symbol Π stands for product. Since h_n can now be changed at each step, the error term can theoretically be completely eliminated in M steps by taking $h_m = -1/\lambda_m$, for $m = 1, 2, \cdots, M$. This concept of eigenvector annihilation is discussed in Section 8.

Let us consider the very important case when all of the λ_m are <u>real</u> and <u>negative</u> (remember that they arise from a conditioned matrix so this constraint is not unrealistic for quite practical cases). Consider one of the error terms taken from

$$\vec{e}_n \equiv \vec{\phi}_n - \vec{\phi}_\infty = \sum_{m=1}^{M} c_m\vec{x}_m \prod_{n=1}^{N}(1+\lambda_m h_n) \tag{7.4.2}$$

and write it in the form

$$c_m\vec{x}_m P_e(\lambda_m) \equiv c_m\vec{x}_m \prod_{n=1}^{N}(1+\lambda_m h_n) \tag{7.4.3}$$

where P_e signifies an "Euler" polynomial. Now focus attention on the polynomial

$$(P_e)_N(\lambda) = (1+h_1\lambda)(1+h_2\lambda)\cdots(1+h_N\lambda) \tag{7.4.4}$$

treating it as a continuous function of the independent variable λ. In the annihilation process mentioned after eq (7.4.1) we considered making the error exactly zero by taking advantage of some knowledge about the discrete values of λ_m for a particular case. Now we pose a less demanding problem. Let us choose the h_n so that the maximum value of $(P_e)_N(\lambda)$ is as small as possible for all λ lying between λ_a and λ_b such that $\lambda_b \leq \lambda \leq \lambda_a \leq 0$. Mathematically stated, we seek

$$\max_{\lambda_b \leq \lambda \leq \lambda_a} |(P_e)_N(\lambda)| = \text{minimum} \quad ; \quad \text{with}(P_e)_N(0) = 1 \tag{7.4.5}$$

This problem has a well known solution due to Markov. It is

$$(P_e)_N(\lambda) = \frac{T_N\left(\frac{2\lambda - \lambda_a - \lambda_b}{\lambda_a - \lambda_b}\right)}{T_N\left(\frac{-\lambda_a - \lambda_b}{\lambda_a - \lambda_b}\right)} \tag{7.4.6}$$

where

$$T_N(y) = \cos(N \arccos y) \tag{7.4.7a}$$

are the Chebyshev polynomials along the interval $-1 \leq y \leq 1$ and

$$T_N(y) = \frac{1}{2}\left[y + \sqrt{y^2 - 1}\right]^N + \frac{1}{2}\left[y - \sqrt{y^2 - 1}\right]^N \tag{7.4.7b}$$

are the Chebyshev polynomials for $|y| > 1$. In relaxation terminology this is generally referred to as Richardson's method, see ref. 6, and it leads to the nonstationary step size choice given by

$$\frac{1}{h_n} = \frac{1}{2}\left\{-\lambda_b - \lambda_a + (\lambda_b - \lambda_a)\cos\left[\frac{(2n-1)\pi}{2N}\right]\right\}, \quad n = 1, 2, \cdots N \tag{7.4.8}$$

Remember that all λ are negative real numbers representing the magnitudes of λ_m in an eigenvalue spectrum.

The error in the relaxation process represented by eq (7.4.1) is expressed in terms of a set of eigenvectors, $\vec{x}_m$, amplified by the coefficients $c_m \prod(1 + \lambda_m h_n)$. With each eigenvector there is a corresponding eigenvalue. Equation (7.4.8) gives us the best choice of a series of h_n that will minimize the amplitude of the error carried in the eigenvectors associated with the eigenvalues between λ_b and λ_a.

As an example for the use of eq (7.4.8), let us consider the following problem:

$$\boxed{\text{Minimize the maximum error associated with the } \lambda \text{ eigenvalues in the interval } -2 \le \lambda \le -1 \text{ using only 3 iterations.}} \tag{7.4.9}$$

The three values of h which satisfy this problem are

$$h_n = 2\Big/\left(3 - \cos\left[\frac{(2n-1)\pi}{6}\right]\right) \tag{7.4.10}$$

and the amplitude of the eigenvector is reduced to

$$(P_e)_3(\lambda) = T_3(2\lambda + 3)/T_3(3) \tag{7.4.11}$$

where

$$T_3(3) = \left\{[3+\sqrt{8}]^3 + [3-\sqrt{8}]^3\right\}/2 \approx 99 \tag{7.4.12}$$

A plot of eq (7.4.12) is given in Fig. 7.4.1 and we see that the amplitudes of all the eigenvectors associated with the eigenvalues in the range $-2 \le \lambda \le -1$ have been reduced to less than about 1% of their initial values.

$$h_1 = 4/(6 - \sqrt{3})$$
$$h_2 = 4/(6 - 0)$$
$$h_3 = 4/(6 + \sqrt{3})$$

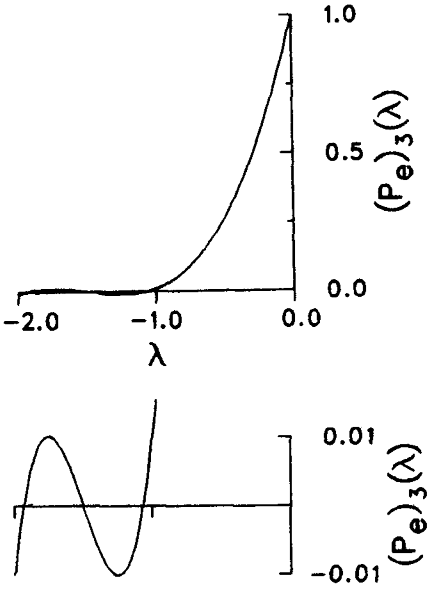


Figure 7.4.1 – Richardson's method for 3 steps, minimized for $-2 \le \lambda \le -1$.

Return now to eq (7.4.1). This was derived from eq (7.2.2) on the condition that the explicit Euler method, eq (7.1.2), was used to integrate the basic ODE. If instead the implicit trapezoidal rule

$$\phi_{n+1} = \phi_n + \frac{1}{2}h(\phi'_{n+1} + \phi'_n) \tag{7.4.13}$$

is used, the nonstationary formula

$$\vec{\phi}_N = \sum_{m=1}^{M} c_m \vec{x}_m \prod_{n=1}^{N} \left(\frac{1 + \frac{1}{2}h_n\lambda_m}{1 - \frac{1}{2}h_n\lambda_m}\right) + \vec{\phi}_\infty \tag{7.4.14}$$

would result. This calls for a study of the rational "trapezoidal" polynomial, P_t,

$$(P_t)_N(\lambda) = \prod_{n=1}^{N} \left(\frac{1 + \frac{1}{2}h_n\lambda}{1 - \frac{1}{2}h_n\lambda} \right) \tag{7.4.15}$$

under the same constraints as before, namely that

$$\begin{aligned} \max_{\lambda_b \le \lambda \le \lambda_a} |(P_t)_N(\lambda)| &= \text{minimum} \;; \\ \text{with } (P_t)_N(0) &= 1 \end{aligned} \tag{7.4.16}$$

The optimum values of h can also be found for this problem, see ref. 6, but we settle here for the approximation suggested by Wachspress

$$\frac{2}{h_n} = -\lambda_b \left(\frac{\lambda_a}{\lambda_b} \right)^{(n-1)/(N-1)} \quad ; \quad n = 1, 2, \cdots, N \tag{7.4.17}$$

This process is also applied to problem (7.4.9). The results for $(P_t)_3(\lambda)$ are shown in Fig. 7.4.2. The error amplitude is about $1/5$ of that found for $(P_e)_3(\lambda)$ in the same interval of λ.

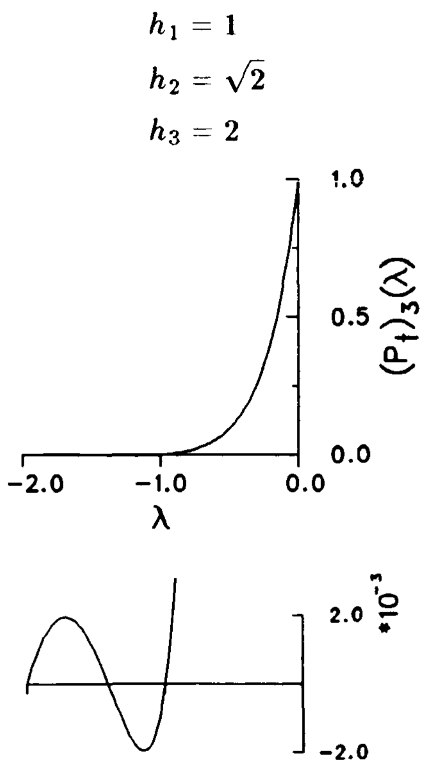


Figure 7.4.2 – Wachpress method for 3 steps, minimized for $-2 \le \lambda \le -1$.

7.5 Eigensystems of the Classical Methods

Before we carry the ODE analysis further, it is instructive to inspect the eigenvectors and eigenvalues in the $[H^{-1}A]$ matrix for the three classical methods represented by eqs (6.2.3), (6.2.4), and (6.2.5). Using eq (7.1.5) we see that this amounts to solving the general eigenvalue problem

$$H\vec{x}_m = \lambda_m A \vec{x}_m \tag{7.5.1}$$

for the special case

$$B(-\beta, \frac{2}{\omega}, 0)\vec{x}_m = \lambda_m B(1, -2, 1)\vec{x}_m \tag{7.5.2}$$

Eq (7.5.2) is a special case of eq (5.2.1) so the solution is given in Section 5. The three special cases are considered below. To illustrate the behavior we take $M = 5$ for the matrix order. This special case makes the general result quite clear.

7.6 The Point–Jacobi System

If $\beta = 0$ and $\omega = 1$ in eq (7.1.5), the ODE matrix reduces to simply $B(\frac{1}{2}, -1, \frac{1}{2})$. The eigensystem for this matrix is given in eqs (5.1.1) through (5.1.8). One finds the following.

$$X = \overbrace{\begin{bmatrix} 1/2 & \sqrt{3}/2 & 1 & \sqrt{3}/2 & 1/2 \\ \sqrt{3}/2 & \sqrt{3}/2 & 0 & -\sqrt{3}/2 & -\sqrt{3}/2 \\ 1 & 0 & -1 & 0 & 1 \\ \sqrt{3}/2 & -\sqrt{3}/2 & 0 & \sqrt{3}/2 & -\sqrt{3}/2 \\ 1/2 & -\sqrt{3}/2 & 1 & -\sqrt{3}/2 & 1/2 \end{bmatrix}}^{\text{Columns of eigenvectors}} \tag{7.6.1a}$$

$$\begin{array}{lccccc} \text{Eigenvector number} & \underbrace{}_{1} & \underbrace{}_{2} & \underbrace{}_{3} & \underbrace{}_{4} & \underbrace{}_{5} \\ \text{Eigenvalue} & \frac{1}{2}[-2+\sqrt{3}] & -\frac{1}{2} & -1 & -\frac{3}{2} & \frac{1}{2}[-2+\sqrt{3}] \end{array}$$

The left hand eigenvector set that is consistent with this normalization is

$$X^{-1} = \frac{1}{3} \begin{bmatrix} 1/2 & \sqrt{3}/2 & 1 & \sqrt{3}/2 & 1/2 \\ \sqrt{3}/2 & \sqrt{3}/2 & 0 & -\sqrt{3}/2 & -\sqrt{3}/2 \\ 1 & 0 & -1 & 0 & 1 \\ \sqrt{3}/2 & -\sqrt{3}/2 & 0 & \sqrt{3}/2 & -\sqrt{3}/2 \\ 1/2 & -\sqrt{3}/2 & 1 & -\sqrt{3}/2 & 1/2 \end{bmatrix} \begin{array}{lc} & \underbrace{\text{Eigenvector number}} \\ \} & 1 \\ \} & 2 \\ \} & 3 \\ \} & 4 \\ \} & 5 \end{array} \tag{7.6.1b}$$

If we inspect eq (7.4.1), we notice that the initial error content is given by

$$\vec{e}_0 = \vec{\phi}_0 - \vec{\phi}_\infty = X\vec{c} \tag{7.6.2a}$$

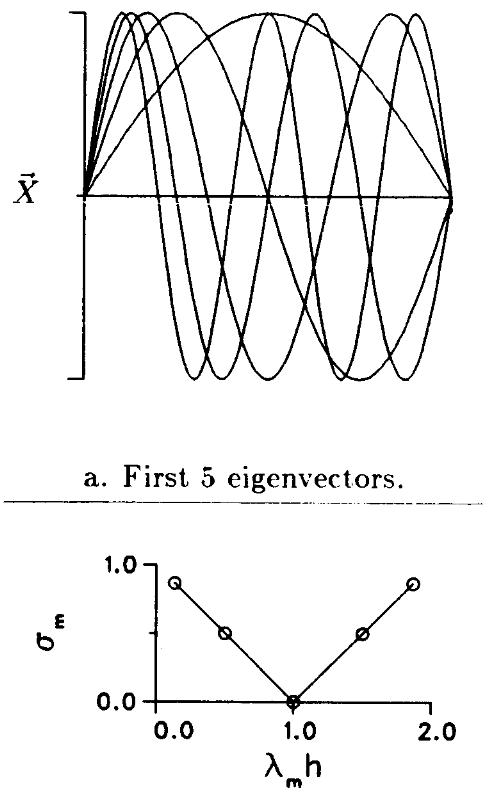
Insert eq (5.1.5) for X, then

$$(e_0)_j = \sum_{m=1}^{M} c_m \sin\left[m\left(\frac{j\pi}{M+1}\right)\right] \quad ; \quad j = 1, 2, \cdots, M \tag{7.6.2b}$$

This simply states that the elements in the initial error vector, or the initial transient terms in the ODE, are given by a sine transform of the amplitudes of the eigenvector content. Similarly

$$\vec{c}_m = \frac{2}{M+1}\sum_{j=1}^{M}(e_0)_j \sin\left[j\left(\frac{m\pi}{M+1}\right)\right] \quad ; \quad m = 1, 2, \cdots, M \tag{7.6.3}$$

represents $\vec{c}$ as a sine synthesis of $\vec{e}_0$. This is a very "well-behaved" eigensystem with linearly independent eigenvectors and distinct eigenvalues. The first 5 eigenvectors, which are simple sine waves, are shown in Fig. 7.6.1a. The eigenvalues are given by the equation



a. First 5 eigenvectors.

b. The σ, λ relation, $h = 1$.

Figure 7.6.1 – Point-Jacobi eigensystem, $M = 5$.

$$\lambda_m = -1 + \cos\left(\frac{m\pi}{M+1}\right) \tag{7.6.4}$$

The functional dependence of σ on λ for the explicit Euler method is $\sigma_m = 1 + \lambda_m h$. Thus the σ,λ relation can be plotted for any h. The plot for $h = 1$, the optimum stationary case, is shown in Fig. 7.6.1b.

7.7 The Gauss-Seidel System

If β and ω are equal to 1 in eq (7.1.5), the matrix eigensystem evolves from the relation

$$B(1,-2,1)\vec{x}_m = \lambda_m B(1,-2,0)\vec{x}_m \tag{7.7.1}$$

which can be studied using the results in Section 5.2. One can show that the $[HA^{-1}]$ matrix for the Gauss-Seidel method, $[A]_{GS}$, is

$$[A]_{GS} \equiv B^{-1}(-1,2,0)B(1,-2,1) =$$

$$\begin{array}{cc}
\left[\begin{matrix}
-1 & 1/2 & & & & & \\
0 & -3/4 & 1/2 & & & & \\
0 & 1/8 & -3/4 & 1/2 & & & \\
0 & 1/16 & 1/8 & -3/4 & 1/2 & & \\
0 & 1/32 & 1/16 & 1/8 & -3/4 & 1/2 & \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 1/2^M & \cdots & & & &
\end{matrix}\right]
&
\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \vdots \\ M \end{matrix}
\end{array} \tag{7.7.2}$$

The eigenvector structure of the Gauss-Seidel ODE matrix is quite interesting. If M is odd there are $(M+1)/2$ distinct eigenvalues with corresponding linearly independent eigenvectors, and there are $(M-1)/2$ defective eigenvalues with corresponding principal vectors. The Jordan canonical form for $M = 5$ is shown in Fig. 7.7.1.

$$X^{-1}[A]_{GS}X = [J]_{GS} = \begin{bmatrix} \begin{bmatrix}\lambda_1\end{bmatrix} & & \\ & \begin{bmatrix}\lambda_2\end{bmatrix} & \\ & & \begin{bmatrix} \lambda_3 & 1 & \\ & \lambda_3 & 1 \\ & & \lambda_3 \end{bmatrix} \end{bmatrix}$$

Figure 7.7.1 - Jordan canonical form for Gauss-Seidel, $M = 5$.

The right-hand column vector set, X, is shown in eq (7.7.3), where the * stands for a principal vector in the defective system, and the corresponding left-hand row vectors, X^{-1}, are shown in eq (7.7.4). Neither set has been normalized in any rational way.

$$X = \begin{bmatrix} 16 & 16 & 1 & 0 & 0 \\ 24 & 8 & 0 & 2 & 0 \\ 24 & 0 & 0 & -1 & 4 \\ 18 & -2 & 0 & 0 & -4 \\ 9 & -1 & 0 & 0 & 1 \end{bmatrix} \tag{7.7.3}$$

$$\begin{array}{lccccc} & \underbrace{\quad} & \underbrace{\quad} & \underbrace{\quad} & \underbrace{\quad} & \underbrace{\quad} \\ \text{Vector number} & 1 & 2 & 3 & * & * \\ \text{Eigenvalue} & -1/4 & -3/4 & -1 & -1 & -1 \end{array}$$

$$X^{-1} = \begin{bmatrix} 0 & 1/144 & 1/72 & 1/54 & 1/54 \\ 0 & 1/16 & 1/8 & 0 & -1/2 \\ 1 & -10/9 & -20/9 & -8/27 & 208/27 \\ 0 & 5/30 & -2/3 & -2/9 & 16/9 \\ 0 & 0 & 0 & -1/6 & 1/3 \end{bmatrix} \tag{7.7.4}$$

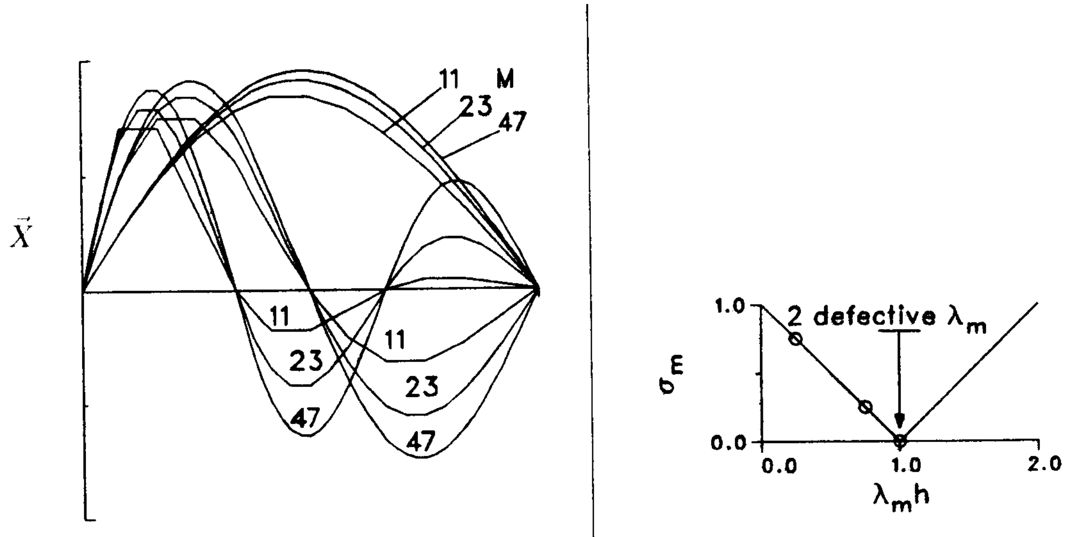
The equation for the nondefective eigenvalues in the ODE matrix is (for odd M)

$$\lambda_m = -1 + \cos^2(\frac{m\pi}{M+1}) \quad ; \quad m = 1, 2, \cdots, \frac{M+1}{2} \tag{7.7.5}$$

and the corresponding eigenvectors are given by

$$\vec{X}_m = \left[\cos\left(\frac{m\pi}{M+1}\right)\right]^{j-1} \sin\left[j\left(\frac{m\pi}{M+1}\right)\right] \quad ; \quad m = 1, 2, \cdots, \frac{M+1}{2} \tag{7.7.6}$$

The eigenvectors are quite unlike the Point-Jacobi set. They are no longer symmetrical. They produce waves that are higher in amplitude on one side (the updated side) than they are on the other. Fig. 7.7.2a shows the eigenvectors corresponding to the lowest three eigenvalues for $M = 11$, 23, and 47. Notice that they do not represent a common family for different values of M. The σ_m produced from the λ_m by the explicit Euler method varies with h. The σ, λ relationship for $h = 1$, the optimum stationary case, is shown in Fig. 7.7.2b.



a. First 3 eigenvectors.

b. The σ, λ relation, $M = 5$, $h = 1$.

Figure 7.7.2 – Gauss–Seidel eigensystem.

7.8 The SOR System

If $\beta = 1$ and $2/\omega = x$ in eq (7.1.5), the ODE matrix is $B^{-1}(-1, x, 0)B(1, -2, 1)$. One can show that this can be written in the form given below for $M = 5$. The generalization to any M is fairly clear. The $[H^{-1}A]$ matrix for the SOR method, $[A]_{SOR} \equiv B^{-1}(-1, x, 0)B(1, -2, 1)$, is

$$\frac{1}{x^5}\begin{bmatrix} -2x^4 & x^4 & 0 & 0 & 0 \\ -2x^3 + x^4 & x^3 - 2x^4 & x^4 & 0 & 0 \\ -2x^2 + x^3 & x^2 - 2x^3 + x^4 & x^3 - 2x^4 & x^4 & 0 \\ -2x + x^2 & x - 2x^2 + x^3 & x^2 - 2x^3 + x^4 & x^3 - 2x^4 & x^4 \\ -2 + x & 1 - 2x + x^2 & x - 2x^2 + x^3 & x^2 - 2x^3 + x^4 & x^3 - 2x^4 \end{bmatrix} \tag{7.8.1}$$

Eigenvalues of the system are given by

$$\left.\begin{aligned} \lambda_m &= -1 + \left(\frac{\omega p_m + z_m}{2}\right)^2 \quad ; \quad m = 1, 2, \cdots M \\ z_m &= [4(1 - \omega) + \omega^2 p_m^2]^{1/2} \\ p_m &= \cos[m\pi/(M + 1)] \end{aligned}\right\} \tag{7.8.2}$$

If $\omega = 1$, the system is Gauss–Seidel. If $4(1-\omega)+\omega^2 p_m < 0, z_m$ and λ_m are complex. If ω is chosen such that $4(1-\omega)+\omega^2 p_1^2 = 0, \omega$ is optimum for the stationary case, and the following conditions hold

1) Two eigenvalues are real, equal and defective
2) If M is even, the remaining eigenvalues are complex and occur in conjugate pairs
3) If M is odd, one of the remaining eigenvalues is real and the others are complex occurring in conjugate pairs.

One can easily show that the optimum ω for the stationary case is

$$\omega_{opt} = 2/\left[1 + \sin\left(\frac{\pi}{M+1}\right)\right] \tag{7.8.3}$$

and for $\omega = \omega_{\text{opt}}$

$$\lambda_m = \varsigma_m^2 - 1$$

$$\vec{x}_m = \varsigma_m^{j-1} \sin\left[j\left(\frac{m\pi}{M+1}\right)\right]$$

where

$$\varsigma_m = \frac{\omega_{opt}}{2}\left[p_m + i\sqrt{p_1^2 - p_m^2}\right] \tag{7.8.4}$$

If the explicit Euler method is used to integrate the ODE, $\sigma_m = 1 - h + h\xi_m^2$, and if $h = 1$, the optimum value for the stationary case, the σ, λ relation reduces to that shown in Fig. 7.8.1. This illustrates the well known fact that for optimum stationary SOR all the $|\sigma_m|$ are identical and equal to $\omega_{opt} - 1$.

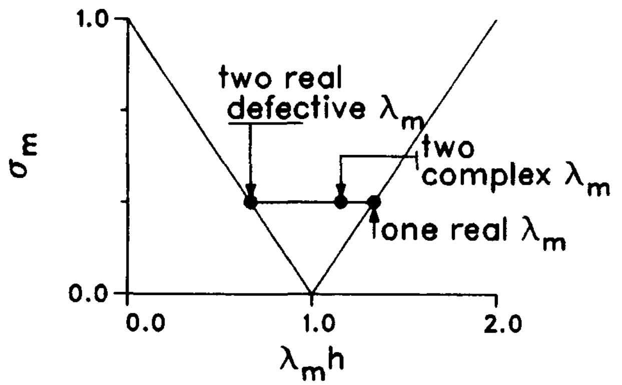


Figure 7.8.1 – The σ, λ relation for optimum stationary SOR, $M = 5$, $h = 1$.

7.9 Solution of the ODE Forms of the Classical Methods

The ODE approach to relaxation can be summarized as follows. The basic equation to be solved came from some time accurate derivation

$$A_b\vec{u} - \vec{f}_b = 0 \tag{7.9.1}$$

This equation is preconditioned in some manner which has the effect of multiplication by a conditioning matrix C giving

$$A\vec{\phi} - \vec{f} - 0 \tag{7.9.2}$$

An iterative scheme is developed to finding the converged, or steady-state, solution of the set of ODE

$$H\frac{d\vec{\phi}}{dt} = A\vec{\phi} - \vec{f} \tag{7.9.3}$$

This solution has the analytic form

$$\vec{\phi}_n = \left\{ \begin{array}{c} \text{Transient term or} \\ \text{error, } \vec{e}_r \end{array} \right\} + \left\{ \begin{array}{c} \text{Steady state term,} \\ \vec{\phi}_\infty \equiv A^{-1}\vec{f} \end{array} \right\} \tag{7.9.4}$$

The three classical methods, Point-Jacobi, Gauss-Seidel, and SOR, are identified for the one-dimensional case by eq (7.1.5) and Table 7.1.1. Their solution for $M = 5$ is written below. For all cases involving linear systems, higher values of M, or higher dimensions, do not alter the fundamental nature of the solutions.

Point-Jacobi: The explicit Euler method is used to solve

$$\frac{d\vec{\phi}}{dt} = \frac{1}{2}\left[B(1, -2, 1)\vec{\phi} - \vec{f}\right] \tag{7.9.5}$$

The eigenvectors are all real and linearly independent, see eq (7.6.1). For $M = 5$ they can be written

$$\vec{x}_1 = \begin{bmatrix} 1/2 \\ \sqrt{3}/2 \\ 1 \\ \sqrt{3}/2 \\ 1/2 \end{bmatrix}, \vec{x}_2 = \begin{bmatrix} \sqrt{3}/2 \\ \sqrt{3}/2 \\ 0 \\ -\sqrt{3}/2 \\ -\sqrt{3}/2 \end{bmatrix}, \vec{x}_3 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}, \vec{x}_4 = \begin{bmatrix} \sqrt{3}/2 \\ -\sqrt{3}/2 \\ 0 \\ \sqrt{3}/2 \\ -\sqrt{3}/2 \end{bmatrix}, \quad \vec{x}_5 = \begin{bmatrix} 1/2 \\ -\sqrt{3}/2 \\ 1 \\ -\sqrt{3}/2 \\ 1/2 \end{bmatrix} \tag{7.9.6}$$

The corresponding eigenvalues are, from eq (7.6.4)

$$\left.\begin{aligned} \lambda_1 &= -1 + \tfrac{\sqrt{3}}{2} = -0.134\cdots \\ \lambda_2 &= -1 + \tfrac{1}{2} \;\;\, = \;\; 0.5 \\ \lambda_3 &= -1 \qquad\;\; = -1.0 \\ \lambda_4 &= -1 - \tfrac{1}{2} \;\;\, = -1.5 \\ \lambda_5 &= -1 - \tfrac{\sqrt{3}}{2} = -1.866\cdots \end{aligned}\right\} \tag{7.9.7}$$

The numerical solution written in full is

$$\begin{aligned} \vec{\phi}_n - \vec{\phi}_\infty = c_1[1 - (1 - \frac{\sqrt{3}}{2})h]^n\vec{x}_1 \\ + c_2[1 - (1 - \frac{1}{2})h]^n\vec{x}_2 \\ + c_3[1 - (1 \qquad)h]^n\vec{x}_3 \\ + c_4[1 - (1 + \frac{1}{2})h]^n\vec{x}_4 \\ + c_5[1 - (1 + \frac{\sqrt{3}}{2})h]^n\vec{x}_5 \end{aligned} \tag{7.9.8}$$

<u>Gauss-Seidel</u>: The explicit Euler method is used to solve

$$\frac{d\vec{\phi}}{dt} = B^{-1}(-1,2,0)\left[B(1,-2,1)\vec{\phi} - \vec{f}\right] \tag{7.9.9}$$

The eigenvectors and principal vectors are all real. For odd M, $(M+1)/2$ of them are linearly independent eigenvectors and $(M+1)/2$ are principal vectors. For $M = 5$ they can be written, see eq (7.7.6)

$$\vec{x}_1 = \begin{bmatrix} 1/2 \\ 3/4 \\ 3/4 \\ 9/16 \\ 9/32 \end{bmatrix}, \vec{x}_2 = \begin{bmatrix} \sqrt{3}/2 \\ \sqrt{3}/4 \\ 0 \\ -\sqrt{3}/16 \\ -\sqrt{3}/32 \end{bmatrix}, \vec{x}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \vec{x}_4 = \begin{bmatrix} 0 \\ 2 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \vec{x}_5 = \begin{bmatrix} 0 \\ 0 \\ 4 \\ -4 \\ 1 \end{bmatrix} \tag{7.9.10}$$

The corresponding eigenvalues are, from eq (7.7.5)

$$\begin{aligned} \lambda_1 &= -1/4 = -0.25 \\ \lambda_2 &= -3/4 = -0.75 \\ \lambda_3 &= -1 \quad = -1.00 \\ &\left.\begin{matrix}(4)\\(5)\end{matrix}\right\} \begin{matrix}\text{Defective, linked to}\\ \lambda_3 \text{ Jordan block}\end{matrix} \end{aligned} \tag{7.9.11}$$

The numerical solution written in full is

$$\begin{aligned} \vec{\phi}_n - \vec{\phi}_\infty &= c_1\left(1 - \frac{h}{4}\right)^n \vec{x}_1 \\ &+ c_2\left(1 - \frac{3h}{4}\right)^n \vec{x}_2 \\ &+ \left[c_3(1-h)^n + c_4 h \frac{n}{1!}(1-h)^{n-1} + c_5 h^2 \frac{n(n-1)}{2!}(1-h)^{n-2}\right]\vec{x}_3 \\ &+ \left[c_4(1-h)^n + c_5 h \frac{n}{1!}(1-h)^{n-1}\right]\vec{x}_4 \\ &+ c_5(1-h)^n \vec{x}_5 \end{aligned} \tag{7.9.12}$$

See Section 8.2 for an interesting experiment with this defective system.

SOR: The explicit Euler method is used to solve

$$\left.\begin{aligned} \frac{d\vec{\phi}}{dt} &= B^{-1}\left(-1, \frac{2}{\omega_0}, 0\right)\left[B(1,-2,1)\vec{\phi} - \vec{f}\right] \\ \text{where } \omega_0 &= 2/\{1 + \sin[\pi/(M+1)]\} \end{aligned}\right\} \tag{7.9.13}$$

For odd M, three of the vectors are real, two of these are eigenvectors and one is a principal vector. The remaining vectors are all complex, linearly independent eigenvectors. For $M = 5$ they can be written, see eq (7.8.4)

$$\vec{x}_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/3 \\ 1/6 \\ 1/18 \end{bmatrix}, \vec{x}_2 = \begin{bmatrix} -6 \\ 9 \\ 16 \\ 13 \\ 6 \end{bmatrix}, \vec{x}_{3,4} = \begin{bmatrix} \sqrt{3}(1 \qquad\quad)/2 \\ \sqrt{3}(1 \pm \ i\sqrt{2})/6 \\ 0 \\ \sqrt{3}(5 \pm \ i\sqrt{2})/54 \\ \sqrt{3}(7 \pm 4i\sqrt{2})/162 \end{bmatrix}, \vec{x}_5 = \begin{bmatrix} 1 \\ 0 \\ 1/3 \\ 0 \\ 1/9 \end{bmatrix} \tag{7.9.14}$$

The corresponding eigenvalues are

$$\begin{aligned} \lambda_1 &= -2/3 \\ (2) & \qquad \text{Defective linked to} \quad \lambda_1 \\ \lambda_3 &= -(10 - 2\sqrt{2}i)/9 \\ \lambda_4 &= \ (10 + 2\sqrt{2}i)/9 \\ \lambda_5 &= -4/3 \end{aligned} \tag{7.9.15}$$

The numerical solution written in full is

$$\begin{aligned} \vec{\phi}_n - \vec{\phi}_\infty = & [c_1(1 - 2h/3)^n + c_2 nh(1 - 2h/3)^{n-1}]\vec{x}_1 \\ & + c_2(1 - 2h/3)^n \vec{x}_2 \\ & + c_3[1 - (10 - 2\sqrt{2}i)h/9]^n \vec{x}_3 \\ & + c_4[1 - (10 + 2\sqrt{2}i)h/9]^n \vec{x}_4 \\ & + c_5(1 - 4h/3)^n \vec{x}_5 \end{aligned} \tag{7.9.16}$$

8. EIGENVECTOR ANNIHILATION

8.1 Introduction

Let us consider a conditioned matrix that has a set of linearly independent eigenvectors. We have seen that the ODE approach to relaxation first casts the process into a form which has the exact analytical solution

$$\vec{\phi}_n = c_1\left(e^{\lambda_1 h}\right)^n \vec{x}_1 + \cdots + c_m\left(e^{\lambda_m h}\right)^n \vec{x}_m + \cdots + c_M\left(e^{\lambda_M h}\right)^n \vec{x}_M + \vec{\phi}_\infty \qquad (8.1.1)$$

and then awaits the choice of a numerical integration method to relate the λ eigenvalues of the differential equation to the σ eigenvalues of the finite difference equations. If there is one σ for each λ, we find

$$\vec{\phi}_n = \sum_{m=1}^{M} \{c_m\} \left\{ \prod_{n=1}^{N} \sigma_m(\lambda_m h_n) \right\} \vec{x}_m + \vec{\phi}_\infty \qquad (8.1.2)$$

where we have taken advantage of the fact that the step size h can be varied from step to step. Clearly the problem is to remove the eigenvector content of the transient solution which is identical to the error of the relaxation process. Three ways of doing this are immediately evident:

(1) Make a good initial guess, that is, make $c_m \approx 0$.

(2) Make all $|\sigma| < 1$ and iterate many times; that is, make N large.

(3) Try to choose h_n so that $|\sigma_m| \approx 0$ at least once in the process.

Of these, (1) is obvious if it is possible, (2) is the typical approach of stationary methods, and (3) is the basis for nonstationary methods.

In addition to the above, there are two more subtle ways for diminishing the error:

(4) Mixing the eigenstructure by applying appropriate conditioning matrices and restarting the process. This makes a sophisticated use of (1) above.

(5) Using multiple grids to surface and destroy the eigenvectors associated with the lower space-frequencies.

These two concepts are discussed in Sections 9 and 10.

8.2 Selective Eigenvector Annihilation

One of the most direct ways to eliminate the eigenvector content of the error in a relaxation scheme is to set $h = -1/\lambda_m$ for every λ_m eigenvalue in the system and iterate M steps. This is an extremely poor strategy in practical application but it is worth discussing from a theoretical viewpoint. It is of course, in principal, a direct solution.

Consider the Point–Jacobi solution given by eq (7.9.8). In the 5-point mesh, 5 sweeps through the mesh with h equal to 7.464..., 2, 1, 0.666..., 0.536... would annihilate all 5 of the eigenvectors. In carrying out the calculation, the worst situation from computer considerations would occur for the fifth eigenvector. Its initial amplitude would be multiplied consecutively by $(-12.93\ldots)$ $(-2.73\ldots)$ $(-.866\ldots)$ $(-.245\ldots)$ (0). In this trivial case the first two multiplies would not cause trouble, but in application to large systems such a strategy would be highly unstable because computer hardware limits the accuracy to which numbers can be represented. Notice, however, that the eigenvectors associated with the highest $(M-1)/2$ eigenvalues can all be annihilated in this manner without amplifying any vector at any step. Only when annihilating the eigenvectors associated with the lowest $(M-1)/2$ eigenvalues are the amplitudes of other vectors amplified.

Consider next the Gauss–Seidel solution given by eq (7.9.12). In an M point mesh, M sweeps are still required to annihilate all the possible error vectors. If M is odd, $(M+1)/2$ of the sweeps with $h = 1$ are required to remove the principal vectors associated with the defective eigenvalues. The remaining annihilations require $(M-1)/2$ sweeps with $h = -1/\lambda_m$. One can observe the initial growth of the principal vectors for $h \neq 1$ caused by the factor $h^k(n)(n-1)\cdots(n-k+1)/k!$ by performing simple numerical experiments. Since the eigenvalues of the Gauss–Seidel system are all real and there are only half as many of them as there are in Point–Jacobi, one might question the second sentence in this paragraph and be led to the conclusion that the reduction in the number of eigenvalues could be used to advantage in annihilation. That such is not the case can be demonstrated by using the simple program written in BASIC and presented in Fig. 8.2.1. Recall the system of eigenvectors and principal vectors give by eq (7.9.10) and consider the pathological nature of eq (7.9.12) when $h = 1$ and the exponent of $(1-h)$ is zero. Run the BASIC program using $\vec{x}_5 = [0,0,4,-4,1]^T$ as input. Notice that after one iteration the second principal vector $\vec{x}_4$ appears. A second iteration produces the eigenvector $\vec{x}_3$ and a third $\left(\frac{(M+1)}{2}\right)$ iteration is required to annihilate the entire Jordan block.

```
10 DIM U(6)
20 PRINT "ENTER INITIAL VALUES"
30 INPUT U(1),U(2),U(3),U(4),U(5)
40 FOR J=1 TO 5
50 U(J) = .5*(U(J-1) + U(J+1))
60 NEXT J
70 PRINT U(1),U(2),U(3),U(4),U(5)
80 PRINT
90 INPUT "TYPE Y TO ITERATE AGAIN, OTHERWISE NEW INI-
TIAL V";CH$
100 IF CH$ = "Y" GOTO 40
110 GOTO 20
120 END
```

Figure 8.2.1 – BASIC program to illustrate defective nature of Gauss-Seidel eigensystem.

Input 0,0,4,-4,1 → 0,2,-1,0,0 → 1,0,0,0,0 → 0

Finally, consider the SOR solution given by eq (7.9.16). It is again true that M sweeps will annihilate all of the possible error vectors, only now one must use complex values for the step size h. In present day computers this requires more arithmetic and storage. Nevertheless, it may have features that make it attractive for practical use. This is being investigated.

We now make some observations. First of all, it is usually unnecessary to remove completely the error from the general solution. We consider it satisfactory to reduce the error below some threshold. Second, the exact values of λ_m are generally unknown and not practical to compute. It is therefore unnecessary and impractical to set $h = -1/\lambda_m$ for $m = 1, 2, \cdots, M$; a few well chosen h's will greatly reduce the amplitudes of whole clusters of eigenvectors associated with nearby λ's in the λ_m spectrum. For example, in the Point–Jacobi case on the model problem using only 3 h's that are given by the $-1/\lambda_m$ at the two ends and in the middle of the highest half of the $|\lambda_m|$ range, reduces the amplitudes of all the eigenvectors associated with that range to about 1% (or less) of their initial values. The amplitudes of all the other eigenvectors coupled into the system are also reduced, but some by very little. Fortunately, this phenomenon is not sensitive to the particular choice of h. The values $h = \frac{1}{2}, \frac{2}{3}, 1$ are about as good as the Chebyshev choice $h = 0.517\ldots, \frac{2}{3}, 0.937\ldots$ given in Fig. 7.4.1. These observations are much more significant when extended to 2 and 3 dimensions. In these cases 3/4 and 7/8 of the error vectors associated with the highest eigenvalues are reduced to about 1% of their initial values by 3 nonstationary Point–Jacobi sweeps with

$$h = 1/|\lambda|_{max}, 1/\frac{3}{4}(|\lambda|_{max}), 1/(\frac{1}{2}|\lambda|_{max}) \qquad (8.2.1)$$

if the eigenvalues are fairly evenly distributed between 0 and λ_{max}. This leads to the concept of selectively annihilating clusters of eigenvectors from the error terms as a part of a total iteration process. This is also the basis for the success of the mixing and multigrid methods discussed below.

8.3 Eigenvector and Eigenvalue Identification with Space Frequencies

Consider the eigensystem of the model matrix $B(\frac{1}{2}, -1, \frac{1}{2})$. The eigenvalues and eigenvectors for $M = 5$ are given in eqs (7.9.6) and (7.9.5), respectively. Notice that as the magnitudes of the eigenvalues increase the space-frequency (number of sign changes) of the corresponding eigenvectors also increase. This has a rational explanation from the origin of the banded matrix. Note that

$$\frac{\partial^2}{\partial x^2} \sin(mx) = -m^2 \sin(mx) \qquad (8.3.1)$$

and recall that

$$\delta_{xx}\vec{\phi} = \frac{1}{\Delta x^2} B(1, -2, 1)\vec{\phi} = X\left[\frac{1}{\Delta x^2} D(\vec{\lambda})\right]X^{-1}\vec{\phi} \qquad (8.3.2)$$

We have seen that $X^{-1}\vec{\phi}$ represents a sine transform, and $X\vec{\phi}$, a sine synthesis. Therefore, the operation $\frac{1}{\Delta x^2} D(\vec{\lambda})$ represents the numerical approximation of the multiplication of the appropriate sine wave by the negative square of its frequency, $-m^2$. One finds that

$$\frac{1}{\Delta x^2}\lambda_m = \left(\frac{M+1}{\pi}\right)^2\left[-2 + 2\cos\left(\frac{m\pi}{M+1}\right)\right] \approx -m^2, m << M \qquad (8.3.3)$$

Hence, the correlation of large magnitudes of λ_m with high space-frequencies (see e.g., eq (7.6.1a)) is to be expected for these particular matrix operators. However, this correlation is not necessary in general. In fact, the complete counterexample of the above association is contained in the eigensystem for $B(\frac{1}{2}, 1, \frac{1}{2})$. For this matrix one finds, from Section 5, exactly the opposite behavior. This is illustrated in eq (8.3.4)

$$X = \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} & 1 & \frac{\sqrt{3}}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & 0 & -\frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ 1 & 0 & -1 & 0 & 1 \\ \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} & 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ \frac{1}{2} & -\frac{\sqrt{3}}{2} & 1 & -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix} \tag{8.3.4}$$

$$\begin{array}{ccccc} \vec{x}_1 & \vec{x}_2 & \vec{x}_3 & \vec{x}_4 & \vec{x}_5 \end{array}$$

$$\boxed{\begin{array}{l}\text{corresponding}\\ \text{eigenvalue}\end{array}} = 1.866\ldots \quad 1.5 \quad 1 \quad 0.5 \quad .134\ldots$$

Notice the eigenvectors with the lowest space-frequencies correspond to the eigenvalues with maximum amplitude. Since the matrix structures that we are working with are subject to quite arbitrary preconditioning, this "inverse" structure can have practical implication.

9. EIGENSYSTEM MIXING

The eigenvectors of $B(1,b,1)$ are given by

$$(x_{jm}) = \sin\left[j\left(\frac{m\pi}{M+1}\right)\right] \tag{9.1}$$

and their structure does not depend upon the value of b. Therefore, the eigenvectors of the product $B(1,b_1,1)B(1,b_2,1)$ are also given by eq (8.2.1), and in general

$$(x_{jm}) = \sin\left[j\left(\frac{m\pi}{M+1}\right)\right] \quad \text{for} \quad \prod_{k=1}^{k} B(1,b_k,1) \tag{9.2}$$

However, the eigenvalues of $B(1,b,1)$ do depend upon b, see eq (5.1.2), and when matrices have a common set of eigenvectors, the eigenvalues of their product are the product of their eigenvalues. Therefore,

$$\lambda_m = \prod_{k=1}^{k}\left[b_k - 2\cos\left(\frac{m\pi}{M+1}\right)\right] for \prod_{k=1}^{k} B(1,b_k,1) \tag{9.3}$$

This provides a way for constructing a process that modifies a matrix without changing its eigenvectors while changing the identities of the eigenvalues attached to them. Such a process is one way to mix the eigenvalue-eigenvector structure of an iteration procedure so that a nonstationary method can remove more of the error without amplifying any of the vectors during the calculation.

A simple example illustrates the point. Consider the linear averaging operator given by

$$\tilde{\phi}_j = \phi_{j-1} + \phi_j + \phi_{j+1} \tag{9.4}$$

when written in point operator form. If used twice on the 1-D model Dirichlet equation in a nonstationary Point-Jacobi sequence, we have

$$\vec{\phi}_{n+1} = \vec{\phi}_n + h_n B(1,1,1)^2\left[B(1,-2,1)\vec{\phi}_n - \vec{f}_n\right] \tag{9.5}$$

the λ_m eigenvalue spectrums of the matrices used above are given analytically by

$$\left.\begin{aligned} \lambda_m &= -2[1 - \cos\theta_m] \quad \text{for} \quad B(1,-2,1) \\ \lambda_m &= -2[1 - \cos(3\theta_m)] \quad \text{for} \quad B(1,1,1)^2 B(1,-2,1) \\ \text{where} \quad \theta_m &= m\pi/(M+1) \end{aligned}\right\} \tag{9.6}$$

and are shown in Fig. 9.1. Remember that the eigenvectors of the two matrices indexed, in both cases, from 1 to M are identical. It is interesting to note that for $M = 7$

$$B(1,1,1)^2 B(1,-2,1) = \begin{bmatrix} -2 & -1 & 0 & 1 & & & \\ -1 & -2 & 0 & 0 & 1 & & \\ 0 & 0 & -2 & 0 & 0 & 1 & \\ 1 & 0 & 0 & -2 & 0 & 0 & 1 \\ & 1 & 0 & 0 & -2 & 0 & 0 \\ & & 1 & 0 & 0 & -2 & -1 \\ & & & 1 & 0 & -1 & -2 \end{bmatrix} \tag{9.7}$$

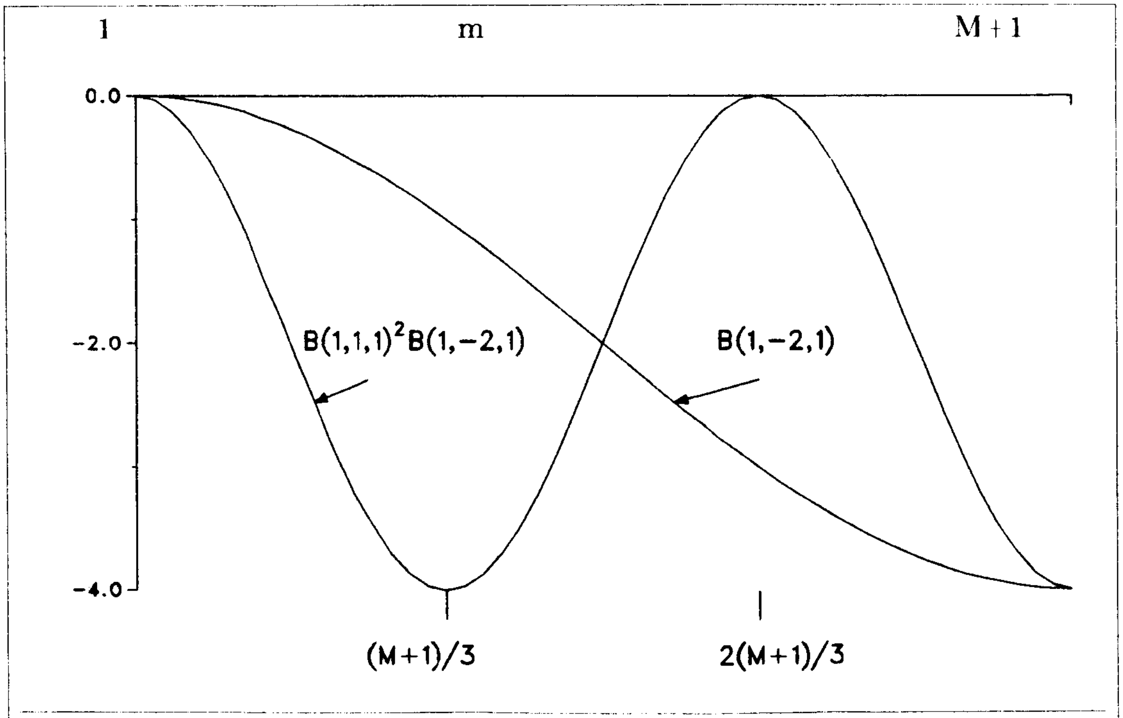


Figure 9.1 – λ_m spectrums of $B(1,-2,1)$ and $B(1,1,1)^2$ times $B(1,-2,1)$ for odd M.

It is not the purpose of this report to invent optimum relaxation techniques, but rather to present some fundamental concepts by which they can be constructed. However, it is instructive to consider two ways in which the model equation

$$B(1,-2,1)\vec{\phi} - \vec{f} = 0 \tag{9.8}$$

can be relaxed. First by a nonstationary Point–Jacobi or Richardson method given by

$$\vec{\phi}_{n+1} = \vec{\phi}_n + h_n\left[B(1,-2,1)\vec{\phi}_n - \vec{f}_n\right] \tag{9.9}$$

which we refer to by $PJ(h_n)$, and then by a combination of this method and the method given in eq (9.5) which we refer to by $VM(h_n)$.

Suppose the error content in the initial guess for $\vec{\phi}$ in eq (9.8) is expressed in terms of the amplitudes of the eigenvectors in (9.1), and suppose that initially each amplitude is given unit weight. That is

$$\begin{aligned} \vec{\phi}_0(x) &= \sum_{m=1}^{M} \sin\left[m(j\Delta x)\right] \\ \text{where} \quad x = j\Delta x \quad &\text{and} \quad \Delta x = \pi/(M+1)\,, \quad \text{and} \\ \widehat{\phi}_0(m) &= \frac{2}{M+1}\sum_{j=1}^{M} \phi_0(x)\sin\left[j(m\Delta x)\right] = (\vec{1}) \end{aligned} \tag{9.10}$$

Apply the Point–Jacobi sequence three times forming

$$\vec{\phi}_3 = PJ\left(\frac{1}{4}\right)\cdot PJ\left(\frac{1}{3}\right)\cdot PJ\left(\frac{1}{2}\right)\vec{\phi}_0 \tag{9.11}$$

The error content vs. eigenvector number (in the sense defined in eq (7.6.1a)) is shown after each iteration in Fig. 9.2, where $\widehat{\phi}$ represents $\vec{\phi}$ in vector space. This is, of course, very similar to Fig. 7.4.1 for Richardson's 3–step method using the Chebyshev h_n sequence, except here the ordinate is the wave number.

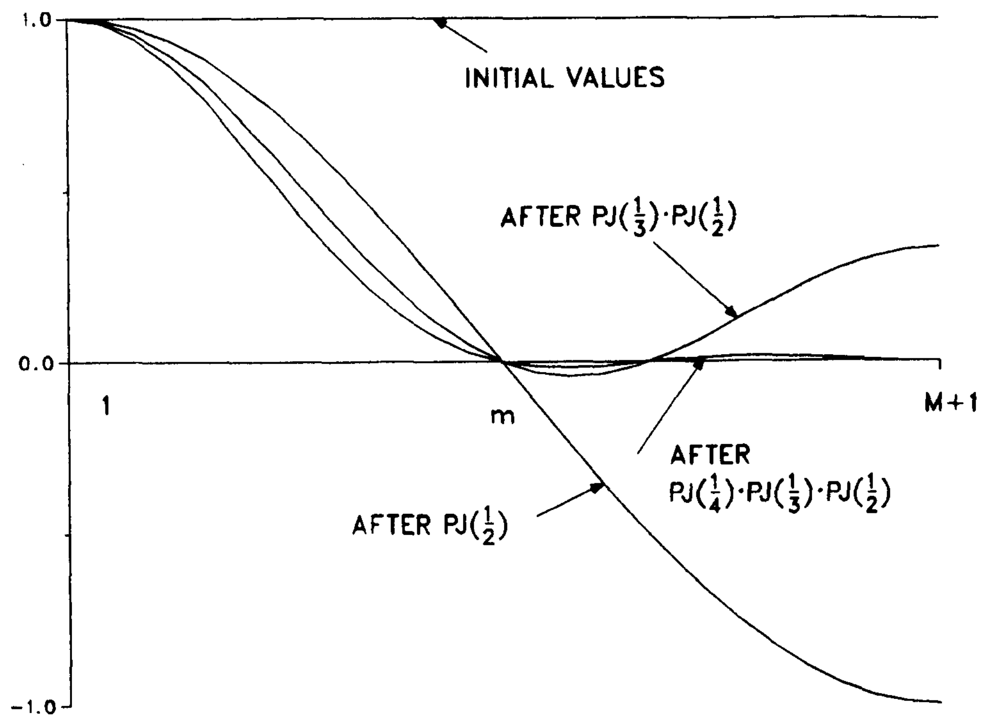


Figure 9.2 – A three-step Richardson sequence, eq (9.11).

Now consider a combination of the PJ and VM sequences given by

$$\vec{\phi}_4 = PJ\left(\frac{1}{3}\right) \cdot PJ\left(\frac{1}{3}\right) \cdot VM\left(\frac{1}{3}\right) \cdot VM\left(\frac{1}{4}\right)\vec{\phi}_0 \qquad (9.12)$$

where the initial conditions are again given by eq (9.10). For the first VM sweep with $h = 1/4$ the amplitudes of the eigenvectors associated with eigenvalues close to -4 are greatly reduced. These occur at $m = M$ and $M/3$. These vicinities are further reduced by the second VM sweep with $h = 1/3$. Spikes of high amplitudes still remain at $m = 0$ and $2M/3$. The spike at $2M/3$ is easily eliminated by the two PJ sweeps using $h = 1/3$ since around $m = 2M/3$ the eigenvalues of the PJ matrix are -3, see Fig. 9.3.

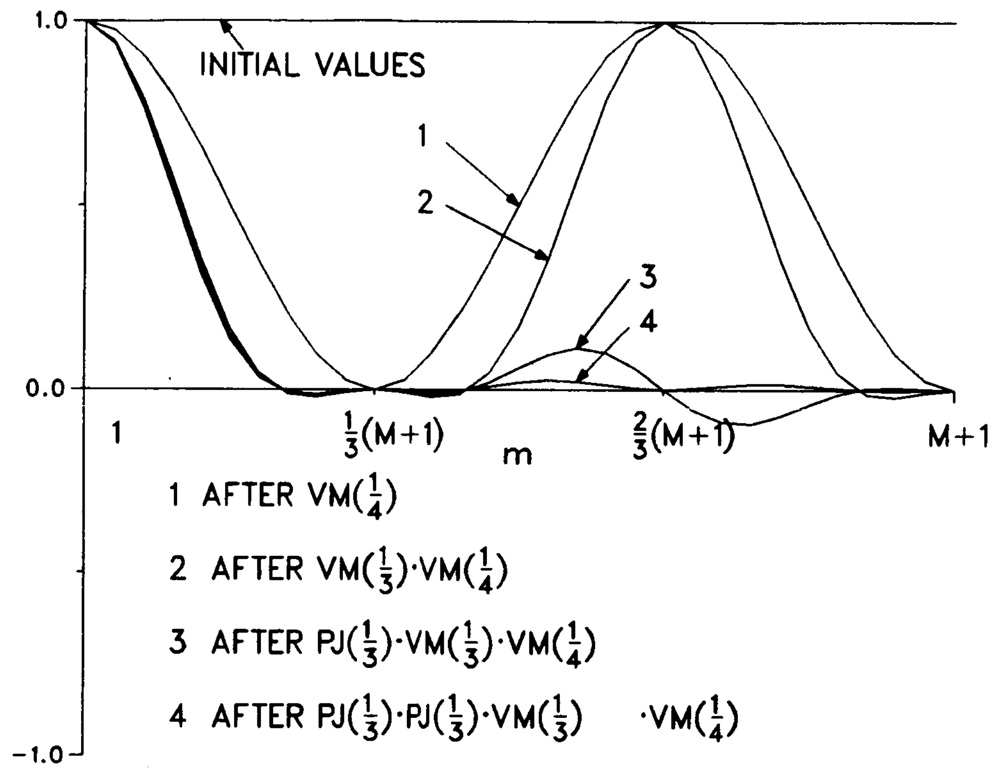


Figure 9.3 – A four-step mixed sequence.

The results of the two processes are compared in Fig. 9.4. A significant portion of the lower frequency content has been removed from the error content by the mixing strategy. The additional work required to do this would have to be compared with alternative methods for doing the same thing; but remember, the constraint imposed in constructing the results in Figures 9.2, 9.3, and 9.4 was that at no time was there any amplification of any eigenvector in the error content.

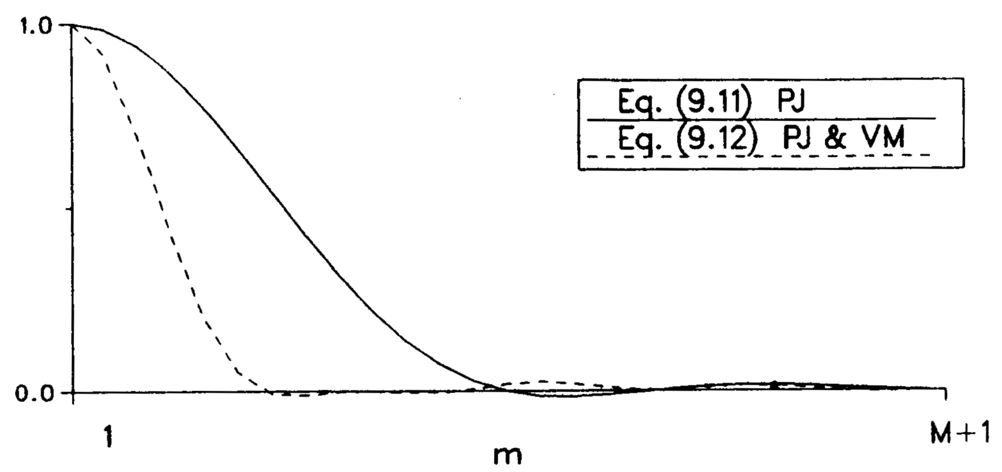


Figure 9.4 – Comparison between a Richardson method and a mixed method.

10. MULTIGRID STRATEGIES

The idea of systematically using sets of coarser grids to accelerate the convergence of iteration schemes that arise from the numerical solution to partial differential equations was made popular in this country by the work of Brandt, see ref. 7. There are many variations of the process, which is by no means unique, and many viewpoints of the underlying theory. The viewpoint presented here is a natural extension of the concepts discussed above.

First of all we assume that the difference equations representing the basic partial differential equations are in a form that can be related to a matrix which has certain basic properties. This form can be arrived at "naturally" by simply replacing the derivatives in the PDE with difference schemes, as in the example given by eq (1.2.2), or it can be "contrived" by further conditioning, as in the examples given by eqs (2.2.9) and (2.2.10). These basic properties are:

(1) The eigenvalues, λ_m, of the matrix are all real and negative.
(2) The λ_m are fairly evenly distributed between their maximum and minimum values.
(3) The eigenvectors associated with the eigenvalues having largest magnitudes can be correlated with high frequencies on the differencing mesh.

(10.1)

These condiiions are sufficient to ensure the validity of the process described next.

Having preconditioned (if necessary) the basic finite differencing scheme by a procedure equivalent to the multiplication by a matrix C, we are led to the starting formulation

$$C[A_b\vec{\phi}_\infty - \vec{f}_b] = 0 \tag{10.2}$$

where the matrix formed by the product CA_b has the three properties in (10.1). In eq (10.2) the vector $\vec{f}_b$ represents the boundary conditions and the forcing function, if any, and $\vec{\phi}_\infty$ is a vector representing the desired exact solution. We start with some initial guess for $\vec{\phi}_\infty$ and proceed through n iterations making use of some iterative process that greatly reduces the amplitudes of the eigenvectors associated with the eigenvalues in the range between $|\lambda|_{max}$ and $\frac{1}{2}|\lambda|_{max}$. We do not attempt to develop an optimum procedure here, but for clarity we suppose that the three-step Richardson method illustrated in Fig. 7.4.1 is used. At the end of the three steps we find $\vec{r}$, the residual, where

$$\vec{r} = C[A_b\vec{\phi} - \vec{f}_b] \tag{10.3}$$

We recall (Section 6) that the $\vec{\phi}$ used to compute $\vec{r}$ is composed of the exact solution $\vec{\phi}_\infty$ and the error $\vec{e}$ in such a way that

$$\left.\begin{aligned} & A\vec{e} - \vec{r} = 0 \\ \text{where} \qquad & A \equiv CA_b \end{aligned}\right\} \tag{10.4}$$

If one could solve eq (10.4) for $\vec{e}$ then

$$\vec{\phi}_\infty = \vec{\phi} - \vec{e} \tag{10.5}$$

Now we can write the exact solution for $\vec{e}$ in terms of the eigenvectors of A, and the σ eigenvalues of the Richardson process in the form (10.6)

$$\vec{e} = \sum_{m=1}^{M/2} c_m\vec{x}_m \prod_{n=1}^{3} [\sigma(\lambda_m h_m)] + \underbrace{\sum_{m=M/2+1}^{M} c_m\vec{x}_m \prod_{n=1}^{3} [\sigma(\lambda_m h_m)]}_{\text{very low amplitude}} \tag{10.6}$$

Combining the properties of the Richardson algorithm and the conditions in (10.1), we can be sure that the high frequency content of $\vec{e}$ has been greatly reduced (about 1% or less of its original value in the initial guess).

Next we construct a permutation matrix which separates a vector into two parts, one containing the odd entries, and the other the even entries of the original matrix (or any other appropriate sorting which is consistent with the interpolation approximation to be discussed below). For example

$$\begin{bmatrix} e_2 \\ e_4 \\ e_6 \\ e_1 \\ e_3 \\ e_5 \\ e_7 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{bmatrix} \quad ; \quad \begin{bmatrix} \vec{e}_e \\ \vec{e}_o \end{bmatrix} = P\vec{e} \tag{10.7}$$

Multiply eq (10.4) from the left by P and, since a permutation matrix has an inverse which is its transpose, we can write

$$PA[P^{-1}P]\vec{e} = P\vec{r} \tag{10.8}$$

The operation PAP^{-1} partitions the A matrix to form

$$\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \begin{bmatrix} \vec{e}_e \\ \vec{e}_0 \end{bmatrix} = \begin{bmatrix} \vec{r}_e \\ \vec{r}_0 \end{bmatrix} \tag{10.9}$$

Notice that

$$A_1\vec{e}_e + A_2\vec{e}_0 = \vec{r}_e \tag{10.10}$$

is an exact expression. At this point we make our one crucial assumption. It is that there is some connection between $\vec{e}_e$ and $\vec{e}_o$ brought about by the smoothing property of the Richardson relaxation procedure. Since the top half of the frequency spectrum has been removed, it is reasonable to suppose that the odd points are the average of the even points. For example

$$\begin{aligned} e_1 &\approx \frac{1}{2}(e_a + e_2) \\ e_3 &\approx \frac{1}{2}(e_2 + e_4) \\ e_5 &\approx \frac{1}{2}(e_4 + e_6) \qquad \text{or } \vec{e}_0 = A_2'\vec{e}_e \\ e_7 &\approx \frac{1}{2}(e_6 + e_b) \end{aligned} \tag{10.11}$$

It is important to notice that e_a and e_b represent errors on the boundaries where the error is zero if the boundary conditions are given. It is also important to notice that we are dealing with the relation between $\vec{e}$ and $\vec{r}$ so the original boundary conditions and forcing function (which are contained in $\vec{f}$ in the basic formulation) no longer appear in the problem. Hence, no aliasing of these functions can occur in subsequent steps. Finally, notice that, in this formulation, the averaging of $\vec{e}$ is our only approximation, no operations on $\vec{r}$ are required or justified.

If the boundary conditions are Dirichlet, e_a and e_b are zero, and one can write for the example case

$$A_2' = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{10.12}$$

With this approximation eq (10.8) reduces to

$$\begin{aligned} & A_c\vec{e}_e - \vec{r}_e = 0 \\ \text{where} \qquad & A_c = [A_1 + A_2A_2'] \end{aligned} \tag{10.13}$$

The form of A_c, the matrix on the coarse mesh, is completely determined by the choice of the permutation matrix and the interpolation approximation. If the original A had been $B(1,-2,1)$, our 7-point example would produce

$$PAP^{-1} = \begin{bmatrix} -2 & & & 1 & 1 & & \\ & -2 & & & 1 & 1 & \\ & & -2 & & & 1 & 1 \\ 1 & 1 & & -2 & & & \\ 1 & 1 & & & -2 & & \\ & 1 & 1 & & & -2 & \\ & & 1 & & & & -2 \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \tag{10.14}$$

and eq (10.13) gives

$$\overbrace{\begin{bmatrix} -2 & & \\ & -2 & \\ & & -2 \end{bmatrix}}^{A_1} + \overbrace{\begin{bmatrix} 1 & 1 & & \\ & 1 & 1 & \\ & & 1 & 1 \end{bmatrix}}^{A_2} \cdot \frac{1}{2} \overbrace{\begin{bmatrix} 1 & & \\ 1 & 1 & \\ & 1 & 1 \\ & & 1 \end{bmatrix}}^{A_2'} = \overbrace{\begin{bmatrix} -1 & 1/2 & \\ 1/2 & -1 & 1/2 \\ & 1/2 & -1 \end{bmatrix}}^{A_C} \tag{10.15}$$

This process is deceptively simple. We started with the equation $B(1,-2,1)\vec{e} = \vec{r}$ on the fine mesh and reduced the problem to the equation $\frac{1}{2}B(1,-2,1)\vec{e}_e = \vec{r}_e$ on the next coarser mesh. It appears as if the data on the odd points had been ignored altogether and a scaling factor had arbitrarily appeared. Such is not the case, however, and except for the assumption in eq (10.11) the process is quite rigorous.

If the boundary conditions are mixed Dirichlet Neumann, A in the 1-D model equation is $B(1,\vec{b},1)$ where $\vec{b} = [-2,-2,...,-2,-1]^T$. The eigensystem is given by eq (5.6.2). It is easy to show that the high space frequencies still correspond to the eigenvalues with high magnitudes, and, in fact, all of the conditions in (10.1) are met. However, the eigenvector structure is different from that shown in Fig. 7.6.1 for Dirichlet conditions. In the present case they are given by

$$x_{jm} = \sin\left[j\left(\frac{(2m-1)\pi}{2M+1}\right)\right] \quad ; \quad m = 1,2,\cdots,M \qquad (10.16a)$$

and are illustrated in Fig. 10.1. All of them go through zero on the left (Dirichlet) side, and all of them reflect on the right (Neumann) side, being symmetrical about the point $m = +\frac{1}{2}$ where $x = \pi$ and their magnitude is 1.

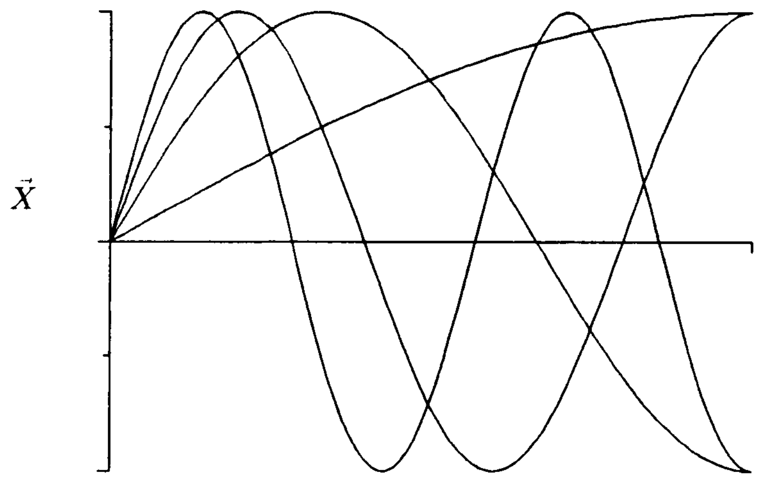


Figure 10.1 - Eigenvectors for the mixed Dirichlet-Neumann case.

For Neumann conditions, the interpolation formula in eq (10.11) must be changed. In the particular case illustrated in Fig. 10.1, e_b is equal to e_M. If Neumann conditions are on the left, $e_a = e_1$. When $e_b = e_M$, the example in eq (10.12) changes to

$$A_2' = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \tag{10.16b}$$

The permutation matrix remains the same and both A_1 and A_2 in the partitioned matrix PAP^{-1} are unchanged (only A_4 is modified by putting -1 in the lower right element). Therefore, we can construct the coarse matrix from

$$\overbrace{\begin{bmatrix} -2 & & \\ & -2 & \\ & & -2 \end{bmatrix}}^{A_1} + \overbrace{\begin{bmatrix} 1 & 1 & & \\ & 1 & 1 & \\ & & 1 & 1 \end{bmatrix}}^{A_2} \cdot \frac{1}{2} \overbrace{\begin{bmatrix} 1 & & \\ 1 & 1 & \\ & 1 & 1 \\ & & 2 \end{bmatrix}}^{A_2'} = \overbrace{\begin{bmatrix} -1 & 1/2 & \\ 1/2 & -1 & 1/2 \\ & 1/2 & -1/2 \end{bmatrix}}^{A_C} \tag{10.17}$$

which gives us what we might have "expected" and shows us that the process is recursive.

The remaining steps required to complete an entire multigrid process are relatively straightforward, but they vary depending on the problem and the user. The reduction can be, and usually is, carried to even coarser grids before returning to the finest level. However, in each case the appropriate permutation matrix and the interpolation approximation define both the down- and up-going paths. The details of finding optimum technique are, obviously, quite important but they are not discussed here.

REFERENCES

1 Ames, W. F.: Numerical Methods for Partial Differential Equations. Barnes & Noble, Inc.. New York, 1969.

2 Boole, G.: Calculus of Finite Differences, Fourth ed., ed. J. F. Moulton. Chelsea Publishing Co., New York.

3 Dahlquist, G.; and Björck, Å.: Numerical Methods. (Ned Anderson, transl.) Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1974.

4 Forsythe, G. E.; and Wasow, W. R.: Finite-Difference Methods for Partial Differential Equations. John Wiley & Sons, Inc., New York, 1960.

5 Isaacson, E.; and Keller, H. B.: Analysis of Numerical Methods. John Wiley & Sons, New York, 1966.

6 Young, D. E.; and Gregory, R. T.: A Survey of Numerical Mathematics. vol. II. Addison-Wesley Publishing Co., Reading, Massachusetts, 1973.

7 Hackbusch, W.; and Trottenberg, U., eds.: Multigrid Methods. Lecture Notes in Mathematics, vol. 960. Springer-Verlag, Berlin, 1982.

Report Documentation Page

1. Report No. TM 88377	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle

Eigensystem Analysis of Classical Relaxation Techniques With Applications to Multigrid Analysis

5. Report Date

March 1987

6. Performing Organization Code

7. Author(s)

Harvard Lomax and Catherine Maksymiuk

8. Performing Organization Report No.

A-86432

10. Work Unit No.

505-60

9. Performing Organization Name and Address

Ames Research Center
Moffett Field, CA 94035

11. Contract or Grant No.

12. Sponsoring Agency Name and Address

National Aeronautics and Space Administration
Washington, D. C. 20546

13. Type of Report and Period Covered

Technical Memorandum

14. Sponsoring Agency Code

15. Supplementary Notes

Point of Contact: Catherine Maksymiuk, Ames Research Center, M/S 202A-1, Moffett Field, CA 94035, (415) 694-4737 or FTS 464-4737

16. Abstract

Classical relaxation techniques are related to numerical methods for solution of ordinary differential equations. Eigensystems for Point-Jacobi, Gauss-Seidel, and SOR methods are presented. Solution techniques such as eigenvector annihilation, eigensystem mixing, and multigrid methods are examined with regard to the eigenstructure.

17. Key Words (Suggested by Author(s)) Numerical methods Relaxation methods Eigensystems Multigrid Eigenvector annihilation	18. Distribution Statement Unclassified-Unlimited Subject Category-64

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of pages	22. Price
Unclassified	Unclassified	82	A05

For sale by the National Technical Information Service, Springfield, Virginia 22161