In the previous chapter we reviewed the formation of difference equations as approximations to differential equations. The purpose of the current chapter is to discuss a variety of methods of obtaining numerical solutions to the approximating equations. As we shall presently see, the actual procedure adopted plays a profound influence on whether or not an approximate solution can be obtained.

For our present interest, we shall concentrate on numerical procedures suitable for digital computers. The reason for the emphasis is evident when one considers the labor involved in attempting to solve nontrivial problems with any degree of accuracy. We shall first consider the problem of numerical integration. Many of the results obtained can be carried over to solving ordinary differential equations. However, the major emphasis of the chapter will be upon solution of partial difference equations.

### 3.1 Numerical Integration

In the introductory example of Chapter II, we considered the problem of approximate integration. The results were obtained in terms of differences of first order. In this section we generalize the previous example. Consider the definite integral

\[ I = \int_a^b dx f(x), \quad (3.1.1) \]

where \( f(x) \) is known, either as an analytic expression or a table of
values at points \( x_j \). The basis for obtaining an approximate solution is to write the integral as

\[
I = \int_a^b dx f(x) \approx \sum_j f(x_j)w_j \Delta x_j ,
\]

(3.1.2)

where the \( w_j \) are weights associated with the integration process. The approximation (3.1.2) is known as a quadrature formula. One possible procedure for generating the quadrature formula is to expand \( f(x) \) in a power series in \( x \) (if possible) and evaluate coefficients using known values of \( f(x_j) \). The usual procedure in this vein leads to the so-called Lagrange interpolation formulas. The integration of the Lagrange formulas of various orders in turn lead to the Newton-Cotes integration formulas. We shall not proceed in this manner, but shall derive equivalent expressions directly in terms of difference operators.

In the previous chapter differential equations and difference equations were discussed. These were solved by integration. In particular, various differential operators were represented, sometimes approximately, by difference operators. In particular, Eq. (2.3.10) shows us an exact relation between the differential operator \( \frac{d}{dx} \) and the forward difference operator \( \Delta \). This relation can be exploited now to express the integral operator in terms of difference operators. Integration may be regarded as the inverse of differentiation. In principle, then, we need only to invert the operator \( \frac{1}{h} \ln(1 + \Delta) \). To proceed we observe, again from the previous chapter, that

\[
\Delta f_j = f(x_{j+1}) - f(x_j) = \frac{d}{dx} \int_{x_j}^{x_{j+1}} dx f(x) = \frac{d}{dx} \int_{x_j}^{x_{j+1}} dx f(x),
\]

(3.1.3)

so that from Eq. (2.3.10) we find

\[
\int_{x_j}^{x_{j+1}} dx f(x) = \frac{h \Delta f_j}{\ln(1 + \Delta)}
\]

(3.1.4)

\[
= h \left[ 1 + \frac{\Delta}{2} - \frac{\Delta^2}{12} + \ldots \right] f_j ,
\]

(3.1.5)

the result desired.

Next we would like to derive several other integration formulas. The integral from \( x_j \) to \( x_{j+2} \) can be derived from the result just obtained, for example, by breaking up the region of integration into two parts, as follows:

\[
\int_{x_j}^{x_{j+2}} dx f(x) = (1 + E) \int_{x_j}^{x_{j+1}} dx f(x) = 2h \left[ 1 + \Delta + \frac{\Delta^2}{6} - \frac{\Delta^4}{180} + \ldots \right] f_j ,
\]

(3.1.6)
3.1 NUMERICAL INTEGRATION

where the last equality has exploited our previous result (3.1.5) and Eq. (2.2.11).

We can also express these two results in terms of backward differences by merely expressing \( \nabla \) in terms of \( \Delta \). By Eqs. (2.2.11) and (2.2.12) we learn that

\[
\Delta = \frac{\nabla}{1 - \nabla}.
\]  

(3.1.7)

This relation may be substituted into Eq. (3.1.4) to find that

\[
\int_{x_j}^{x_{j+1}} dx f(x) = h \left[ 1 + \frac{\nabla}{2} + \frac{5}{12} \nabla^2 + \ldots \right] f_j.
\]  

(3.1.8)

Again, by use of this result and the shift operator \( E \) expressed in terms of \( \nabla \) by Eq. (2.2.12), we find

\[
\int_{x_j}^{x_{j+2}} dx f(x) = 2h \left[ 1 + \nabla + \frac{7}{6} \nabla^2 + \frac{4}{3} \nabla^3 + \ldots \right] f_j.
\]  

(3.1.9)

Additional formulas are readily obtained for other difference operators and for different intervals of integration.

The significant feature of the integration formulas is the truncation error associated with termination of the expansions. If we cut off the expansion to order \( \Delta^n \), then the truncation error is given by the order of the next term, i.e., \( h^{n+1} \). The coefficient \( h \) in these expansions then implies that the truncation error is of order \( h^{n+2} \). Recall that the truncating of the approximation to a derivative at \( n \)th order yielded a truncation error of order \( h^n \). We conclude that numerical integration formulas are more accurate than differentiation formulas of the same order, as indeed is usually true. The increased accuracy is easily understood geometrically when we realize that integration is a smoothing process, whereas differentiation tends to exaggerate fluctuations.

A formula of particular interest is readily obtained from Eq. (3.1.6) truncated to terms of order three. The resulting formula

\[
\int_{x_j}^{x_{j+2}} dx f(x) = 2h \left[ 1 + \Delta + \frac{\Delta^2}{6} \right] f_j = \frac{h}{3} [f_j + 4f_{j+1} + f_{j+2}] + O(h^5)
\]  

(3.1.10)

is the well-known Simpson's rule.

We shall use Simpson's rule for a sample problem. If we desire to integrate \( f(x) \) from \( a \) to \( b \) (Fig. 3.1.1), we divide the interval at \( J \) equally
spaced points such that \( x_1 = a, x_J = b \). We must choose \( J \) to be odd. The integration formula is then

\[
\int_a^b dx f(x) = \int_{x_1}^{x_2} dx f(x) + \int_{x_2}^{x_3} dx f(x) + \ldots + \int_{x_{J-1}}^{x_J} dx f(x), \tag{3.1.11}
\]

or

\[
\int_a^b dx f(x) = \frac{h}{3} [f_1 + 4f_2 + 2f_3 + \ldots + f_J] = \sum_j f(x_j)w_jh. \tag{3.1.12}
\]

In each of the 3-point intervals, the function \( f(x) \) is approximated as a parabola. It is readily apparent that the integration formula is indeed equivalent to a power series expansion of \( f(x) \). The same is true for higher order expansion. An illustration of the accuracy of the formula is left for the problems.

### 3.2 Ordinary Differential Equations

The results of the previous section are readily adapted to the solution of ordinary differential equations. Thus, the simple equation

\[
\frac{dy(x)}{dx} = f(x, y) \tag{3.2.1}
\]

is equivalent to the indefinite integral

\[
y(x) = y(a) + \int_a^x dx f(x, y). \tag{3.2.2}
\]

The numerical solution of Eq. (3.2.1) may then be written in terms of differences. In order to advance a solution, it is usually convenient to
use backward differences in order to use previously computed results. An approximation of the form

\[ y_{j+1} = y_j + hg(\nabla)f(x_j, y_j), \quad (3.2.3) \]

where \( g \) is some polynomial in \( \nabla \), is usually referred to as an Adams formula. We shall term such integration formulas explicit since all the quantities on the right-hand side are known by the time we try to evaluate the left-hand side. The truncation error in the integration formula is readily determined to be of \( O(h^{n+2}) \) where \( n \)th order differences are retained. In particular cases the order of the truncation error may be improved. Recall that the difference approximation \( (h/2)(y_{i+1} - y_{i-1}) \), was an approximation to the first derivative to \( O(h^2) \). Similar formulas are obtainable for integration. For instance,

\[
\begin{align*}
y_{j+1} &= y_j + \int_{x_j}^{x_{j+1}} dx f(x, y) = y_j + h \left[ 1 + \frac{\nabla}{2} + \frac{5}{12} \nabla^2 + \ldots \right] f_j \quad (3.2.4) \\
y_i &= y_{i-1} + \int_{x_{i-1}}^{x_i} dx f(x, y) = y_{i-1} + h \left[ 1 - \frac{\nabla}{2} - \frac{1}{12} \nabla^2 + \ldots \right] f_j \quad (3.2.5)
\end{align*}
\]

yield, together,

\[ y_{j+1} = y_{j-1} + h \left[ 2 + \frac{5}{12} \nabla^2 + \ldots \right] f_j. \quad (3.2.6) \]

Further formulas are readily obtainable for larger intervals. The truncation error, if we terminate before the \( \nabla^2 \) term, is of \( O(h^3) \). Either Eq. (3.2.4) or (3.2.5) is of \( O(h^2) \) if truncated after the first term. The formula (3.2.5) is an example of an implicit formula since the right-hand side is to be evaluated at the same point as the left-hand side. Generally implicit formulas have less truncation error than corresponding explicit formulas. In the above example, note the coefficient of \( \nabla^2 \) is smaller in the implicit formula. However, to use an implicit formula one must usually iterate to obtain the appropriate right-hand side.

Although the error term indicates a given truncation error, the actual error in computing may be increased since the values \( y_j \), used to compute \( f(x_j, y_j) \), are in error. Thus, the errors may propagate in a manner which reduces the order of approximation. As an extreme example, we reconsider the problem

\[ \frac{dy}{dx} = -\alpha y, \quad (3.2.7) \]

and use Eq. (3.2.4), terminated after first order differences. We have

\[ y_{j+1} = y_j - \alpha h \left[ \frac{3}{2} y_j - \frac{y_{j-1}}{2} \right]. \quad (3.2.8) \]
The solution of the difference equation yields the root pair
\[
\beta_1 = \frac{1 - \frac{3}{2} \alpha h}{2} + \frac{1}{2} \sqrt{1 - \alpha h + \frac{9}{4} \alpha^2 h^2} \quad (3.2.9a)
\]
\[
= 1 - \alpha h + \frac{\alpha^2 h^2}{2} + \frac{1}{4} \alpha^3 h^3 + ... \quad (3.2.9b)
\]
and
\[
\beta_2 = \frac{1 - \frac{3}{2} \alpha h}{2} - \frac{1}{2} \sqrt{1 - \alpha h + \frac{9}{4} \alpha^2 h^2} \quad (3.2.10a)
\]
\[
= -\frac{1}{2} \alpha h - \frac{1}{2} \alpha^2 h^2 - \frac{1}{2} \alpha^3 h^3 + ... \quad (3.2.10b)
\]
The solution is then
\[
y_j \approx c_0 \left(1 - \alpha h + \frac{\alpha^2 h^2}{2} + \frac{1}{4} \alpha^3 h^3\right)^j + c_1 (-)^j \left(\frac{1}{2} \alpha h + \frac{\alpha^2 h^2}{2}\right)^j, \quad (3.2.11)
\]
\[
y_j \approx c_0 \exp \left[-j \left(\alpha h - \frac{5}{12} \alpha^2 h^2\right)\right] + c_1 \left(-\frac{\alpha h}{2}\right)^j (1 + \alpha h)^j. \quad (3.2.12)
\]
Since the correction to the lowest order term in our result is \((-5/12) \alpha^2 h^2\) times that term, the solution is accurate to \(O(h^2)\). A more important result is to note that the second solution, the parasitic solution, decreases in magnitude and hence the calculation is stable, for sufficiently small \(h\). This result is in contrast to the earlier result (2.6.15). This result indicates a possible technique for combating instabilities in the numerical solution of differential equations. The two different difference equations, for the same problem, are
\[
y_{j+1} + 2 \alpha h y_j - y_{j-1} = 0 \quad \text{(unstable)} \quad (3.2.13)
\]
and
\[
y_{j+1} - \left(1 - \frac{3}{2} \alpha h\right) y_j - \frac{\alpha h}{2} y_{j-1} = 0 \quad \text{(stable)} \quad (3.2.14)
\]
Both approximations are accurate to \(O(h^2)\), but the second is stable. In many problems it is possible to adjust coefficients in a difference relation to obtain a stable solution. Sometimes the adjustment in coefficients may increase the truncation error. Nevertheless, the adjustment may make a solution possible. We shall encounter the problem of instabilities in partial differential equations also and shall consider further techniques for obtaining a stable solution.\(^1\)

In using higher order difference approximations to differential equations, a problem arises that does not occur in lowest order. The

\(^1\) A very illustrative study of stability may be found in Reference 3.
high order difference equation requires the values of the dependent variable to be known at many more points than a low order difference equation in order to find the value of the dependent variable at the next point in the sequence. The boundary conditions, which must be known in order that the solution be specific, supply the value of the dependent variables at only a few points. Accordingly, in order to carry out the integration, using either an explicit or an implicit method, several values of the dependent variable must be computed initially. Several methods are of frequent use, and we outline only one.

A sufficient number of initial values of the dependent variables are assumed in order to construct the necessary differences for forward integration.

The assumed values of the dependent variable can be related to values at other points by means of the integration formulas, such as Eqs. (3.1.5), (3.1.6), (3.2.4), or (3.2.5). With the known boundary values, and the other assumed values, we may compute the value of each dependent variable assumed earlier by means of the relevant difference relation mentioned above.

In general, the first guess for the values of the dependent variable will not agree with those computed; the procedure is iterated until the assumed values and the computed values agree. After a consistent set of initial values are found, the integration proceeds in a straightforward manner.

Certain numerical integration formulas have been developed which have the advantage of being self-starting. Among the most widely known methods are the Runge-Kutta methods. The basic idea behind the Runge-Kutta formulas is to develop an integration formula for Eq. (3.2.1) of the form

$$y_{n+1} = y_n + h \sum_i \alpha_i f(x_i, y_i), \quad (3.2.15)$$

where the $\alpha_i$, $x_i$, $y_i$ are chosen to make the integration formula agree with the Taylor series expansion of $y(x)$ to some order. The algebra involved in deriving the various formulas is quite involved (see Reference 2). We shall be content here to display an integration formula of 4th order:

$$y_{n+1} = y_n + (h/6)(k_1 + 2k_2 + 2k_3 + k_4), \quad (3.2.16)$$

where

$$k_1 = f(x_n, y_n),$$
$$k_2 = f(x_n + h/2, y_n + k_1/2),$$
$$k_3 = f(x_n + h/2, y_n + k_2/2),$$
$$k_4 = f(x_n + h, y_n + k_3).$$

The set of formulas is accurate to order $h^4$.  

3.2 ORDINARY DIFFERENTIAL EQUATIONS 93
In order to use the above-mentioned Runge-Kutta formula, one must evaluate the function \( f(x, y) \) four different times for each point of the solution. This may constitute a serious drawback to a problem; however, with high-speed computers even the evaluation of involved functions is relatively simple. Equations such as (3.2.16) involve only two points initially and hence are completely self-starting.

Higher order Runge-Kutta methods require additional evaluation of constants, \( (k_1, k_2, \text{etc.}) \), and ultimately the effort involved makes their use impractical. A more serious drawback with the methods are the difficulties in finding expressions for the errors. With the explicit and implicit methods, the evaluation of the higher differences gave some estimate of the errors in the integration. With the Runge-Kutta methods, one normally does not keep a running check on the error, and hence additional computations are necessary.

In the numerical solution of differential equations of order higher than first order, methods may be derived in an analogous manner. Alternatively, the higher order equation may be reduced to a set of simultaneous first order equations and the methods considered thus far used. The choice of approach depends upon the equation involved and any particular properties that may be exploited. We shall indicate the reduction to simultaneous first order equations and then consider the direct derivation of an approximate integration formula for a particular second order equation.

The \( n \)th order differential equation

\[
\frac{d^n y}{dx^n} = \frac{d^{n-1} y}{dx^{n-1}} + \cdots + g_0(x) = y(x) \quad (3.2.17)
\]

may be reduced by a simple change of variable. First we write (3.2.17) as

\[
\frac{d^n y}{dx^n} = y^n(x) = f(x, y, y', \ldots, y^{n-1}), \quad (3.2.18)
\]

where prime denotes differentiation with respect to \( x \). We define the variables \( y_0, y_1, \ldots, y_{n-1} \), by

\[
\begin{align*}
y_0(x) & = y, \\
y_1(x, y_0) & = y', \\
y_2(x, y_0, y_1) & = y'', \\
& \quad \vdots \\
y_{n-1}(x, y_0, y_1, \ldots, y_{n-2}) & = y^{(n-1)}.
\end{align*}
\]
We then obtain the set of \( n \) first order equations by differentiation of Eq. (3.2.19)

\[
\begin{align*}
y_0' &= y_1(x, y_0), \\
y_1' &= y_2(x, y_0, y_1), \\
y_2' &= y_3(x, y_0, y_1, y_2), \\
\vdots & \quad \vdots \\
y_{n-1}' &= f(x, y_0, y_1, y_2, \ldots, y_{n-1}).
\end{align*}
\tag{3.2.20}
\]

Each equation of the set (3.2.20) may now be integrated in order by the first order integration formulas.

Frequently one encounters second order equations in which the first derivative does not occur, i.e.,

\[
\frac{d^2y}{dx^2} + q(x)y = f(x). \tag{3.2.21}
\]

A second order equation of the form

\[
\frac{d^2u}{dx^2} + p(x) \frac{du}{dx} + g(x)u = s(x) \tag{3.2.22}
\]

can be transformed into the form (3.2.21) by the change of variable

\[
u = \exp \left( -\frac{1}{2} \int p(x) \, dx \right) y. \tag{3.2.23}
\]

The numerical integration of (3.2.21) can be accomplished by a double integration and by use of the expansions for functions of the difference operators. To this end, the double inverse of the differentiation operator \( d/dx \) is needed. Further, for our present purposes it will be convenient to express this operator in terms of the backwards difference operator, as in Eq. (2.3.14). Next we observe that, by repeated application of the technique used in obtaining Eq. (3.1.4),

\[
y_{j+1} - y_j - hy_j' = \int_{x_j}^{x_{j+1}} dx' \int_{x_j}^{x'} dx \frac{d^2y(x)}{dx^2} = \frac{d^2}{dx^2} \int_{x_j}^{x_{j+1}} dx' \int_{x_j}^{x'} dx y(x) = \left[ \frac{\nabla}{1 - \nabla} - h \frac{d}{dx} \right] y_j. \tag{3.2.24}
\]
From this last equality and Eq. (2.3.14), we find that
\[
\int_{x_j}^{x_{j+1}} dx'' \int_{x_j}^{x''} dx \, y(x) = h^2 \left( \frac{1}{2} + \frac{V}{6} + \frac{3}{24} V^2 + \ldots \right) y_j
\]
(3.2.25)
and that
\[
y_{j+1} = y_j + hy_j' + h^2 \left( \frac{1}{2} + \frac{V}{6} + \frac{3}{24} V^2 + \ldots \right) y_j''.
\]
(3.2.26)
Analogous implicit formulas and expressions using other difference operators are easily found.

Thus far we have considered only initial value problems. Frequently one encounters ordinary differential equations for which values of the dependent variable are given at the different boundaries of the domain of interest. It is possible to solve boundary value problems by the methods previously considered. Usually it is necessary to make an estimate of the starting slope and march to the far boundary. Any discrepancy between the computed and desired end condition must be eliminated by adjusting the starting slope. Pictorially the procedure is displayed in Fig. 3.2.1. We assume the function \( y(a) \) and \( y(b) \) known. The first trial is a trajectory computed by some integration rule. The second trial results from correcting the initial slope. Trial three is a better one, while trial four would represent the solution. Except under the most extraordinary conditions, the solution of a boundary value problem by the means just described is iterative.

An alternative approach to the problem is to consider replacing derivatives with appropriate difference equations and solving the resultant set of simultaneous equations. This procedure has some advantages over the "trajectory" method mentioned above. For instance, the problem of
extraneous solutions, which might contaminate a forward integration, can be controlled. Furthermore, for certain specific difference approximations, rapid methods for solving the simultaneous equations are possible. In later sections of this chapter, we shall discuss methods of solving boundary value problems in some detail.

### 3.3 Partial Differential Equations

The numerical solution of partial differential equations is usually considerably more difficult than the solution of ordinary differential equations. Different numerical procedures have evolved for different classes of partial differential equations. The most general second order, linear, partial differential equations can be written

\[
A \frac{\partial^2 f}{\partial x^2} + 2B \frac{\partial^2 f}{\partial x \partial y} + C \frac{\partial^2 f}{\partial y^2} = F(x, y, \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}). \tag{3.3.1}
\]

An equation of the form (3.3.1) is termed elliptic, parabolic, or hyperbolic according to the nature of the discriminant, where

\[
\Gamma = B^2 - AC. \tag{3.3.2}
\]

When \( \Gamma > 0 \) we call the equation hyperbolic, when \( \Gamma = 0 \) the equation is parabolic; and for \( \Gamma < 0 \) the equation is elliptic. If the coefficients \( A, B, \) and \( C \) depend upon position, then the nature of the equation may also depend upon position. It is possible, for instance, for an equation to be hyperbolic in some region and parabolic in another. If the equation is of one type, then the relations for \( \Gamma \) must hold everywhere. Classical examples of the three types of equations are:

1. the wave equation

\[
\frac{\partial^2 f}{\partial t^2} - \frac{\partial^2 f}{\partial x^2} = 0, \tag{3.3.3}
\]

which is hyperbolic;

2. the heat-flow equation

\[
\frac{\partial f}{\partial t} - \frac{\partial^2 f}{\partial x^2} = 0, \tag{3.3.4}
\]

which is parabolic;

3. Laplace's equation

\[
\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0, \tag{3.3.5}
\]

which is elliptic.
Equations of the hyperbolic and parabolic type are usually associated with initial value problems, whereas elliptic equations are associated with boundary value problems.

The types of boundary conditions for a problem are classified in a rather simple manner. If the value of a function along some boundary is given, we speak of the condition as being a Dirichlet boundary condition. In particular, if the function is zero all along the boundary, the condition is termed homogeneous Dirichlet, otherwise it is an inhomogeneous Dirichlet condition. If the derivative of the function is specified along the boundary, the condition is termed a Neumann condition. It is possible to have homogeneous or inhomogeneous Neumann boundary conditions. If the boundary conditions contain values of the function and derivative, we speak of mixed boundary conditions.

For all of the examples to be considered subsequently, we shall be concerned with two properties of the numerical solution. First, we shall want to know if the solution of the finite difference approximation is a reasonable representation of the analytic solution. In other words, if the relevant mesh spacings are made smaller and smaller, does the difference solution approach the differential solution. If the difference solution does approach the differential solution, we say the approximation converges, and the study of this property is termed convergence. The second property of interest is the behavior of any errors introduced into the calculation, for instance by round-off. An error may grow in an unbounded fashion and destroy a solution. Such a situation is called an instability. The general study of error behavior is called the stability problem.

An example of the convergence problem was given in Section 2.6. We learned there that the approximation to the heat-flow equation was convergent under a stringent condition on the spacing ratio. It may happen that the coefficients of the particular harmonics which violate the convergence criterion are zero for certain initial conditions. In this case the difference solution would converge in principle to the differential solution. However, if a round-off error introduced the nonconvergent harmonics, the solution may degenerate. In this latter case, we would say the problem is convergent but unstable. The requirement for stability is exactly the same as the requirement for convergence in this particular example.

It is not necessarily true that the convergence and stability requirements are the same for a given problem. It has been shown (see Reference 6) however, that for certain difference approximations to initial value problems with a wide variety of boundary conditions, the stability and
convergence requirements are the same. Proof of this important result is beyond the scope of this text. For our purpose we shall pay particular attention to the stability problem.

3.4 Hyperbolic Equations

3.4.1 The Wave Equation

For the study of hyperbolic equations, we consider the simple wave equation

\[
\frac{\partial^2 \phi(x, t)}{\partial t^2} = c^2 \frac{\partial^2 \phi(x, t)}{\partial x^2}, \quad a \leq x \leq b, \quad 0 \leq t
\]  

(3.4.1)

with initial conditions

\[
\phi(x, 0) = f_0(x), \quad \frac{\partial \phi(x, 0)}{\partial t} = g_0(x).
\]  

(3.4.2)

Perhaps the simplest difference approximation is obtained by using central differences in space and time. The equations become

\[
\frac{\phi_{j,k+1} - 2\phi_{j,k} + \phi_{j,k-1}}{h_t^2} = c^2 \frac{\phi_{j+1,k} - 2\phi_{j,k} + \phi_{j-1,k}}{h_x^2},
\]  

(3.4.3)

where \( j \) denotes the space index, \( k \) the time index, \( h_t \) and \( h_x \) are the time and space mesh spacings respectively, assumed constant. We denote the ratio \( c^2 h_t^2/h_x^2 = r^2 \), and factor Eq. (3.4.3) in the form

\[
\phi_{j,k+1} = r^2 \left[ \phi_{j+1,k} + \phi_{j-1,k} - 2 \left( 1 - \frac{1}{r^2} \right) \phi_{j,k} \right] - \phi_{j,k-1}.
\]  

(3.4.4)

Equation (3.4.4) is a 5-point difference relation and is shown schematically in Fig. 3.4.1.

We interpret the relation (3.4.4) as an algorithm to permit a march-out of the solution from \( k = 0 \) and \( k = 1 \) to later times. The procedure is explicit since all of the past values (smaller \( k \)) are known as we compute values along the time line \( k + 1 \).

The truncation error in the approximation is \( O(h_t^2) + O(h_x^2) \). To study the stability of the approximation, there are several possible approaches. We shall discuss one procedure here and in the next section consider a more general technique. We note first that for \( r^2 > 1 \), Eq. (3.4.4) takes on an interesting character, i.e., the sign of the term in \( \phi_{j,k} \) is positive.
We might expect that such an occurrence gives rise to some problems with the solution. To illustrate this fact, we consider the classical arguments presented by Courant et al. (see Reference 9).

The differential equation (3.4.1) is satisfied by any functions of the form

\[ \phi_1(x, t) = q(x - ct), \]
\[ \phi_2(x, t) = s(x + ct). \]

From the initial conditions we have

\[ q(x) + s(x) = f_0(x), \]
\[ -q'(x) + s'(x) = g_0(x). \]

Differentiating and subtracting, we have

\[ 2q'(x) = f'_0(x) - g_0(x), \]

or

\[ q(x) = \frac{1}{2} \left[ f_0(x) - \int_0^x du g_0(u) \right] + C_1. \]

Similarly,

\[ s(x) = \frac{1}{2} \left[ f_0(x) + \int_0^x du g_0(u) \right] + C_2. \]
Any linear combination of $\phi_1(x, t)$ and $\phi_2(x, t)$ is a solution and therefore

$$
\phi(x, t) = \frac{1}{2} \left[ f_0(x + ct) + f_0(x - ct) + \int_{x-ct}^{x+ct} du \, g_0(u) \right] + C_3. \tag{3.4.12}
$$

The condition at $t = 0$ requires that $C_3 = 0$.

Equation (3.4.12) affords an interesting interpretation of the stability requirements. The lines $x + ct = \text{constant}$ and $x - ct = \text{constant}$ are lines along which the function $f_0$ is constant. These lines are called characteristics of the differential equation. At a given point, say $x_0, t_0$, the characteristics are given by

$$
x - ct = x_1 \tag{3.4.13}
$$

$$
x + ct = x_2. \tag{3.4.14}
$$

The characteristics are sketched in Fig. 3.4.2. The characteristics extend to the $x$ axis at the points $x_1$ and $x_2$. The triangle with vertices at the points $x_1, x_2, x_0$ is called the region of determination of the solution at the point $x_0, t_0$. Notice that any initial conditions outside the interval $x_1$ to $x_2$ is not in the region of determination of $x_0, t_0$. We see that the solution at $x_0, t_0$ is not dependent upon the data outside the interval $x_1$ to $x_2$. The slope of the characteristic is $1/c$.

In order to study the stability of our difference approximation as a function of the ratio $r^2$, we now consider the region of determination on the network approximating the domain of interest as shown in Fig. 3.4.3.

When $r^2 = 1$ the slope of the line bounding the region of determination is $1/c$. Hence the boundary lines intersect the $x$ axis at the points $x_1$
and $x_2$. For $r^2 > 1$ the boundary lines have a slope greater than $1/c$ and define an interval within the interval $x_1$ to $x_2$ on the $x$ axis. The converse result is obtained for $r^2 < 1$.

For the case $r^2 \leq 1$, the difference solution is determined by as much (or more) of the initial data as that which determines the analytic solution. We should expect that such a solution would be a reasonable representation of the analytic solution. On the other hand, for $r^2 > 1$

![Figure 3.4.3](image-url)

**Fig. 3.4.3.** Regions of determination of the solution for various ratios of the spacing $r^2$.

the region of determination for the difference solution is smaller than that of the differential solution. This means that a portion of the data is not being used for the difference calculation that is necessary for the analytic solution. Consequently, we should expect that $r^2 > 1$ yields an unrealistic calculation. Indeed such is the case as we shall see.
Let us obtain the analytic solution of the difference equation (3.4.3). By the usual separation of variables, we have

\[ \phi_{j,k} = R_j T_k. \] (3.4.15)

Inserting in Eq. (3.4.3) and denoting the separation constant as \(- (c\alpha/h_x)^2\), we have the difference equations

\[ T_{k+1} - 2 \left(1 - \frac{\alpha^2 r^2}{2}\right) T_k + T_{k-1} = 0, \] (3.4.16)

and

\[ R_{j+1} - 2 \left(1 - \frac{\alpha^2}{2}\right) R_j + R_{j-1} = 0. \] (3.4.17)

To simplify matters, we shall assume \( \phi(a, t) = \phi(b, t) = 0 \). The spatial solution is then of the form

\[ R_j = A_n \sin \frac{n\pi j}{J}, \] (3.4.18)

with

\[ \frac{\alpha^2}{2} = \left(1 - \cos \frac{n\pi}{J}\right). \] (3.4.19)

Using this result in Eq. (3.4.16), we have

\[ T_{k+1} - 2 \left[1 - r^2 \left(1 - \cos \frac{n\pi}{J}\right)\right] T_k + T_{k-1} = 0. \] (3.4.20)

For \( r^2 \leq 1 \) Eq. (3.4.20) has trigonometric solutions which are similar to the trigonometric solutions of the differential equation. However, for \( r^2 > 1 \) some solutions of (3.4.20) will be exponential and would fail to represent the analytic solution. In order for the procedure to be stable (and convergent in this case), we must have

\[ r^2 \leq 1 \] (3.4.21)

or, equivalently,

\[ \frac{c^2 h_t^2}{h_x^2} \leq 1. \] (3.4.22)

The stability requirement (3.4.22) places an upper bound on the size of the time step for a given spatial mesh. In particular, if we decrease \( h_x \) (to reduce truncation error), we must also reduce the time increment. For particularly small meshes, the allowed maximum time step may be so small as to make the computation impractical.
In order to avoid the restriction, recourse is made to other difference approximations. For instance, the approximation

\[
\frac{\phi_{j,k+1} - 2\phi_{j,k} + \phi_{j,k-1}}{h_t^2} = c^2 \frac{\phi_{j+1,k+1} - 2\phi_{j,k+1} + \phi_{j-1,k+1}}{h_x^2}
\]

which has a truncation error \(O(h_t^2), O(h_x^t)\). The point pattern for the equation is shown in Fig. 3.4.4. Notice that we cannot solve for the point \(\phi_{j,k+1}\) explicitly in this case. In fact we must solve for the time line \(k + 1\) at all \(j\) simultaneously. A difference equation such as (3.4.23) is called implicit since we cannot solve for each point explicitly.

The advantage of the formulation based on the difference relation (3.4.23) is that the equation is unconditionally stable. In order to prove this fact, we introduce a particularly simple means for studying stability in the next section.

![Fig. 3.4.4. Five-point, implicit differencing pattern for the wave equation.](image)

3.4.2 The Von Neumann Method

To analyze the stability of a difference approximation, we must study the behavior of errors introduced into the computation. For a given difference equation, we may express the exact solution to the difference equation in the form

\[
\phi(\text{exact}) = \phi(\text{computed}) + \epsilon(\text{error}).
\]

The error may be due to round-off, a computational mistake, etc. The propagation of errors through the computation is obviously governed...
by the original difference equation itself with one notable distinction. Since we presume initial values and boundary values are known, the initial and boundary values for the error are obviously zero. That is, the equation governing the propagation of errors is the homogeneous form of the defining difference equation.

For difference relations involving constant coefficients, the errors may be expanded in a finite Fourier series. Thus for example, in the difference approximation to the wave equation as given in Eq. (3.4.23), the spatial dependence of the error can be written

$$\epsilon_{jk} = \sum_n B_{nk} \epsilon^{in\theta_j},$$

(3.4.25)

with $\theta_j = \pi x_j/L_x$. The time dependence may be included by presuming a coefficient of the form

$$B_{nk} = A_n \zeta^k(n).$$

(3.4.26)

The error at any point $j, k$ is expressed as

$$\epsilon_{jk} = \sum_n A_n \zeta^k(n) \epsilon^{in\theta_j}.$$  

(3.4.27)

This expression for the error was first used by Von Neumann (see Reference 10). The problem of stability is studied by noting the behavior of the coefficients $\zeta(n)$. If any $\zeta(n)$ is such that

$$|\zeta(n)| > 1$$

(3.4.28)

for any $n$, then we expect that the corresponding error harmonic would grow beyond limit for increasing $k$. To see this behavior we consider using the Von Neumann method for the difference approximation (3.4.3) and (3.4.23).

In the first case we have for the $n$th harmonic,

$$e^{in\theta_j}[\zeta^{k+1} - 2\zeta^k + \zeta^{k-1}] = r^2 \epsilon^{in\theta_j} \epsilon^{n\theta_j} + e^{in\theta_j} e^{i\theta_j}$$

or

$$(\zeta - 2 + \zeta^{-1}) = -4r^2 \sin^2 \frac{n\theta_j}{2}.$$  

(3.4.29)

(3.4.30)

Solutions of this equation obey the quadratic form

$$\zeta^2 - 2 \left(1 - 2r^2 \sin^2 \frac{n\theta_j}{2} \right) \zeta + 1 = 0.$$  

(3.4.31)
We have for the roots
\[ \zeta = \left[ 1 - 2r^2 \sin^2 \left( \frac{n\theta_1}{2} \right) \right] \pm \sqrt{\left[ 1 - 2r^2 \sin^2 \left( \frac{n\theta_1}{2} \right) \right]^2 - 1}. \tag{3.4.32} \]
For \( r^2 > 1 \) one of the roots is of absolute value greater than unity for large \( n \). Consequently, we expect amplification of errors at successive time steps. Conversely, for \( r^2 \leq 1 \), we see there is no amplification of errors. Note that for \( r^2 \leq 1 \), the roots occur as complex pairs of magnitude unity. In such a case we speak of a linear instability. Obviously if errors are not diminished in magnitude, errors at successive steps may accumulate. For such cases the amount of round-off error may ultimately become the dominant factor in the calculation.

For the implicit difference relation, the error equation for the \( n \)th harmonic is easily seen to be
\[ \left( 1 + 4r^2 \sin^2 \left( \frac{n\theta_1}{2} \right) \right) \zeta^2 - 2\zeta + 1 = 0. \tag{3.4.33} \]
For any real \( r^2 \) the roots of this equation are complex conjugates of maximum magnitude unity. Hence the implicit relation is unconditionally stable.

In general, the Von Neumann method is conservative in predictions about stability. The range of values for the harmonic index \( n \) is taken to be \( -\infty \) to \( +\infty \) in the usual application of the method. For a problem in the finite domain, the actual number of Fourier harmonics needed to describe the error is finite. By a very careful analysis using the finite series, it is possible to generate very accurate stability criteria.

Although the method is strictly applicable to difference equations with constant coefficients, it is sometimes used heuristically for equations with variable coefficients. In later sections we shall show further application of the method.

3.5 Parabolic Equations

3.5.1 Introduction

In the previous chapter, we considered the analytic solution to the heat-flow equation and derived a convergence criterion. In this section we shall show the stability condition for the given difference equation is the same as the convergence criterion. We shall also discuss some implicit approximations to the heat-flow equation. We postpone discussion of the age diffusion equation (which is parabolic) until the discussion of multigroup methods in Chapter IV.
The stability of the explicit difference equation

\[
\phi_{j,k+1} = \phi_{j,k} + \frac{h_t}{h_x^2} [\phi_{j+1,k} - 2\phi_{j,k} + \phi_{j-1,k}] \tag{3.5.1}
\]

or

\[
\phi_{j,k+1} = r^2(\phi_{j+1,k} + \phi_{j-1,k}) + (1 - 2r^2)\phi_{j,k} \tag{3.5.2}
\]

with \( r^2 = h_t/h_x^2 \) is easily examined. Notice first that Eq. (3.5.2) is a 4-point expression with the point pattern as shown in Fig. 3.5.1. A

![Four-point difference relation for the heat-flow equation.](image)

A formula such as Eq. (3.5.2) is sometimes called a two-level formula in contrast to the hyperbolic difference equations which were all three-level. The starting conditions for a two-level formula require initial data only along one time line.

For a two-level formula, the Von Neumann method is quite simple. By substituting the trial function

\[
\epsilon_{jk} = \zeta^k(n)e^{in\theta_j} \tag{3.5.3}
\]

into Eq. (3.5.2), we have

\[
\zeta = 2r^2 \cos n\theta_1 + (1 - 2r^2). \tag{3.5.4}
\]

The growth factor \( \zeta \) is bounded as

\[
1 - 4r^2 \leq \zeta \leq 1. \tag{3.5.5}
\]
In order to keep the magnitude of $\zeta \leq 1$, we must require $r^2 \leq \frac{1}{2}$, or equivalently

$$h_t \leq \frac{h_x^2}{2}.$$  \hspace{1cm} (3.5.6)

Equation (3.5.6) is precisely the same condition found earlier.

A wide variety of implicit approximations to the heat-flow equation have been studied. As usual, implicit relations are considered in order to overcome the restrictive condition imposed for stability of the simple explicit difference equation. Consider the following simple two-level approximation.

$$\frac{\phi_{j,k+1} - \phi_{jk}}{h_t} = \alpha \frac{\delta_x^2 \phi_{j,k+1}}{h_x^2} + (1 - \alpha) \frac{\delta_x^2 \phi_{jk}}{h_x^2}, \quad 0 \leq \alpha \leq 1.$$  \hspace{1cm} (3.5.7)

Notice that the second difference is applied along the time lines $k$ and $k + 1$. The truncation error for the relation (3.5.7) is easily seen to be $O(h_t) + O(h_x^2)$. Formula (3.5.7) is a six-point formula and has the point pattern shown in Fig. 3.5.2.

![Six-point difference relation for an implicit approximation to the heat-flow equation.](image)

**FIG. 3.5.2.** Six-point difference relation for an implicit approximation to the heat-flow equation.

The stability of Eq. (3.5.7) is studied by the Von Neumann method. By the usual process we have

$$\zeta - 1 = -4r^2 \alpha \zeta \sin^2 \frac{n\theta_1}{2} - 4(1 - \alpha)r^2 \sin^2 \frac{n\theta_1}{2}.$$  \hspace{1cm} (3.5.8)
Considering the extremum of the sine function, we find Eq. (3.5.8) factors into
\[ r^2 \leq \frac{1}{2(1 - 2\alpha)}, \quad 0 \leq \alpha < \frac{1}{2}, \]  
(3.5.9)
and \( r^2 \) unrestricted for \( \frac{1}{2} \leq \alpha \leq 1 \). Note that in the limit \( \alpha = 1 \), the difference equation becomes a simple four-point implicit formula.

### 3.5.2 Alternating-Direction Implicit Method

A very important method, first derived by Peaceman and Rachford (see Reference 11), will be considered in this section. The method is useful for solving two-dimensional parabolic equations (and also elliptic equations). We consider the simple heat-flow equation
\[ \frac{\partial \phi}{\partial t} = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2}, \quad 0 \leq x \leq a, \quad 0 \leq y \leq b, \quad t \geq 0, \]  
(3.5.10)
with boundary conditions
\[ \phi(0, y, t) = \phi(a, y, t) = 0, \]
\[ \phi(x, 0, t) = \phi(x, b, t) = 0, \]
\[ \phi(x, y, 0) = f(x, y). \]

The simple explicit difference equation is
\[ \frac{\phi_{j,k,n+1} - \phi_{j,k,n}}{h_t} = \frac{\delta^2 \phi_{j,k,n}}{h_x^2} + \frac{\delta^2 \phi_{j,k,n}}{h_y^2}. \]  
(3.5.12)

It is easily seen that the stability criterion for Eq. (3.5.12) is
\[ \frac{h_t}{h_x^2} + \frac{h_t}{h_y^2} \leq \frac{1}{2}. \]  
(3.5.13)

To derive a less restrictive condition, we consider the implicit equation
\[ \frac{\Delta \phi_{j,k,n}}{h_t} = \frac{\delta^2 \phi_{j,k,n+1}}{h_x^2} + \frac{\delta^2 \phi_{j,k,n+1}}{h_y^2}. \]  
(3.5.14)

It is easy to show that Eq. (3.5.14) is unconditionally stable; however, the solution requires the simultaneous solution of values along the entire plane of time \( n + 1 \). The basis of the Peaceman-Rachford method is to
make the equations implicit along only one line at a time. Thus, we use the two difference equations
\[
\frac{\Delta \phi_{j,k,n}}{h_t} = \frac{\delta_z^2 \phi_{j,k,n+1}}{h_x^2} + \frac{\delta_y^2 \phi_{j,k,n}}{h_y^2} \tag{3.5.15}
\]
and
\[
\frac{\Delta \phi_{j,k,n+1}}{h_t} = \frac{\delta_z^2 \phi_{j,k,n+1}}{h_x^2} + \frac{\delta_y^2 \phi_{j,k,n+2}}{h_y^2} \tag{3.5.16}
\]
For simplicity we take \( h_x = h_y \) and \( h_t/h_x^2 = r^2 \). Written in component form, the equations become
\[
\phi_{j+1,k,n+1} - (2 + 1/r^2)\phi_{j,k,n+1} + \phi_{j-1,k,n+1} = -\phi_{j-1,k,n} + (2 - 1/r^2)\phi_{j,k,n} - \phi_{j,k-1,n}, \tag{3.5.17}
\]
\[
\phi_{j,k+1,n+2} - (2 + 1/r^2)\phi_{j,k,n+2} + \phi_{j,k-1,n+2} = -\phi_{j+1,k,n+1} + (2 - 1/r^2)\phi_{j,k,n+1} - \phi_{j-1,k,n+1}. \tag{3.5.18}
\]
The two equations are used successively and hence the name alternating direction. The stability is analyzed in the usual manner. We assume an error function of the form
\[
\epsilon_{j,k,n} = \zeta(m,l)e^{im\theta}e^{i\psi k}. \tag{3.5.19}
\]
From Eq. (3.5.17) we have
\[
\zeta(m,l) \left[ \frac{1}{r^2} + 4 \sin^2 \frac{m\theta}{2} \right] = \left[ \frac{1}{r^2} - 4 \sin^2 \frac{m\theta_1}{2} \right] \tag{3.5.20}
\]
or
\[
\zeta(m,l) = \frac{\frac{1}{r^2} - 4 \sin^2 \frac{m\theta_1}{2}}{\frac{1}{r^2} + 4 \sin^2 \frac{m\theta_1}{2}} \tag{3.5.21}
\]
For some values of \( m, l, \) and \( r^2 \), the growth factor might be considerably greater than unity. Thus applying the implicit equation in only one direction is unstable in general. On the other hand, using the alternating equations we have
\[
\zeta^2(m,l) = \frac{\frac{1}{r^2} - 4 \sin^2 \frac{m\theta_1}{2}}{\frac{1}{r^2} + 4 \sin^2 \frac{m\theta_1}{2}} \cdot \frac{\frac{1}{r^2} - 4 \sin^2 \frac{m\theta_1}{2}}{\frac{1}{r^2} + 4 \sin^2 \frac{m\theta_1}{2}} \tag{3.5.22}
\]
For two successive steps the resulting growth is always bounded at unity. Consequently the alternating-direction method is unconditionally stable.

The second advantage of the method stems from the fact that the equations are implicit along one line at a time. The resulting equations are three-point relations. The matrix form of the equations along one line is tridiagonal and hence can be solved by the method of matrix factorization. The result is a very fast stable method for solving two-dimensional parabolic equations. In the discussion of iterative procedures, we shall return to the alternating-direction method and show its application to elliptic equations.

3.6 Elliptic Equations and Iterative Methods

3.6.1 Introduction

The treatment of initial-value problems involved the formation of a difference operator that permitted the initial conditions to be extended into the domain of interest of the problem. Such a procedure is generally not useful for elliptic equations where the boundary conditions are given over the entire region of interest. The numerical solution of elliptic equations is usually accomplished by solution of simultaneous equations with a variety of methods.

One possible means of solving a set of simultaneous equations is by the Gauss reduction scheme. Unfortunately the reduction process for \( N \) equations in \( N \) unknowns requires approximately \( N^3 \) operations. Furthermore, a certain amount of round-off in each operation may cause the solution to degenerate for large \( N \). On the other hand, a direct reduction procedure is determinate in that a fixed number of steps are needed (in theory) to find the solution.

An alternative approach to the solution of elliptic equations is an iterative procedure. In general, iterative methods require an infinite number of steps to solve a problem exactly. However, for practical purposes it is usually possible to terminate an iteration after a finite number of steps which are fewer in number than those required for reduction methods. Furthermore, iterative procedures have certain advantages with respect to round-off over direct reduction. To make these matters clearer, we consider a simple introductory example.

We desire the solution of the problem

\[
\frac{d^2y}{dx^2} = 0 \tag{3.6.1}
\]
III. NUMERICAL SOLUTIONS OF EQUATIONS

with boundary conditions

\[ y(0) = y_0; \quad y(a) = y_a. \]

We divide the interval 0 to \( a \) into \( K \) subintervals of equal width \( h \) as in Fig. 3.6.1. For simplicity we replace the second derivative with a second central difference. We have then

\[ y_{k-1} - 2y_k + y_{k+1} = 0, \quad 1 \leq k \leq K - 1. \]  

(3.6.2)

The set of equations (3.6.2) could be solved by reduction (in this case by matrix factorization, see Section 1.14) in a straightforward manner.

The iterative solution of the equations is achieved by first assuming a trial solution at all the interior points of the mesh. We then write Eq. (3.6.2) in the form

\[ y_k = \frac{y_{k+1} + y_{k-1}}{2}. \]

(3.6.3)

Equation (3.6.3) provides an algorithm for computing values of \( y_k \) in terms of its nearest neighbors. Let us denote the values of \( y_k \) for the initial trial with a superscript 0. To compute the "new" values of \( y_k \) from the "old," we might consider the rule

\[ y_k^1 = \frac{y_{k+1}^0 + y_{k-1}^0}{2} \]  

(3.6.4)

and in general

\[ y_k^p = \frac{y_{k+1}^{p-1} + y_{k-1}^{p-1}}{2}. \]  

(3.6.5)

The following important questions now occur:

1. Will the iteration rule (3.6.5) ever yield a solution?
2. How will we know when we have achieved a solution?
3. How long will it take?

We assume that the difference approximation itself has been adequate for the problem.

To help answer these questions, let us denote the actual solution to the difference equation as $y_k^\infty$. We define the initial error, say $e_k^0$, as

$$e_k^0 = y_k^0 - y_k^\infty$$  \hfill (3.6.6)

and after $p$ steps

$$e_k^p = y_k^p - y_k^\infty.$$  \hfill (3.6.7)

The answer to the first question can be given in terms of the errors. Thus, if the sequence $e_k^0$, $e_k^1$, ..., $e_k^p$ approaches zero, for all $k$, then the iteration does yield a solution. To answer the second question is somewhat more difficult in that we never know the errors (else we would also know the solution). The usual test for determining when the solution has been achieved is based upon the following criterion. We define a function, say $r_k^p$, as

$$r_k^p = y_k^{p+1} - y_k^p.$$  \hfill (3.6.8)

In terms of the errors we have

$$r_k^p = e_k^{p+1} - e_k^p.$$  \hfill (3.6.9)

If $r_k^p$ is small, for all $k$, then one assumes the approximate solution has been found. Obviously this criterion is not always valid since a small difference of errors does not necessarily imply small errors. A practical way to be reasonably sure of convergence is to choose a sufficiently small criterion and to iterate a few times perhaps beyond the criterion. This second feature guards against the rate of convergence as a function of the iteration index having a minimum. For example, the convergence rate may initially increase only to later decrease. A minimum might result while still far from converged.

The question of how long an iteration will take is very important, and many different iteration methods have been devised to hasten the rate at which a solution is obtained. We shall introduce a variety of methods in later sections, and one of the criteria for judging the merit of a method will be the rate of convergence.
3.6.2 Stability of Iterations

In this section we shall attempt to provide a unified basis for the study of iterative methods. We assume the set of simultaneous equations are written in the form

\[ Ax = y, \]  

where we further assume a solution does exist but otherwise do not restrict the matrix \( A \). The exact solution to the problem, say \( x^\infty \), is

\[ x^\infty = A^{-1}y. \]  

We begin the iteration by considering a trial solution \( x^0 \) and then operate on \( x^0 \) to produce a new trial \( x^1 \). We assume the iteration can be written

\[ x^1 = Bx^0 + z, \]  

where the matrix \( B \) and the vector \( z \) are taken independent of the iteration index. Such an iteration is called stationary. The iterative procedure can be extended to successive trials in the form

\[ x^{n+1} = Bx^n + z. \]  

The matrix \( B \) is called the iteration matrix, and the ability of achieving a solution and the corresponding rate of convergence are intimately related to the properties of the iteration matrix.

The properties we desire of our iteration are that the sequence of vectors \( x^n \) approach \( x^\infty \), and further, that iteration with \( x^\infty \) reproduces itself. In mathematical terms the last requirement is

\[ x^\infty = Bx^\infty + z. \]  

Assuming Eq. (3.6.14) is valid, and defining the error vector \( e^n \) as

\[ e^n = x^n - x^\infty, \]  

we have from Eq. (3.6.13)

\[ e^n = Be^{n-1}. \]  

Equation (3.6.16) states that the error vector obeys the homogeneous form of the iteration equation. Now, if the sequence of vectors \( x^n \) is to approach \( x^\infty \), then the sequence of vectors \( e^n \) must approach zero. From Eq. (3.6.16) we have

\[ e^n = Be^{n-1} = B^n e^0. \]
We require
\[
\lim_{p \to \infty} \epsilon^p = \lim_{p \to \infty} B^p \epsilon^0 \to 0. \tag{3.6.18}
\]

Let us now assume the matrix $B$ has a complete set of eigenvectors, say $e_i$, and corresponding eigenvalues $\lambda_i$. Since the eigenvectors are complete, we may expand $\epsilon^0$ in the form
\[
\epsilon^0 = \sum_i \alpha_i e_i. \tag{3.6.19}
\]

Operation on $\epsilon^0$ by $B$ yields $\epsilon^1$ as
\[
\epsilon^1 = B \epsilon^0 = \sum_i \alpha_i \lambda_i e_i. \tag{3.6.20}
\]

By induction we have
\[
\epsilon^p = \sum_i \alpha_i \lambda_i^p e_i. \tag{3.6.21}
\]

In order for the error vector to vanish, we must require
\[
|\lambda_i| < 1, \quad \text{all } i. \tag{3.6.22}
\]

The above result is very important for our future considerations and is also quite general. We shall refer to Eq. (3.6.22) as the stability condition for iterative methods. If an eigenvalue, say $\lambda_r$, was of magnitude greater than unity, the iteration would diverge, and in our terminology is unstable. Even if the initial error vector were orthogonal to the eigenvector, say $e_r$, of the eigenvalue $\lambda_r$, round-off in computations would introduce components along $e_r$ and ultimately ruin the iteration.

The stability condition also provides a means of computing the convergence rate of an iteration. If the eigenvalues are ordered such that
\[
1 > |\lambda_1| > |\lambda_2| > \ldots > |\lambda_K|,
\]
then for sufficiently large $p$, the error is approximately
\[
\epsilon^p \approx \alpha_1 \lambda_1^p e_1 = \alpha_1 e^{p \ln \lambda_1} e_1. \tag{3.6.23}
\]

The term $e^{p \ln \lambda_1}$ is the decay factor for the errors. In particular we define
\[
-\ln \lambda_1 = \nu, \tag{3.6.24}
\]

\(^2\) The requirement of completeness of the eigenvectors of $B$ to derive condition (3.6.22) is over restrictive. By consideration of the Jordan canonical form, the same result can be achieved by any real matrix. See problem 7.
as the convergence rate of the iteration. For small $\lambda_1$ the convergence rate is large, meaning a rapid reduction of the error with increasing number of iterations. The worth of an iteration scheme is partly measured in terms of the factor $\nu$. We shall consider detailed examples later. We note, however, that in general $\nu$ decreases for increasing number of unknowns for most iterations. That is, the larger the problem, the longer it takes to solve. In fact, for many methods convergence is reciprocally related to $N^2$, where $N$ is the number of unknowns. That the relation goes as $N^2$ can be heuristically seen by the following argument. Since the eigenvalues must be distributed between $-1$ and $+1$, an increase in $N$ should move the ones of largest magnitude toward the end points of the interval. Further, since there are more points to be solved, the number of operations grows with $N$. Thus we expect the convergence to be related somehow to $N^2$.

The above argument is strictly heuristic and not necessarily true. Later examples will be considered to illustrate the general behavior of the convergence factor.

Thus far our results have been derived by strictly algebraic considerations. It is interesting and instructive to consider a geometric interpretation of iterations. To this end, we introduce the residual vector $r^p$ defined as

$$r^p = x^{p+1} - x^p. \quad (3.6.25)$$

In terms of the error vector, Eq. (3.6.25) becomes

$$r^p = (B - I)e^p. \quad (3.6.26)$$

Since the residual vector represents the error vector in a transformed space, the convergence criteria for the residuals are the same as for the errors. Furthermore, the asymptotic behavior is the same.

The residuals are calculable at any stage of the iteration. From the defining equation for the iteration (3.6.13), we have

$$r^p = (B - I)x^p + z. \quad (3.6.27)$$

We interpret Eq. (3.6.25) in the form

$$x^{p+1} = x^p + r^p \quad (3.6.28)$$

as stating that the vector $x^p$ is corrected by addition of a vector $r^p$ which in turn is defined by an algorithm (3.6.27). Different iteration methods consist of different algorithms for computing the correction vector. Notice that if $r^p$ is chosen to change only one component of $x^p$, the vector $r^{p+1}$ may still have all of its components changed since
$x^{p+1}$ is transformed by an operator that is not diagonal. In any event, convergence of the solution requires that $x^p$ approach a limit vector and $r^p$ approach the null vector.

The change of the trial vector $x^p$ by the correction vector $r^p$ is called an iterate or one iteration. The change of each component separately is called a relaxation or displacement. Thus, after relaxing each component of $x^p$, once and only once, we complete one iteration.

### 3.6.3 Stationary Iterations

In this section we shall discuss several very common iteration methods and consider their stability properties. For each of the iteration methods introduced, we shall apply the method to the simple Laplace equation in the square.\(^3\) That is, we consider the equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0, \quad 0 \leq x \leq a, \quad 0 \leq y \leq a \quad (3.6.29)$$

with boundary conditions

$$\begin{align*}
\phi(0, y) &= \phi(a, y) = 0, \\
\phi(x, 0) &= 0, \\
\phi(x, a) &= f(x).
\end{align*} \quad (3.6.30)$$

For all iteration methods we use the difference equation

$$\frac{\delta_x^2 \phi}{h_x^2} + \frac{\delta_y^2 \phi}{h_y^2} = 0, \quad h_x = h_y = h \quad (3.6.31)$$

#### A. Method of Simultaneous Displacements

Let us assume we are solving the matrix equation

$$Ax = y, \quad (3.6.32)$$

where we assume a solution does exist. The matrix $A$ is written in the form

$$A = L + D + U, \quad (3.6.33)$$

where $L$ is strictly lower triangular, $U$ strictly upper triangular, and $D$ diagonal. We assume $D \neq 0$. We now define the iteration

$$(L + U)x^p + Dx^{p+1} = y, \quad (3.6.34)$$

\(^3\) The numerical solution of the two-dimensional Laplace equation is perhaps the most thoroughly studied problem in iterative analysis. The problem is sometimes referred to as the model problem.
or
\[ x^{n+1} = -D^{-1}(L + U)x^n + D^{-1}y. \]  
(3.6.35)

The method of simultaneous displacements then consists in solving the difference equation by means of the iteration in Eq. (3.6.35).

We next observe that the method of iteration in consistent with a vector being a solution of our original Eq. (3.6.32) for if at some stage the correct answer were found, then, apart from round-off errors, iterating the solution does nothing more than give it back to us again. In particular, if at some stage \( x^n \) were the correct solution
\[ x^n = A^{-1}y, \]  
(3.6.36)
then by substituting this solution into the iteration scheme (3.6.35), we find that
\[ x^{n+1} = A^{-1}y, \]
and we recover the original solution.

The important question of whether or not an arbitrary starting vector, say \( x^0 \), will approach \( A^{-1}y \) depends upon the iteration matrix \(-D^{-1}(L + U)\). The eigenvalue spectrum of the iteration matrix must be such that \( |\lambda_i| < 1 \) for all \( i \), where the \( \lambda_i \) are eigenvalues of the matrix. Thus, the roots of the equation
\[ | -D^{-1}(L + U) - \lambda I | = 0 \]  
(3.6.37)
or, equivalently
\[ | L + \lambda D + U | = 0, \]  
(3.6.38)
must have magnitude less than unity.

In general the solution of (3.6.38) is quite difficult. However, for many cases of interest, it is possible to determine the nature of the eigenvalue spectrum without solving the secular equation. We assume that the problem is well posed in the sense that \( A \) is irreducible, otherwise we factor \( A \) into two separate problems. The iteration matrix can be written
\[
D^{-1}(L + U) = \begin{bmatrix}
0 & a_{12} & a_{13} & \cdots & a_{1n} \\
a_{21} & 0 & a_{23} & \cdots & a_{2n} \\
a_{31} & a_{32} & 0 & \cdots & a_{3n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nn} & 0 & 0
\end{bmatrix}. \]  
(3.6.39)
By Gerschgorin's theorem, the spectral radius of $D^{-1}(L + U)$ is bounded by

$$|\lambda_{\text{max}}| \leq \max_{i} \sum_{j} \left| \frac{a_{ij}}{a_{ii}} \right|. \quad (3.6.40)$$

Gerschgorin's theorem may be strengthened if the matrix (3.6.39) is irreducible. If the maximum value of the sum on the right-hand side of Eq. (3.6.40) is denoted $\rho$, then Gerschgorin's theorem states

$$|\lambda_{\text{max}}| \leq \rho. \quad (3.6.41)$$

If, for any $i$, and for an irreducible matrix

$$\sum_{j} \left| \frac{a_{ij}}{a_{ii}} \right| < \rho, \quad (3.6.42)$$

then

$$|\lambda_{\text{max}}| < \rho, \quad (3.6.43)$$

i.e., there is strict inequality. As a consequence, we see that if the matrix $A$ is such that

$$|a_{ii}| \geq \sum_{j} |a_{ij}| \quad (j \neq i) \quad (3.6.44)$$

with inequality for some $i$, then the method of simultaneous displacements converges.

If the diagonal elements of $A$ are such that (3.6.42) is true, then we say $A$ has diagonal dominance. Consequently, for an irreducible matrix with diagonal dominance, the method of simultaneous displacements converges. Fortunately most elliptic difference equations of reactor interest have this property. Note that the condition is sufficient for convergence but not necessary. For instance, the matrix

$$A = \begin{bmatrix} -1 & 2 \\ 2 & -5 \end{bmatrix} \quad (3.6.45)$$

does not have diagonal dominance, yet the iteration matrix

$$-D^{-1}(L + U) = \begin{bmatrix} 0 & 2 \\ 2/5 & 0 \end{bmatrix} \quad (3.6.46)$$

has eigenvalues with magnitude less than unity.

\* For proof see Reference 8, Chapter 1.
III. NUMERICAL SOLUTIONS OF EQUATIONS

We now consider application of the method of simultaneous displacements to the model problem. The equation for the error function is the homogeneous form of the equation with homogeneous boundary conditions. The iteration algorithm for the error $\epsilon_{j,k}^p$ is, from Eq. (3.6.34)

$$
\epsilon_{j+1,k}^p + \epsilon_{j-1,k}^p - 2\epsilon_{j,k}^p + \epsilon_{j,k+1}^p + \epsilon_{j,k-1}^p - 2\epsilon_{j,k}^{p+1} = 0, \quad (3.6.47)
$$
or

$$
\epsilon_{j,k}^{p+1} = \frac{1}{4}[\epsilon_{j,k+1}^p + \epsilon_{j,k-1}^p + \epsilon_{j+1,k}^p + \epsilon_{j-1,k}^p]. \quad (3.6.48)
$$

The stability of the iteration can be studied by examining the behavior of the various eigenfunctions of the iteration operator. From past results we know the error (with homogeneous boundary conditions) must be of the form

$$
\epsilon_{j,k}^{(m,n)} = A(m, n) \sin \frac{m\pi j}{K} \sin \frac{n\pi k}{K}, \quad 1 \leq j, k \leq K - 1 \quad (3.6.49)
$$

where $(m, n)$ are the indices of the eigenfunction. If we assume $\epsilon_{j,k}^{(m,n)}$ is of the form (3.6.49) and insert in (3.6.48), we have

$$
\epsilon_{j,k}^{p+1} = \frac{A(m, n)}{4} \left[ \sin \frac{m\pi j}{K} \left( \sin \frac{m\pi (k + 1)}{K} + \sin \frac{m\pi (k - 1)}{K} \right) \\
+ \sin \frac{n\pi k}{K} \left( \sin \frac{n\pi (j + 1)}{K} + \sin \frac{n\pi (j - 1)}{K} \right) \right]. \quad (3.6.50)
$$

After an elementary reduction we have

$$
\epsilon_{j,k}^{p+1} = A(m, n) \sin \frac{m\pi j}{K} \sin \frac{n\pi k}{K} \left[ \cos \frac{m\pi}{K} + \cos \frac{n\pi}{K} \right] \quad (3.6.51)
$$

$$
= \zeta \epsilon_{j,k}^p. \quad (3.6.52)
$$

The scale factor $\zeta$ determines the rate of growth or decay for the errors. For this case the errors all decay with differing rates for the different harmonics. The largest value of $\zeta$ occurs for $m = n = 1$ or $m = n = K - 1$. For such a case we have

$$
\zeta = \cos \frac{\pi}{K} \approx 1 - \frac{\pi^2}{2K^2}. \quad (3.6.53)
$$

The iteration converges with the convergence rate

$$
v = \frac{\pi^2}{2K^2}. \quad (3.6.54)
$$

The argument below follows that of Frankel, Reference 12.
Notice that the convergence rate is proportional to $K^2$, the number of unknowns. For a mesh of 20 by 20 points, the convergence rate is approximately

$$\nu = 0.01234$$

and hence about 80 iterations are required to reduce the error by a factor of $e$.

Notice that the method of simultaneous displacements requires the vectors $\mathbf{x}^{p+1}$ and $\mathbf{x}'$ in the computation simultaneously. This means that for a vector of $N$ dimensions, $2N$ numbers must be retained. For large problems this condition may be important in working with digital computers.

The method of simultaneous displacements is easily described in terms of the residuals. By Eqs. (3.6.35) and (3.6.28) we have

$$\mathbf{r}^p = -D^{-1}[\mathbf{A}\mathbf{x}^p - \mathbf{y}]. \quad (3.6.55)$$

The quantity $\mathbf{A}\mathbf{x}^p - \mathbf{y}$ is a measure of the error in a given iteration. Some authors call $\mathbf{A}\mathbf{x}^p - \mathbf{y}$ the residual.

As a final note we remark that the method of simultaneous displacements is also called the Jacobi method or the Richardson method.

**B. Method of Successive Displacements**

The method of successive displacements (also called the Gauss-Seidel method or the Liebmann method) is quite similar to the previous scheme. We again consider the problem of solving the matrix equation (3.6.32). In this case we assume an iteration of the form

$$(\mathbf{L} + \mathbf{D})\mathbf{x}^{p+1} + \mathbf{U}\mathbf{x}^p = \mathbf{y} \quad (3.6.56)$$

or

$$\mathbf{x}^{p+1} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}\mathbf{x}^p + (\mathbf{L} + \mathbf{D})^{-1}\mathbf{y}, \quad (3.6.57)$$

which is the iteration algorithm that defines the method of successive displacements.

As before, it is easily seen that if at some state $\mathbf{x}^p = \mathbf{A}^{-1}\mathbf{y}$, then $\mathbf{x}^{p+1} = \mathbf{A}^{-1}\mathbf{y}$.

We now consider the convergence properties of the iteration and derive expressions for the eigenvalues of the iteration operator. Assume $|\mathbf{D}| \neq 0$, and since $\mathbf{L}$ is strictly lower triangular, the matrix $(\mathbf{L} + \mathbf{D})^{-1}$ exists and is of the form

$$(\mathbf{L} + \mathbf{D})^{-1} = \mathbf{L}' + \mathbf{D}^{-1}, \quad (3.6.58)$$
where $L'$ is strictly lower triangular. Convergence of the iteration depends upon the eigenvalues of the iteration matrix, i.e., the roots $\lambda$ of

$$|-(L + D)^{-1}U - \lambda I| = 0 \quad (3.6.59)$$

or

$$|\lambda L + \lambda D + U| = 0. \quad (3.6.60)$$

The roots of Eq. (3.6.60) must all have magnitude less than unity for convergence. It is usually difficult to solve for the roots explicitly; however, the rule concerning diagonal dominance of irreducible matrices applies for the method of successive displacement also.

In the method of successive displacements, we are always using the latest computed values for the unknown and hence the name successive displacement. It is interesting to compare the convergence rates of the method of simultaneous and successive displacements for the same problem. When the iteration matrices are non-negative, we shall prove shortly that the two methods converge or diverge together. Further, if they converge, then the successive displacements technique converges more rapidly.

To prove this result we consider first the method of simultaneous displacement. We assume that the iteration matrix is nonnegative which is possible if the matrices $L$ and $U$ have nonnegative elements and $D$ has all negative nonzero components, or conversely, for instance. For the method of simultaneous displacements, the iteration matrix is then

$$-(D^{-1}(L + U)) = R + T, \quad (3.6.61)$$

where $R$ is strictly lower triangular and $T$ strictly upper triangular. Similarly, the iteration matrix for the method of successive displacements is

$$-(L + D)^{-1}U = (I - R)^{-1}T. \quad (3.6.62)$$

The proof (from Reference 13) of divergence is straightforward. Let $\lambda$ be the positive eigenvalue of $R + T$ of greatest magnitude, and similarly $\sigma$ be the largest eigenvalue of $(I - R)^{-1}T$. Let $z$ be the eigenvector corresponding to $\lambda$; that is,

$$(R + T)z = \lambda z. \quad (3.6.63)$$

Therefore

$$\left(I - \frac{R}{\lambda}\right)^{-1}(R + T)z = \lambda \left(I - \frac{R}{\lambda}\right)^{-1}z. \quad (3.6.64)$$
Now the matrix \((I - \frac{R}{\lambda})^{-1}\) can be written
\[
(I - \frac{R}{\lambda})^{-1} = I + \frac{R}{\lambda} + \left(\frac{R}{\lambda}\right)^2 + \ldots + \left(\frac{R}{\lambda}\right)^m,
\] (3.6.65)
where \(R\) is assumed \((m + 1)\) by \((m + 1)\). Thus terms beyond the \(m\)th power of \(R\) vanish, since \(R\) is strictly lower triangular. Note that all the elements of the sum are nonnegative in view of the hypothesis. Using the expansion (3.6.65) in Eq. (3.6.64), we have
\[
(I - \frac{R}{\lambda})^{-1}Tz = \lambda z.\]
(3.6.66)
Thus \(\lambda\) is an eigenvalue of \((I - \frac{R}{\lambda})^{-1}T\). From the properties of nonnegative matrices, if \(\lambda > 1\), then
\[
(I - \frac{R}{\lambda})^{-1}T < (I - R)^{-1}T.
\] (3.6.67)
Consequently \(\sigma > \lambda > 1\), and the two iterations diverge together. If \(\lambda = 1\), then \(\sigma = 1\), while for \(\lambda < 1\), \(\sigma < \lambda < 1\), which proves the result.

Conversely we could reverse the arguments. Let \(\sigma\) be the largest positive eigenvalue of \((I - R)^{-1}T\) and \(z\) the corresponding eigenvector. Thus
\[
(I - R)^{-1}Tz = \sigma z,
\] (3.6.68)
\[
(\sigma R + T)z = \sigma z.
\] (3.6.69)
Therefore \(\sigma\) is also an eigenvalue of \(\sigma R + T\). If \(\sigma > 1\), then \(\sigma R + T > R + T\) and \(\sigma > \lambda > 1\). For \(\sigma = 1\), then \(\lambda = 1\). Finally if \(\sigma < 1\)
\[
\sigma R + T < R + T.
\] (3.6.70)
Therefore
\[
\sigma < \lambda < 1.
\] (3.6.71)
This last result is the important result as it shows that for nonnegative iteration matrices the method of successive displacements asymptotically converges faster than the method of simultaneous displacements.

For matrices which are not nonnegative, it is possible for one method to work and not the other, and vice versa (see problem 9).
III. NUMERICAL SOLUTIONS OF EQUATIONS

We now illustrate the method by considering the model

\[ \epsilon_{j,k}^{p+1} = \lambda_{m,n} \epsilon_{j,k}^p. \]  \hspace{1cm} (3.6.72)

Stability requires that \(|\lambda_{m,n}| < 1\) for all \(m, n\). Using Eq. (3.6.72) in the difference equation (3.6.56) and factoring, we have

\[ \lambda_{m,n} \epsilon_{j,k}^p = \frac{1}{4} [\epsilon_{j+1,k}^p + \lambda_{m,n} \epsilon_{j-1,k}^p + \epsilon_{j,k+1}^p + \lambda_{m,n} \epsilon_{j,k-1}^p]. \]  \hspace{1cm} (3.6.73)

Following Frankel (Ref. 12) we assume eigenfunctions of the form

\[ \epsilon_{j,k}^{(m,n)} = A^j \sin \frac{n\pi j}{K} B^k \sin \frac{m\pi k}{K}, \]  \hspace{1cm} (3.6.74)

where \(A\) and \(B\) are assumed constants to be determined. Expanding the trigonometric functions in Eq. (3.6.73) yields the equation

\[ \lambda_{m,n} \epsilon_{j,k}^p = \frac{1}{4} [(A^{j+1} + \lambda_{m,n} A^{j-1}) B^k \cos \frac{n\pi j}{K} \sin \frac{m\pi k}{K} \sin \frac{m\pi k}{K} \\
+ (B^{k+1} + \lambda_{m,n} B^{k-1}) A^j \cos \frac{n\pi j}{K} \cos \frac{m\pi k}{K} \cos \frac{m\pi k}{K} \\
+ \frac{1}{4} [(A^{j+1} - \lambda_{m,n} A^{j-1}) B^k \sin \frac{n\pi j}{K} \sin \frac{m\pi k}{K} \cos \frac{n\pi j}{K} \cos \frac{m\pi k}{K} \\
+ (B^{k+1} - \lambda_{m,n} B^{k-1}) A^j \sin \frac{n\pi j}{K} \sin \frac{n\pi j}{K} \cos \frac{m\pi k}{K} \cos \frac{m\pi k}{K}] . \]  \hspace{1cm} (3.6.75)

Since the error must be zero on the boundary, viz. at \(k = 0\) or \(k = K\), the terms in the cosine must vanish. Hence we require

\[ A^{j+1} - \lambda_{m,n} A^{j-1} = 0 \]  \hspace{1cm} (3.6.76)

\[ B^{k+1} - \lambda_{m,n} B^{k-1} = 0. \]  \hspace{1cm} (3.6.77)

Therefore,

\[ A^2 = B^2 = \lambda_{m,n}. \]  \hspace{1cm} (3.6.77)

Equation (3.6.75) then becomes

\[ \lambda_{m,n} = \frac{1}{4} \left( \cos \frac{n\pi}{K} + \cos \frac{m\pi}{K} \right)^2. \]  \hspace{1cm} (3.6.78)
The maximum value is again found for $m = n = 1$. Expanding we have
\[ \lambda_{m,n} \approx 1 - \frac{\pi^2}{K^2}. \]  \hspace{1cm} (3.6.79)

The asymptotic decay rate is then
\[ \nu = \frac{\pi^2}{K^2}. \]  \hspace{1cm} (3.6.80)

Notice that this rate is twice as large as for the method of simultaneous displacements (see Eq. (3.6.54)). Thus, we expect the method of successive displacements to take roughly half as long for the model problem as the method of simultaneous displacements. This result is consistent with the general result for nonnegative irreducible matrices since the Laplace difference equation (3.6.31) leads to a nonnegative irreducible iteration matrix.

In terms of the residuals the method of successive displacements becomes
\[ x_{p+1} = x_p + r_p, \]  \hspace{1cm} (3.6.81)

where
\[ r_p = -(L + D)^{-1}(Ax_p - y). \]  \hspace{1cm} (3.6.82)

It is interesting to write out the component form of the residual for the model problem: we have
\[ r_{p}^{j,k} = \frac{1}{4}[x_{j+1,k}^{p} + x_{j-1,k}^{p} + x_{j,k+1}^{p} + x_{j,k-1}^{p} - 4x_{j,k}^{p}] \]  \hspace{1cm} (3.6.83)

A similar result applies for the method of simultaneous displacements where all terms with superscript $p + 1$ are replaced by like terms with superscript $p$. Notice that the residual can be interpreted as the inbalance between the function $x_{p}^{j,k}$ and the value of the difference relation operating on the function at the $p$th iterate.

\section*{C. Successive Over-Relaxation}

For both of the previous methods, the iteration algorithm could be written
\[ x_{p+1} = x_p + r_p \]  \hspace{1cm} (3.6.84)

with different $r_p$ for different methods. From the discussion above, we interpret the residual as correcting the function at each point, say $j, k$, so as to satisfy the difference equation. Obviously, if any neighboring
III. NUMERICAL SOLUTIONS OF EQUATIONS

Point to \( j, k \) is changed, the residual also changes, as illustrated by Eq. \((3.6.84)\). We might anticipate further changes in the function by over-correcting (or perhaps under-correcting) in hopes of speeding convergence of the iteration. The iteration might then be written

\[
x^{\nu+1} = x^{\nu} + \alpha \varepsilon^{\nu},
\]

where \( \alpha \) is a real number. For \( \alpha > 1 \) we speak of over-relaxation; for \( \alpha < 1 \) we speak of under-relaxation. The method of successive over-relaxation (also called the extrapolated Liebmann method) is defined (see problem 10) as

\[
x^{\nu+1} = x^{\nu} + \alpha\left[-D^{-1}Lx^{\nu+1} - D^{-1}Ux^{\nu} + D^{-1}y\right] - x^{\nu}.
\]

The iteration can also be written

\[
x^{\nu+1} = (D + \alpha L)^{-1}\left[(1 - \alpha)D - \alpha U\right]x^{\nu} + \alpha(D + \alpha L)^{-1}y.
\]

Notice for \( \alpha = 1 \) we recover the method of successive displacements. Again if \( x^{\nu} = A^{-1}y \), then the iteration yields \( x^{\nu+1} = A^{-1}y \), proving consistency.

To illustrate the utility of over-relaxation, consider again the model problem. The iteration algorithm for the errors is

\[
\varepsilon^{\nu+1}_{jk} = (1 - \alpha)\varepsilon^{\nu}_{jk} + \frac{\alpha}{4}\left[\varepsilon^{\nu}_{j+1,k} + \varepsilon^{\nu+1}_{j-1,k} + \varepsilon^{\nu}_{j,k+1} + \varepsilon^{\nu}_{j,k-1}\right].
\]

We again try for the \( m, n \)th component of the error, an expression in the form

\[
\varepsilon^{\nu}_{jk} = A_j \sin \frac{m\pi j}{K} B_k \sin \frac{n\pi k}{K}
\]

and also assume the form

\[
\varepsilon^{\nu+1}_{jk} = \lambda_{mn} \varepsilon^{\nu}_{jk}.
\]

Then we have, after the usual algebra,

\[
\lambda_{mn} = (1 - \alpha) + \frac{\alpha}{2} \sqrt{\lambda_{mn}} \left(\cos \frac{m\pi}{K} + \cos \frac{n\pi}{K}\right).
\]

The largest value of \( \lambda \) will occur for \( m = n = 1 \). Therefore let

\[
\eta = \cos \frac{\pi}{K}.
\]
We then have
\[ \lambda_{11} = (1 - \alpha) + \alpha \eta (\lambda_{11})^{1/2}, \tag{3.6.93} \]
and
\[ (\lambda_{11})^{1/2} = \frac{\alpha \eta + \sqrt{\alpha^2 \eta^2 - 4(\alpha - 1)}}{2}, \tag{3.6.94} \]
for the largest real value of \((\lambda_{11})^{1/2}\). To study the behavior of \(\lambda\) as a function of \(\alpha\), note that for \(\alpha = 1\), we have
\[ \lambda_{11} = \eta^2, \tag{3.6.95} \]
which is consistent with Eq. (3.6.80). To find the slope of \(\lambda_{11}\) versus \(\alpha\), we write Eq. (3.6.93) in the form (with the subscript 11 suppressed)
\[ (\lambda^{1/2})^2 = (1 - \alpha) + \alpha \eta \lambda^{1/2} \tag{3.6.96} \]
and differentiate with respect to \(\alpha\). We have
\[ \frac{d\lambda^{1/2}}{d\alpha} = \eta \lambda^{1/2} - \frac{1}{2 \lambda^{1/2} - \alpha \eta}. \tag{3.6.97} \]
We also have from Eqs. (3.6.96) and (3.6.97)
\[ \frac{d\lambda}{d\alpha} = -1 + \eta \lambda^{1/2} + \alpha \eta \left[ \frac{\eta \lambda^{1/2} - 1}{2 \lambda^{1/2} - \alpha \eta} \right], \tag{3.6.98} \]
or
\[ \frac{d\lambda}{d\alpha} = -\frac{\lambda^{1/2} - \eta \lambda}{\lambda^{1/2} - \frac{\alpha \eta}{2}}. \tag{3.6.99} \]

At \(\alpha = 1\), \(\lambda^{1/2} = \eta\) with \(\eta < 1\). Therefore, as \(\alpha\) increases, \(\lambda\) decreases with ever steeper slope until the denominator vanishes. From Eq. (3.6.94) we see that the minimum \(\lambda\) occurs when
\[ \alpha^2 \eta^2 = 4(\alpha - 1) \tag{3.6.100} \]
or
\[ \alpha = \frac{2 - 2 \sqrt{1 - \eta^2}}{\eta^2} = \frac{2}{1 + \sqrt{1 - \eta^2}}. \tag{3.6.101} \]

With the above value of \(\alpha\), \(\lambda\) becomes
\[ \lambda = \alpha - 1 = \frac{1 - \sqrt{1 - \eta^2}}{1 + \sqrt{1 - \eta^2}}. \tag{3.6.102} \]
To obtain an order of magnitude estimate we expand the expression for $\eta$ to find

$$
\alpha \approx \frac{2}{1 + \frac{\pi}{K}},
$$

(3.6.103)

and with the given (optimum) $\alpha$

$$
\lambda \approx \frac{1 - \pi/K}{1 + \pi/K} \approx 1 - \frac{2\pi}{K}.
$$

(3.6.104)

Using the optimum over-relaxation factor, we have an asymptotic convergence rate of

$$
\nu \approx \frac{2\pi}{K}.
$$

(3.6.105)

Notice that this decay rate is an enormous increase over that obtained for the method of successive displacements. For a 20 by 20 problem, the asymptotic decay rate is approximately

$$
\nu = 0.314.
$$

(3.6.106)

This is greater than a factor of 10 better than the results obtained by using successive displacements. Such gains are characteristic of the method of successive over-relaxation with the optimum over-relaxation parameter.

It is interesting to study the behavior of the method as a function of the over-relaxation parameter. For a given $\eta$ we consider various $\alpha$’s and $\lambda$’s. At $\alpha = 0$, $\lambda = 1$, but this is evident since then no iteration occurs. The slope of the $\lambda$ versus $\alpha$ curve grows steeper from $\alpha = 0$. As $\alpha$ increases from 1, $\lambda$ decreases, the slope reaching minus infinity at the optimum $\alpha$. Beyond the optimum $\alpha$, $\lambda$ begins to increase. Note that for $\alpha$ greater than the optimum, $\lambda$ is complex. The derivative of the magnitude of $\lambda$ is then

$$
\frac{\partial |\lambda|}{\partial \alpha} = \frac{\partial}{\partial \alpha} \left[ \frac{\alpha^2 \eta^2}{4} + \frac{1}{4} (4(\alpha - 1) - \alpha^2 \eta^2) \right] = 1,
$$

(3.6.107)

and hence the magnitude of $\lambda$ increases linearly with $\alpha$. The general shape of the $\lambda$ versus $\alpha$ curve is shown in Fig. 3.6.2.

The results obtained thus far are more general than the derivation indicates. For difference approximations to a wide variety of elliptic
equations, over-relaxation is very worth-while. It can be shown\(^6\) that the results obtained previously for the asymptotic convergence rate and optimum over-relaxation factor apply to more general problems than the model problem. In particular, notice that the definition of \(\eta\), Eq. (3.6.92), is the same as the largest eigenvalue for the method of simultaneous displacements, Eq. (3.6.53). The results, Eqs. (3.6.96) and (3.6.97), are actually expressed in terms of the eigenvalues of the method of simultaneous displacements.

In actual practice one seldom has available the eigenvalues \(\eta\) to predict accurately the optimum over-relaxation parameter. Usually an experimental approach is undertaken where one attempts to find the optimum factor by trial and error. The process of looking for the optimum factor is time consuming, but the actual gain in convergence rate is usually so large that the total time of solution is significantly reduced. It is usually best to choose an \(\alpha\) a little too large than a little too small, because the value of \(\lambda\) will then be closer to the minimum. In general, the optimum \(\alpha\) is unknown so a small error will usually be made in choosing it.

3.6.4 Nonstationary Methods

Frequently it is possible to hasten the convergence of an iteration by considering nonstationary iterations. These are iteration methods in which the technique used depends upon the step of the iteration. In this section we shall consider several such techniques.

First, assume we have a convergent iteration with iteration matrix \(B\). Since the iteration is convergent, the eigenvalues of \(B\), say \(\lambda\), are all

\(^6\) See Forsythe and Wasow, Reference 5, for a very lucid discussion of the more general theory of successive over-relaxation.
less than one in magnitude. We assume we have a dominant eigenvalue, say $\lambda_1$. The errors are asymptotically given in the form

$$e^{n+1} \approx \lambda_1 e^n,$$  \hspace{1cm} (3.6.108)

as shown in Section 3.6.2. Since the errors are not known at any stage of the iteration, we consider the residuals, which are known. The residuals obey the recurrence relation

$$r^{n+1} = (B - I)B(B - I)^{-1}r^n = Br^n,$$  \hspace{1cm} (3.6.109)

as may be seen from Eqs. (3.6.9) and (3.6.16). Consequently the asymptotic behavior of the residuals is also of the form

$$r^{n+1} \approx \lambda_1 r^n.$$  \hspace{1cm} (3.6.110)

Once the residuals have assumed their asymptotic behavior (and this is measurable), we may predict the total of the future displacements. Thus

$$r^{n+2} = \lambda_1^2 r^n.$$  \hspace{1cm} (3.6.111)

The total displacement, say $r^t$, is then

$$r^t = r^n + \lambda_1 r^n + \ldots + \lambda_1^\infty r^n = \left( \frac{1}{1 - \lambda_1} \right) r^n,$$  \hspace{1cm} (3.6.112)

and the solution approximated as

$$x^t = x^n + r^t.$$  \hspace{1cm} (3.6.113)

We cannot expect the acceleration to yield the true answer since the factor $1/(1 - \lambda_1)$ will amplify harmonics other than the first. Nevertheless, the acceleration is frequently useful. The measurement of $\lambda_1$ is usually done experimentally. That is, one monitors the ratio

$$\frac{|r^{n+1}|}{|r^n|} \approx \lambda_1$$

for successive values of $p$. When the variation in the approximate value of $\lambda_1$ is small, then the acceleration may be attempted. The use of this technique for solving linear algebraic equations is analogous to a method for solving nonlinear algebraic equations, in which case the acceleration is known as Aitken's $\delta^2$ method.
### 3.6 Elliptic Equations and Iterative Methods

#### A. The Alternating Direction Method

The alternating direction implicit method of Peaceman and Rachford may also be applied to the iterative solution of elliptic equations. The following development is from Varga (Reference 8). We must require that the matrix operator be of block tridiagonal form and contain at most five nonzero elements per row. The basis of the method is to assume the iteration is analogous to a time variable. If we desire to solve the equation

$$Ax = y,$$  \hspace{1cm} (3.6.114)

then we consider the equation

$$\frac{dx}{dt} = Ax - y.$$  \hspace{1cm} (3.6.115)

For any $x$ satisfying Eq. (3.6.114), the derivative of $x$ must be zero. Using a simple difference approximation we have

$$\frac{x^{p+1} - x^{\nu}}{\Delta t_p} = Ax^{\nu} - y,$$  \hspace{1cm} (3.6.116)

where $\Delta t_P$ corresponds to a single time step. For elliptic equations in two dimensions, the spacing $h^2$ appears as a factor in the matrix $A$, consequently the ratio $\Delta t_p/h^2 = r_p^2$ is also a factor of $A$. We now write $A$ in the form

$$A = (Y + Z)/h^2,$$  \hspace{1cm} (3.6.117)

where we assume $Y, Z$ symmetric and positive definite, and with at most three elements per row. The two matrices $Y, Z$ are taken to correspond to second differences in the two spatial dimensions $y$ and $z$ respectively. The iteration consists of time steps (iterations) in alternate directions and of the form

$$x^{p+1} = x^p + r_p^2[Yx^{p+1} + Zx^p - h^2y],$$  \hspace{1cm} (3.6.118)

and

$$x^{\nu+2} = x^{\nu+1} + r_p^2[Yx^{\nu+1} + Zx^{\nu+2} - h^2y].$$  \hspace{1cm} (3.6.119)

The above equations can be written as

$$x^{p+1} = (I - r_p^2Y)^{-1}(I + r_p^2Z)x^p - h^2r_p^2(I - r_p^2Y)^{-1}y,$$  \hspace{1cm} (3.6.120)
and
\[ x^{p+2} = (I - r_p^2 Z)^{-1}(I + r_p^2 Y)x^{p+1} - h_{p+2}^2(I - r_p^2 Z)^{-1}y, \] (3.6.121)
which define the iteration algorithm.

Before applying the iteration to the model problem, several comments are pertinent. We assume the factor \( r_p^2 \) may be a function of \( p \), and in fact is chosen so as to speed convergence. However, \( r_p^2 \) must not vary in going from \( x^p \) to \( x^{p+2} \). That is, \( r_p^2 \) must be the same for the two alternating direction sweeps. Also, we take the operators \( Y \) and \( Z \) to be three-point relations so the rapid method of matrix factorization of Section 1.14 may be used.

The virtues of this method become evident by applying it to the model problem for which we can find an analytical relation. Applied to this problem, the method yields the error equations
\[ [2 + (1/r_p^2)]e_{jk}^{p+1} - e_{j-1,k}^{p+1} - e_{j+1,k}^{p+1} = e_{j,k}^p - e_{j,k-1}^p - [2 - (1/r_p^2)]e_{jk}^p \] (3.6.122)
and
\[ [2 + (1/r_p^2)]e_{jk}^{p+2} - e_{j,k-1}^{p+2} - e_{j,k+1}^{p+2} = e_{j+1,k}^{p+1} + e_{j-1,k}^{p+1} - [2 - (1/r_p^2)]e_{jk}^{p+1}. \] (3.6.123)

From the results of Section 3.5, we can analytically find the eigenvalues of the two-step iteration to be
\[ \lambda_{m,n} = \frac{(1/r_p^2) - 4 \sin^2 \frac{n\pi}{2K}}{(1/r_p^2) + 4 \sin^2 \frac{n\pi}{2K}} - \frac{(1/r_p^2) - 4 \sin^2 \frac{m\pi}{2K}}{(1/r_p^2) + 4 \sin^2 \frac{m\pi}{2K}}. \] (3.6.124)

It is evident from Eq. (3.6.124) that if \( r_p^2 \) (i.e., \( \Delta t_p \)) is properly chosen, we can make one eigenvalue equal to zero, and hence eliminate that error component by means of each two-step iteration. If the total number of unknowns is \( N \), then in \( 2N \) steps the iterations will have converged quite precisely. However, in general it is not necessary to perform so many iterations; a sufficiently small error, though not zero error, can be found in rather fewer steps. To this end, an average factor \( r_p^2 \) is used to reduce significantly particular harmonics; the factor \( r_p^2 \) must be chosen, of course, so that other harmonics are not increased. Indeed, in the general problem, the most difficult part of the alternating direction method is the choice of a good average value for \( r_p^2 \). With enough average values the errors can be significantly reduced by using them cyclically.
Appropriate average factors $1/r_p^2$ can be easily found for the model problem. These will give an idea of the rate of convergence of the method. To find the appropriate average factors, we note from Eq. (3.6.124) that the values of $1/r_p^2$ found analytically in the model problem consist of a discrete spectrum between 0 and 4. Our present objective is to replace this spectrum with fewer, but representative, entries. We have two alternatives available: we can either choose the number of subintervals, or we can choose the error-reduction factor. In either case there is an optimum way of averaging over each subinterval in choosing a representative value of $1/r_p^2$ for this interval.

We note that all of the harmonics have factors of the form $\sin^2(n\pi/2K)$. Let

$$\bar{\beta}_n = \sin^2\frac{n\pi}{2K}. \quad (3.6.125)$$

The factor $\bar{\beta}$ ranges from near zero, say $a$, to near unity, say $b$. We now subdivide the interval $a - b$ into subintervals $\beta_i$, $i = 0, 1, \ldots, I$, and such that

$$\sin^2\frac{(K-1)\pi}{2K} = b = \beta_0 > \beta_1 > \beta_2 > \ldots > \beta_I = a = \sin^2\frac{\pi}{2K}.$$

Suppose we desire to choose the factors $1/r_i^2$ such that an iteration with the given $1/r_i^2$ reduces all harmonics between $\beta_i$ and $\beta_{i+1}$ by a factor of at least $\alpha$. We require then

$$\max \left| \frac{1/r_i^2 - 4\beta}{1/r_i^2 + 4\beta} \right| = \alpha, \quad \beta_{i+1} \leq \beta \leq \beta_i. \quad (3.6.126)$$

Consequently, the best choice of the $\beta_i$ is that for which the maximum occurs at the end points, i.e.,

$$\frac{1/r_i^2 - 4\beta_i}{1/r_i^2 + 4\beta_i} = \frac{1/r_{i+1}^2 - 4\beta_{i+1}}{1/r_{i+1}^2 + 4\beta_{i+1}} = \alpha. \quad (3.6.127)$$

By eliminating $1/r_i^2$ from the equation, we find the intervals must be so chosen that

$$\beta_{i+1} = \left(\frac{1 - \alpha}{1 + \alpha}\right)^2 \beta_i \quad (3.6.128)$$

and therefore

$$\beta_i = \left(\frac{1 - \alpha}{1 + \alpha}\right)^{2i} b, \quad i = 0, 1, \ldots, I. \quad (3.6.129)$$
Therefore, having selected the reduction factor $\alpha$ and knowing the interval in which the function $\beta$ varies, we find the number of intervals needed and the widths of each. Note that for small $\alpha$, many narrow intervals are needed, whereas for large $\alpha$ (but less than unity), few intervals are needed. Having found the interval widths, we use Eq. (3.6.127) to find $r_i^2$. The result is

$$\frac{1}{r_i^2} = 4b \left(\frac{1 - \alpha}{1 + \alpha}\right)^{2i+1}, \quad i = 1, 2, ..., l,$$  

(3.6.130)

which are the average factors $1/r_i^2$ desired. With the appropriate values of $r_i^2$, one complete cycle of iterations reduces all the errors by at least $\alpha^2$ as seen from Eq. (3.6.124).

To find the number of intervals needed to reduce each error by at least a factor of $\alpha^2$, we must take a number of intervals (and hence iterations) given by Eq. (3.6.129) as

$$\beta_I \equiv \sin^2 \frac{\pi}{2K} = \left(\frac{1 - \alpha}{1 + \alpha}\right)^{2I} \sin^2 \frac{(K - 1)\pi}{2K}.$$  

(3.6.131)

By the usual expansions, for $K$ this becomes

$$\left(\frac{\pi}{2K}\right)^2 \approx \left(\frac{1 - \alpha}{1 + \alpha}\right)^{2I}$$

or

$$I = \frac{\ln K - \ln \frac{\pi}{2}}{\ln \left(\frac{1 + \alpha}{1 - \alpha}\right)} \approx \frac{\ln K}{2\alpha}.$$  

(3.6.132)

Therefore, to reduce the error by a factor of $e$, i.e., $\alpha = e^{-1/2}$, approximately $(.8)\ln K$ iterations are required. For $K = 100$ this means approximately 4 iterations. For comparison, the method of over-relaxation would require approximately 15 iterations. In general, the method is better than the method of over-relaxation by a factor of $(K/2)\ln K$, approximately.

The alternating direction method is admirable for the model problem. The large gain in convergence more than offsets the additional algebra due to the implicit nature of the equations. For more general problems, proof of the large gain in convergence has not been established. Nevertheless the method is used to good advantage for other two-dimensional problems.
3.6 ELLIPTIC EQUATIONS AND ITERATIVE METHODS

B. Method of Steepest Descent

We consider again the matrix equation

$$Ax = y$$

(3.6.133)

and an associated iteration algorithm

$$x^{p+1} = x^p + \alpha_p z^p,$$  (3.6.134)

where $\alpha_p$ is a scalar and $z^p$ a correction vector. In the previously considered methods, the correction vector was obtained from the residuals in a straightforward manner. The method of steepest descent is derived by considering a geometric interpretation of the iteration. To this end we define the function

$$s^p = Ax^p - y,$$  (3.6.135)

which is the amount by which the trial solution $x^p$ fails to satisfy the original equation (3.6.133). The function

$$f(x^p) = (s^p)^T(s^p) = (Ax^p - y)^T(Ax^p - y) = \text{constant}$$  (3.6.136)

is the equation for an ellipsoid in the $n$ dimensional space of the problem, $n$ being the order of the operator $A$. The center of the ellipsoid is the point $x = A^{-1}y$. The solution of the matrix equations is then equivalent to finding the center of the ellipsoids defined by Eq. (3.6.136). At each step of the iteration, we desire a correction vector which will move us near the center of the ellipsoids. From elementary vector analysis, we know the direction of greatest change from a surface is the direction of the gradient vector. Hence a reasonable choice for $z^p$ is the gradient of $f(x^p)$.

To simplify the algebra we shall assume $A$ is positive definite and symmetric. In this case $A$ can be written

$$A = B^T B,$$  (3.6.137)

an equation that has a solution since there are more elements in $B$ than in $A$. The original equation (3.6.133) is equivalent to

$$Bx = u,$$  (3.6.138)

$$B^Tu = y.$$

We define the quadratic surface

$$f(x) = (Bx - u)^T(Bx - u) = x^T A x - 2x^T y + u^T u.$$  (3.6.139)
This function vanishes for $x = A^{-1}y$. The gradient to the surface is
\[
\frac{df(x)}{dx_i} = 2(Ax - y)_i,
\] (3.6.140)
hence for $x = x^p$, the gradient is
\[
\frac{df(x^p)}{dx_i} = 2s_i^p.
\] (3.6.141)

We take the correction $z^p$ proportional to $s^p$. The choice of $\alpha_p$ can be made in the following manner. We wish to add a correction to the approximate solution, in the direction $s$, such that the magnitude of the function $f(x^{p+1})$ is minimized. Thus we desire
\[
\frac{\partial}{\partial \alpha} [f(x^p + \alpha s^p)] = 0.
\] (3.6.142)

By simple operations we find
\[
\alpha_p = -\frac{(s^p)^T(s^p)}{(s^p)^T A(s^p)}.
\] (3.6.143)

If one treats the method of steepest descent as a stationary problem with $\alpha_p$ the appropriate constant, then the method is the same as the method of simultaneous displacements. The convergence properties are thus similar to the same properties for simultaneous displacements. However, the geometric interpretation lends itself to a much improved gradient method which we consider next.

**C. Method of Conjugate Gradients**

The method of conjugate gradients is based upon the concept that along a line the shortest distance to a point (to be taken as the origin) is the perpendicular through that point to the line. In constructing a path to that point, we may proceed to the foot of the perpendicular along the line. The path to the origin may be traced in Fig. 3.6.3. Of course, only

![Fig. 3.6.3. A typical path followed in the conjugate gradient procedure.](image)
3 dimensions are illustrated in this figure. From this foot we never again have to travel in any direction with a component along the given line. Indeed, we need move only in a plane through the point and perpendicular to the given line. (This plane contains the original perpendicular, of course). In this plane, we may strike out in some arbitrary direction, which will be perpendicular to the original line. The shortest distance to the point from this arbitrary direction is again a new perpendicular to this direction (and also the original line) through the point. The previous process can then be repeated again. In a space of \( n \) dimensions, only \( n \) repetitions of this procedure are required to bring us precisely to the point desired—apart from round-off error.

The actual selection of the directions of correction is relatively simple. We desire that successive correction vectors be orthogonal to all previous corrections and further, we desire to minimize the error as much as possible at each step. We again consider the quadratic surface \( f(x^0) \) [as defined by Eq. (3.6.139)] and the gradient vector \( s^0 \) [as defined by Eq. (3.6.140)]. We correct \( x^0 \) in the direction \( s^0 \) and such that \( f(x^1) \) is minimized. The vector from the origin to \( x^1 \) lies in a plane orthogonal to \( s^0 \). For the next correction we find the component of the gradient of \( f(x^1) \) which lies in the plane orthogonal to \( s^0 \). We move along the correction vector to \( x^2 \) a distance that minimizes \( f(x^2) \). We continue in this manner taking corrections as components of the gradient which are orthogonal to the previous directions of correction.

In general the process of orthogonalization of a set of vectors is quite complicated. Since we are dealing with projections of gradients, however, a remarkable simplification is possible. To illustrate, consider the starting vector \( x^0 \) and the initial gradient \( z^0 = Ax^0 - y \). We choose \( x^1 \) as in the steepest descent, i.e.,

\[
x^1 = x^0 + \alpha_0 s^0,
\]

whence

\[
\alpha_0 = -\frac{(z^0)^T z^0}{(z^0)^T A z^0}.
\]

For \( z^1 \) we take the gradient

\[
z^1 = Ax^1 - y = z^0 + \alpha_0 Az^0.
\]

Note that \( z^1 \) is orthogonal to \( z^0 \) since

\[
(z^0)^T z^1 = (z^0)^T z^0 + \alpha_0 (z^0)^T A z^0 = 0
\]
for $\alpha_0$ as in Eq. (3.6.145). If we take $x^2$ in the form

$$x^2 = x^1 + \alpha_1 z^1 = x^0 + \alpha_0 z^0 + \alpha_1 z^1,$$

(3.6.148)

then the gradient is

$$z^2 = Ax^2 - y = z^1 + \alpha_1 Az^1,$$

(3.6.149)

and this vector is not orthogonal to $z^0$ for any $\alpha_1 \neq 0$. We desire to correct $x^2$ such that the vector $z^2$ is orthogonal to $z^0$ and $z^1$. To this end it is convenient to write the residual in the form

$$z^2 = a_1(z^1 + b_0 z^0 + \alpha_1 Az^1).$$

(3.6.150)

The orthogonality requires that

$$\alpha_1 = -\frac{(z^1)^T z^1}{(z^1)^T Az^1},$$

(3.6.151)

and

$$b_0 = -\alpha_1 \frac{(z^0)^T Az^1}{(z^0)^T z^0}. $$

(3.6.152)

The correction is derived by substituting for $z^i$ in the form $z^i = Ax^i - y$. We have

$$Ax^2 - y = a_1(Ax^1 - y + b_0(Ax^0 - y) + \alpha_1 Az^1).$$

(3.6.153)

Thus,

$$x^2 = a_1[x^1 + b_0 x^0 + \alpha_1 z^1]$$

(3.6.154)

for $a_1$ such that

$$a_1(1 + b_0) = 1.$$

(3.6.155)

Having computed $x^2$ such that $z^2$ is orthogonal to $z^0$ and $z^1$, we now consider correcting $x^3$ such that $z^3$ is orthogonal to $z^0$, $z^1$, $z^2$. We expect $z^3$ of the form

$$z^3 = a_2(z^2 + b_1 z^1 + c_0 z^0 + \alpha_2 Az^2).$$

(3.6.156)

However, the coefficient $c_0$ must be zero since we have

$$(z^0)^T z^3 = a_2(c_0(z^0)^T z^0 + \alpha_2(z^0)^T Az^2)$$

$$= a_2(c_0(z^0)^T z^0 + \alpha_2(z^0)^T Az^0).$$

(3.6.157)

Notice the vector $Az^0$ is a linear combination of $z^1$ and $z^0$. By the previous orthogonality $(z^0)^T z^0 = (z^2)^T z^1 = 0$. Hence $c_0 = 0$. In general one can show by the above argument that $z^j$ consists only of terms in $z^{j-1}$,
3.6 ELLIPTIC EQUATIONS AND ITERATIVE METHODS

Consequently the expressions for the correction vectors and successive estimates are only three term recursions. The general rule is of the form

\[ z_j = a_{j-1} z_j^{-1} + b_{j-2} z_{j-2} + \alpha_{j-1} A z_j^{j-1} \tag{3.6.158} \]

and

\[ x_j = a_{j-1} x_j^{-1} + b_{j-2} x_{j-2} + \alpha_{j-1} A x_j^{j-1}. \tag{3.6.159} \]

The coefficients are computed in the form

\[ \alpha_{j-1} = -\frac{(z_j^{-1})^T z_j^{-1}}{(z_j^{-1})^T A z_j^{j-1}}, \tag{3.6.160a} \]

\[ b_{j-2} = -\frac{\alpha_{j-1} (z_j^{-1})^T A z_j^{j-2}}{(z_j^{-2})^T A z_j^{j-2}}, \tag{3.6.160b} \]

and

\[ a_{j-1}(1 + b_{j-2}) = 1. \tag{3.6.160c} \]

By the orthogonality of the \( z^i \), we must have \( z^N = 0 \) for \( N \) equal the dimension of the problem, and hence \( x^{N-1} \) is the desired solution.

The method of conjugate gradients is an \( N \)-step process which would yield an exact result if there were no round-off. With round-off an additional step may be necessary to correct the approximate solution. At each step we require something like \( N^2 \) operations (from the product \( A z^i \) and \( N \) steps. Consequently, the total number of steps is proportional to \( N^3 \). For a sparse matrix the number of operations is much reduced.

3.6.5 EIGENVALUE PROBLEMS

Of frequent interest to nuclear engineers is the study of critical assemblies. The mathematical formulations of such problems lead to eigenvalue problems. In general, the nuclear engineer or physicist is interested in the smallest eigenvalue, that is, the first critical mode. Rarely is he concerned with the entire set of eigenvalues of a reactor system.

This limited interest greatly simplifies the numerical study of reactive assemblies. Although a variety of methods have been derived for generating the eigenvalues and eigenvectors of a matrix, many simplifications are possible when one is interested only in the largest or smallest eigenvalue. Hereafter we shall be concerned with criticality problems and assume we are interested in approximating the smallest eigenvalue of a given assembly.
For an introductory example, consider the simple one-group, one-region diffusion equation

\[
\frac{d^2 y}{dx^2} + B^2 y = 0 \quad (3.6.161)
\]

with boundary conditions

\[ y(0) = y(a) = 0. \]

The simple difference approximation

\[
\delta^2 y_k + B^2 h^2 y_k = 0 \quad (3.6.162)
\]

can be easily solved analytically. We know the lowest mode has the solution

\[
y_k = \sin \frac{kx}{K}, \quad (3.6.163a)
\]

\[
B_1^2 = \frac{2}{h^2} \left(1 - \cos \frac{x}{K}\right), \quad (3.6.163b)
\]

where the subscript 1 is the mode index. For small \( h \), the results \( (3.6.163a, b) \) approach the analytic solution.

One approach to solving the problem is to write Eq. (3.6.162) in matrix form

\[
Ay = \lambda y, \quad (3.6.164)
\]

where

\[
A = \begin{bmatrix}
2 & -1 & 0 & \ldots & 0 & 0 \\
-1 & 2 & -1 & \ldots & 0 & 0 \\
0 & -1 & 2 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 2 & -1 \\
0 & 0 & 0 & \ldots & -1 & 2
\end{bmatrix} \quad (3.6.165)
\]

and \( \lambda = B^2 h^2 \).

To generate the solution we might take a trial vector \( y^0 \), operate with the matrix \( A \) repeatedly until we achieve a repeating distribution. This technique is called the power method. If we assume \( A \) has a complete set of eigenvectors, then we expand \( y^0 \) in terms of the eigenvectors, say \( e_r \). We have

\[
y^0 = \sum_{r=1}^{n} a_r e_r. \quad (3.6.166)
\]
Successive applications of the matrix \( \mathbf{A} \) yield

\[
y^p = \mathbf{A} y^{p-1} = \mathbf{A}^p y^0
\]  

(3.6.167)

\[
= \lambda_1^p \left[ a_1 e_1 + \sum_{r=2}^{n} a_r \left( \frac{\lambda_r}{\lambda_1} \right)^p e_r \right].
\]  

(3.6.168)

If \( |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n| \), then ultimately we have

\[
y^{p+1} \approx a_1 \lambda_1^p e_1.
\]  

(3.6.169)

We generate a repeating function which is the eigenvector \( e_1 \) corresponding to the eigenvalue of largest magnitude. Unfortunately, we desire the eigenfunction corresponding to the smallest eigenvalue. There are several elementary methods of recasting the problem so that the smallest eigenvalue can be obtained. For instance, if we know the largest eigenvalue is bounded by some scalar, say \( \beta \), then the matrix, \( \beta \mathbf{I} - \mathbf{A} \), has eigenvalues, \( \mu_r = \beta - \lambda_r \), and the largest \( \mu_r \) corresponds to the smallest \( \lambda_r \) of \( \mathbf{A} \).

Our primary concern in this section will be to consider the use of iteration methods for solving eigenvalue problems and some associated difficulties. We shall use the one-group diffusion equation in Cartesian coordinates as a model problem to illustrate techniques.

We consider first the basic eigenvalue problem

\[
\mathbf{A} \mathbf{x} = \lambda \mathbf{x},
\]  

(3.6.170)

where we assume we are interested in the smallest eigenvalue, say \( \lambda_{\text{min}} \). We write the basic equation in the form

\[
(\mathbf{A} - \lambda \mathbf{I}) \mathbf{x} = 0,
\]  

(3.6.171)

and again factor \( \mathbf{A} \) into its upper, lower, and diagonal portions. The method of simultaneous displacements then gives the iteration rule

\[
\mathbf{x}^{p+1} = -(\mathbf{D} - \lambda \mathbf{I})^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^p,
\]  

(3.6.172)

provided \( (\mathbf{D} - \lambda \mathbf{I})^{-1} \) exists (which it will unless \( \mathbf{D} \) is the diagonal matrix some of whose elements are equal to \( \lambda \)). Define the iteration matrix as

\[
\mathbf{B} = -(\mathbf{D} - \lambda \mathbf{I})^{-1}(\mathbf{L} + \mathbf{U}).
\]  

(3.6.173)

If all the eigenvalues of \( \mathbf{B} \), say \( \mu \), are less than unity in magnitude, then the iteration (3.6.172) will always converge to the trivial solution.
On the other hand, if any eigenvalue of $B$ is greater than unity, the iteration diverges. To achieve a nontrivial, but finite, solution, we must require that the largest eigenvalue of $B$ be exactly unity.

The matrices formed from the difference approximations to elliptic differential equations are frequently of a very special form called a Stieltjes matrix. A Stieltjes matrix is a symmetric, positive definite, irreducible matrix with nonpositive off-diagonal elements. The matrix (3.6.165) is a typical example. We now let $A$ be a Stieltjes matrix and $\lambda_r$ and $e_r$ such that

$$Ae_r = \lambda_r e_r.$$  

(3.6.174)

We then have

$$(A - \lambda_r I)e_r = 0,$$

$$(D - \lambda_r I)e_r = -(L + U)e_r,$$

and finally

$$e_r = -(D - \lambda_r I)^{-1}(L + U)e_r.$$  

(3.6.175)

Therefore, $e_r$ is an eigenvector of $B$ with eigenvalue unity. Now let $\lambda_r = \lambda_{\text{min}}$. In this case the matrix $B$ is nonnegative and of the form

$$B = -\begin{bmatrix}
0 & \frac{a_{12}}{a_{11} - \lambda_{\text{min}}} & \frac{a_{13}}{a_{11} - \lambda_{\text{min}}} & \cdots & \frac{a_{1n}}{a_{11} - \lambda_{\text{min}}} \\
\frac{a_{21}}{a_{22} - \lambda_{\text{min}}} & 0 & \frac{a_{23}}{a_{22} - \lambda_{\text{min}}} & \cdots & \frac{a_{2n}}{a_{22} - \lambda_{\text{min}}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{a_{n1}}{a_{nn} - \lambda_{\text{min}}} & \frac{a_{n2}}{a_{nn} - \lambda_{\text{min}}} & \frac{a_{n3}}{a_{nn} - \lambda_{\text{min}}} & \cdots & 0
\end{bmatrix}$$  

(3.6.176)

(see problem 13).

The eigenvector of $A$ with eigenvalue $\lambda_{\text{min}}$ is also an eigenvector of $B$ as defined in Eq. (3.6.176) and with eigenvalue unity, as may be seen from Eq. (3.6.175). Since $A$ is positive definite, the other eigenvalues of $A$ are greater than $\lambda_{\text{min}}$. From the results of Section 1.13, if the value of $\lambda$ in Eq. (3.6.175) decreases from $\lambda_{\text{min}}$, the spectral radius decreases from unity; conversely, if the value of $\lambda$ increases from $\lambda_{\text{min}}$, the spectral radius increases. Hence the iteration using the method of simultaneous displacements is admirably suited for finding the smallest eigenvalue of a Stieltjes matrix. The eigenvalue of $A$ may be found from the eigenvector by Eq. (3.6.174).

In a manner similar to that above, it is also possible to show that the methods of successive displacements and successive over-relaxation
(with appropriate relaxation parameters) provide a means of finding the smallest eigenvalue of a Stieltjes matrix.

The actual process of finding the eigenvalue and eigenvector is relatively complicated. Assuming we have no prior knowledge concerning the desired eigenvalue, it is probable that any guess will yield an iteration matrix that diverges or else converges to zero. Thus, our iteration must not only attempt to find a solution to the equations, but must also continually improve the approximation to the eigenvalue. In effect we have two concurrent iterations; one iteration to find the eigenvector and one iteration to find the eigenvalue.

The iteration techniques outlined earlier are suitable for finding the eigenvector, and we now consider techniques for locating the eigenvalue. Let the iteration be of the form

$$x^{p+1} = Bx^p.$$  \hspace{1cm} (3.6.177)

We recall that the eigenvalues of $B$, say $\gamma$, are related to the eigenvalues $\lambda$ of the matrix $A$, where $A$ is the matrix we desire to solve. If we expand the trial solution $x^0$ in terms of the eigenvectors of $B$, say $v_r$, with expansion coefficients $a_r$, then we have

$$x^p = \sum_r a_r(\gamma_r)^p v_r.$$  \hspace{1cm} (3.6.178)

After a sufficient number of iterations, we have approximately

$$x^p \approx \gamma_1x^{p-1},$$  \hspace{1cm} (3.6.179)

where $\gamma_1$ is the largest eigenvalue of $B$. The approximate value of $\gamma_1$ may be obtained by comparing successive values of some component of the vector $x$. Having obtained a value of $\gamma_1$, we can adjust the value of $\lambda$. Thus, from a previous result of this section concerning the matrix $B$, if $\gamma_1 > 1$, we decrease $\lambda$, while if $\gamma_1 < 1$, we increase $\lambda$. The convergence of this method is dependent upon the ratio of the eigenvalues of $B$. In general we have

$$x^p = \gamma_1^p \left[ \sum_r a_r \left( \frac{\gamma_r}{\gamma_1} \right)^p v_r \right].$$  \hspace{1cm} (3.6.180)

If $\gamma_2$ is the second largest eigenvalue of $B$, then the convergence is given by the factor $(\gamma_2/\gamma_1)^p$.

An alternative technique is to use Eq. (3.6.178) and compute the quantity

$$(x^p)^T(x^p) = (x^p)^T Bx^{p-1}.$$  \hspace{1cm} (3.6.181)
After a sufficient number of iterations, we have

\[ \frac{(x^p)^T(x^p)}{(x^{p-1})^T(x^{p-1})} \approx \gamma_1. \]  

(3.6.182)

Equation (3.6.182) is known as the Rayleigh quotient. If the eigenvectors of \( \mathbf{B} \) are orthogonal, then the method converges faster than one in which a single component is examined. We have

\[ (x^p)^T(x^p) = \sum_{r=1}^{R} a_r^2 \gamma_r^{2p} = \gamma_1^{2p} \sum_{r=1}^{R} a_r^2 \left( \frac{\gamma_r}{\gamma_1} \right)^{2p} \]

\[ (x^p)^T(x^{p-1}) = \sum_{r=1}^{R} a_r^2 \gamma_r^{2p-1} = \gamma_1^{2p-1} \sum_{r=1}^{R} a_r^2 \left( \frac{\gamma_r}{\gamma_1} \right)^{2p-1} \]

and hence

\[ \frac{(x^p)^T(x^p)}{(x^p)^T(x^{p-1})} \approx \gamma_1 \left[ \frac{a_1^2 + a_2^2 \left( \frac{\gamma_2}{\gamma_1} \right)^{2p}}{a_1^2 + a_2^2 \left( \frac{\gamma_2}{\gamma_1} \right)^{2p-1}} \right]. \]  

(3.6.184)

The convergence to \( \gamma_1 \) is thus given by a factor \( (\gamma_2/\gamma_1)^{2p} \), which is twice as fast as given by Eq. (3.6.180).

We now describe an equivalent procedure which is quite suited to eigenvalue problems. We assume an iteration of the form

\[ x^{p+1} = \mathbf{B} x^p \]  

(3.6.185)

and the initial trial solution \( x^0 \). We define the scalar \( \alpha_0 \) as

\[ [(x^0)^T(x^0)]^{1/2} = \alpha_0. \]  

(3.6.186)

We now compute \( x^1 \) from the iteration rule (3.6.185) and also the factor

\[ \alpha_1 = [(x^1)^T(x^1)]^{1/2}. \]  

(3.6.187)

We now require that \( x^1 \) be scaled as \( \tilde{x}^1 \) such that

\[ [(\tilde{x}^1)^T(\tilde{x}^1)]^{1/2} = \alpha_0. \]  

(3.6.188)

Thus we multiply \( x^1 \) by the factor \( \sqrt{\alpha_0/\alpha_1} = \beta_1 \). We continue in this manner generating the successive trial vectors and scaling the results by a factor

\[ \beta_p = \sqrt{\frac{\alpha_0}{\alpha_p}}. \]  

(3.6.189)
The factors $\alpha_p$ approach an asymptotic value $\gamma_1$ and at a rate dependent on $(\gamma_2/\gamma_1)^{2p}$, as may be seen from the discussion of the previous paragraph, which is applicable here. Thus we can approximately determine $\gamma_1$ by monitoring the quantity $\alpha$. Furthermore, the scaling of the trial vectors prevents the solution from degenerating before we reach an asymptotic behavior even if $\gamma_1 \neq 1$. The advantage of this technique is consequently two-fold; a rapid convergence and a nondegenerate trial solution. The computation of the scale factors takes time, of course, but the increased convergence rate usually makes up for the additional computations.

To illustrate the steps in solving a simple problem, we consider the equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + B^2 \phi = 0, \quad 0 \leq x \leq a, \quad 0 \leq y \leq a, \quad (3.6.190)$$

with boundary conditions

$$\phi(0, y) = \phi(a, y) = 0, \quad (3.6.191)$$
$$\phi(x, 0) = \phi(x, a) = 0.$$

We desire the smallest critical buckling $B_{mn}$. We consider the difference approximation

$$-\frac{\delta^2 \phi_{j,k}}{h_x^2} - \frac{\delta^2 \phi_{j,k}}{h_y^2} - B^2 \phi_{j,k} = 0. \quad (3.6.192)$$

Let $h_x = h_y = h$. Denote the initial estimate of $B_{mn}$ as $B_0$. For simplicity, we use the method of simultaneous displacements. The iteration becomes

$$\phi_{j,k}^{n+1} = \frac{1}{4 - B^2 h^2} \left[ \phi_{j+1,k}^n + \phi_{j-1,k}^n + \phi_{j,k+1}^n + \phi_{j,k-1}^n \right]. \quad (3.6.193)$$

The errors obey a similar iteration rule. The stability of the iteration can be studied by expanding the errors in the usual Fourier series. The resulting computation yields

$$\lambda_{mn} = \frac{2}{4 - B^2 h^2} \left[ \cos \frac{n\pi}{K} + \cos \frac{m\pi}{K} \right]. \quad (3.6.194)$$

We again consider the first harmonic, i.e., $m = n = 1$. We have

$$\lambda_{11} = \frac{4 \cos \frac{\pi}{K}}{4 - B^2 h^2}. \quad (3.6.195)$$
In order for the largest eigenvalue of the iteration matrix to be exactly unity, we require the buckling to be

\[ B^2 = \frac{4}{h^2} \left[ 1 - \cos \frac{\pi}{K} \right], \]  

(3.6.196)

which is well known.

Note that for \( B^2 \) greater than the right-hand side of Eq. (3.6.196), the iteration diverges. Conversely for \( B \) less than the requisite quantity, the iteration converges to zero. Denote the proper value of \( B^2 \) [from Eq. (3.6.196)] as \( B^2_0 \). We then have

\[
\begin{align*}
B^2 > B^2_0, & \quad \text{divergence;} \\
B^2 < B^2_0, & \quad \text{trivial solution;} \\
B^2 = B^2_0, & \quad \text{unique solution.}
\end{align*}
\]

The nuclear analyst recognizes the above three conditions as corresponding to a super-critical, sub-critical, and just critical reactor. Consequently, the iteration displays the same properties we expect the reactor to have. A generalization of this result to more involved equations and iteration methods is found in Birkhoff and Varga (Reference 14).

References

There are many excellent texts dealing with the numerical solution of equations. Particularly useful are 1, 2, 3, and 4. There are few elementary texts that deal with the numerical solution of partial differential equations in detail. One of the best texts is 5; however, the level is above elementary. An excellent treatment of difference methods for parabolic and hyperbolic equations is 6; again the level is not elementary. A very readable discussion of partial difference equations is found in Chapter III of 7. An excellent survey of iterative methods, at a relatively advanced level, is 8. Specific techniques are reviewed in many papers throughout the literature. Generalizations of results quoted in the text are cited.


**Problems**

1. Find the truncation error for the integration of \( \sin x \) from \( 0 \leq x \leq \pi \) by Simpson’s rule. Compare the result with the trapezoidal rule [i.e., truncate the integration at first differences].

2. Derive an integration formula over the interval \( x_i \) to \( x_{i+4} \) in terms of the central difference operator.

3. Discuss the stability of the numerical solution of the equation

\[
\frac{d^2y}{dx^2} = y, \quad y(0) = 0, \quad y'(0) = 1.
\]

Use an approximate difference quotient and also an integration formula.

4. Derive a difference approximation to the one-dimensional wave equation that is accurate to \( O(h^4) + O(h^4) \). Sketch the point pattern.

5. Analyze the stability and accuracy of the following approximation to the heat-flow equation.

![Heat-flow equation diagram](image-url)
6. What order truncation error does the approximation
\[
\frac{A - \nabla}{2h_t} T_{i,k} = \frac{\delta^2}{h_z^2} T_{i,k}
\]
have? What stability criterion is applicable?

7. Let the iteration operator \( B \) have the form
\[
B = \begin{bmatrix}
\lambda_1 & 1 & 0 & 0 & 0 \\
0 & \lambda_1 & 1 & 0 & 0 \\
0 & 0 & \lambda_1 & 0 & 0 \\
0 & 0 & 0 & \lambda_2 & 0 \\
0 & 0 & 0 & 0 & \lambda_3
\end{bmatrix}, \quad 1 > \lambda_1 > \lambda_2 > \lambda_3.
\]
Prove the asymptotic behavior of the iteration is proportional to \( \lambda_1 \). How many iterations are necessary to achieve the asymptotic behavior? Generalize the result.

8. It is desired to solve the Helmholtz equation
\[
\delta^2 \phi + k^2 \phi = 0, \quad (h_z = 1)
\]
with \( \phi(0) = \phi(4) = 0 \). The mesh is shown below.

(a) Find the eigenvalues and eigenfunctions of the difference equation.
(b) Use the first eigenvalue from part (a) and the initial estimate \( \phi(1) = \phi(2) = \phi(3) = 1 \). Solve the problem by the method of simultaneous displacements and by successive displacements. Explain.
(c) Repeat using the second eigenvalue and explain.

9. Show that the method of simultaneous displacements converges, whereas the method of successive displacements diverges for the problem
\[
\begin{bmatrix}
1 & 1 & -1 & \vdots & x_1 \\
1 & 1 & -2 & \vdots & x_2 \\
0 & 1 & 1 & \vdots & x_3
\end{bmatrix} = \begin{bmatrix}
y_1 \\
y_2 \\
y_3
\end{bmatrix}.
\]

10. Derive Eq. (3.6.86) from the following considerations: the residual is defined as
\[
x_p = x^{p+1/2} - x^p
\]
and \( x^{p+1/2} \) is the solution computed without over-relaxation, i.e.,
\[
L x^{p+1} + D x^{p+1/2} + U x^p = y.
\]
By elimination of the intermediate solution, the desired algorithm is obtained.

11. Show that the difference approximation to the Laplace equation
\[
\frac{\delta^2 \phi_{j,k}}{h_x^2} + \frac{\delta^2 \phi_{j,k}}{h_y^2} = 0
\]
is unstable if the problem is treated as an initial value problem.
12. Consider a square mesh with \( h_x = h_y = h \) and the difference equation from problem 11. If \( \phi_{j,k} \) is changed by an amount \( r_{j,k} \), derive an expression for the influence of the change at point \( j + n, k + m \). Assume the iteration is the method of successive displacements and that \( j, k \) is relaxed before \( j + n, k + m \).

13. Derive an expression for the growth factors of the Peaceman-Rachford iteration method for the Helmholtz equation in a rectangle.

14. Prove that \( B \) in Eq. (3.6.176) is nonnegative. Also: Examine the behavior of the spectral radius of \( B \) as a function of \( \lambda \) in place of \( \lambda_{\text{min}} \) and apply the Perron-Frobenius theorem of Chapter I.