THE NUMERICAL SOLUTION
OF PARTIAL DIFFERENTIAL EQUATIONS

John Gary

National Center for Atmospheric Research*

Boulder, Colorado 80302

---

*The National Center for Atmospheric Research is sponsored by the National Science Foundation.

NCAR Manuscript No. 69–54

THE NUMERICAL SOLUTION
OF PARTIAL DIFFERENTIAL EQUATIONS

John Gary

National Center for Atmospheric Research*
Boulder, Colorado 80302

April 1969

*The National Center for Atmospheric Research is sponsored by the National Science Foundation.

NCAR Manuscript No. 69-54

TABLE OF CONTENTS

1. MATHEMATICAL BACKGROUND

The tools required to undertake the numerical solution of partial differential equations include a reasonably good knowledge of the calculus and some facts from the theory of partial differential equations. Also, the reader should have some knowledge of matrix theory. A good reference for the analysis is "Advanced Calculus" by Kaplan, and for matrix theory the reader might try "Linear Algebra and Matrix Theory" by Nering. Of course, there are many other suitable references. In this first chapter, we will review some of the concepts we will need for the remaining chapters. We will assume some familiarity with advanced calculus, including limits, uniform convergence, continuity, etc.

1.1 A Few Analytical Tools

1.1.1 Taylor series expansions for functions of one and two variables.

We will first consider a Taylor series expansion with remainder for the function $f(x)$. We assume the derivatives of f d^kf/dx^k are continuous in the interval $a-r < x < a+r$ for $0 \le k \le n+1$ $(r > 0)$. Then for each x in this interval there is at least one point x_1 contained in the interval from a to x $(a < x_1 < x$ or $x < x_1 < a)$ such that

$$f(x) = f(a) + (x-a)f^{(1)}(a) + \frac{(x-a)^2}{2} f^{(2)}(a) + \dots + \frac{(x-a)^n}{n!} f^{(n)}(a)$$

$$+ \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(x_1)$$

Here $f^{(k)}(x) = \frac{d^kf}{dx^k}$ denotes the k^{th} derivative of $f(x)$. For further details see any advanced calculus text, in particular page 357 in the book by Kaplan. A similar formula exists for functions of several variables. For example,

$$f(x,y) = f(a,b) + (x-a)f_x + (y-b)f_y + \frac{1}{2!}\Big[(x-a)^2 f_{xx} + 2(x-a)(y-b)f_{xy}$$

$$+ (y-b)^2 f_{yy}\Big] + \frac{1}{3!}\Big[(x-a)^3 f^*_{x^3} + 3(x-a)^2(y-b)f^*_{x^2y}$$

$$+ 3(x-a)(y-b)^2 f^*_{xy^2} + (y-b)^3 f^*_{y^3}\Big]$$

where $f_x = \frac{\partial f}{\partial x}(a,b)$, $f_{xy} = \frac{\partial^2 f}{\partial x \partial y}(a,b)$, $f^*_{x^3} = \frac{\partial^3 f}{\partial x^3}(x^*,y^*)$, etc.

$$x^* = a + \tau(x-a), \quad y^* = b + \tau(y-b) \qquad \text{with } 0 < \tau < 1.$$

1.1.2 Polynomial interpolation. To construct finite difference approximations to derivatives, we will usually use polynomial interpolation. This is generally a trivial matter and does not require even as much theory as we will describe in this section. Usually we can obtain these difference approximations from a Taylor series expansion. We will first illustrate interpolation problems by a few simple examples.

Suppose we have a function $f(x)$ and we happen to know the values of this function $f(x_j)$ for the points x_j, $0 \le j \le 2$, $x_0 < x_1 < x_2$. Then we could approximate this function by the second degree polynomial $P_2(x) = a_2x^2 + a_1x + a_0$ which agrees with $f(x)$ at these points; that is, $P_2(x_j) = f(x_j)$ $0 \le j \le 2$. A simple way to write this polynomial in terms of the points x_j and the values $f_j = f(x_j)$ is the following Lagrangian form of the interpolation polynomial:

$$P_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f_2 \qquad (1.1\text{-}1)$$

The reader should inspect this formula and note that the expression is a second degree polynomial and furthermore $P_2(x_j) = f_j = f(x_j)$, for $0 \le j \le 2$.

Problem 1.1-1. Suppose $x_0 = -h$, $x_1 = 0$, $x_2 = h$. Evaluate the coefficients a_2, a_1, a_0 in terms of the values f_0, f_1, f_2. Given 4 points $x_0 < x_1 < x_2 < x_3$, write out the Lagrangian formula for the interpolation polynomial. Now generalize this to a polynomial $P_n(x) = a_nx^n + a_{n-1}x^{n-1} + \ldots + a_0$. You may wish to use the product notation

$$\prod_{i=0}^{2} (x-x_i) = (x-x_0)(x-x_1)(x-x_2)$$

and

$$\prod_{\substack{i=0\\ i\neq 0}}^{2} (x-x_i) = (x-x_1)(x-x_2)$$

We could obtain the coefficients a_2, a_1, a_0 in the above example directly from the conditions $P_2(x_j) = f_j$ for $0 \leq j \leq 2$. From these conditions we obtain three equations for the three unknowns a_2, a_1, a_0.

$$a_0 + a_1x_0 + a_2x_0^2 = f_0$$

$$a_0 + a_1x_1 + a_2x_1^2 = f_1$$

$$a_0 + a_1x^2 + a_2x_2^2 = f_2$$

Problem 1.1-2. Show that these three equations will have a unique solution provided $x_0 < x_1 < x_2$. (You might show that the determinant of the matrix is non-zero.) Solve these equations for $x_0 = -h$, $x_1 = 0$, $x_2 = h$.

Next we will consider a related interpolation problem. Suppose we wish to find an approximation to the second derivative of $f(x)$ at $x = x_1$ using the values $f(x_j)$ at three points x_0, x_1, x_2. We could do this by differentiating the interpolating polynomial $P_2(x)$ which we have written down in equation (1.1-1).

Problem 1.1-3. Differentiate equation (1.1-1) to obtain an approximation for $f^{(2)}(x_1)$. If $x_0 = -h$, $x_1 = 0$, $x_2 = h$, show this yields

$$f^{(2)}(x_1) \simeq (f_0 - 2f_1 + f_2)/h^2.$$

This formula can be obtained with less effort by use of the Taylor series for f(x). This approach will also yield an estimate for the error. We assume $x_0 = -h$, $x_1 = 0$, $x_2 = h$. From the Taylor series we have

$$f(x_0) = f(x_1) + (x_0-x_1)f^{(1)}(x_1) + \frac{(x_0-x_1)^2}{2} f^{(2)}(x_1)$$

$$+ \frac{(x_0-x_1)^3}{3!} f^{(3)}(x_1) + \frac{(x_0-x_1)^2}{4!} f^{(4)}(\xi_0)$$

where $x_0 < \xi_0 < x_1$. A similar expansion holds for $f(x_2)$. We thus obtain

$$f_0 = f_1 - hf_1^{(1)} + \frac{h^2}{2} f_1^{(2)} - \frac{h^3}{6} f_1^{(3)} + \frac{h^4}{24} f_{\xi_0}^{(4)}$$

$$f_2 = f_1 + hf_1^{(1)} + \frac{h^2}{2} f_1^{(2)} + \frac{h^3}{6} f_1^{(3)} + \frac{h^4}{24} f_{\xi_2}^{(4)}$$

If we add these equations we can obtain an expression for the second derivative.

$$f_1^{(2)} = \frac{(f_0 - 2f_1 + f_2)}{h^2} - \frac{h^2}{24}\left(f_{\xi_0}^{(4)} + f_{\xi_2}^{(4)}\right)$$

This expression is exact, but it contains the error term $-h^2 f_\xi^{(4)}/12$. Normally, we will not know the value of the fourth derivative, but the knowledge that the error term contains the factor h^2 is of great value.

We can also use this interpolating polynomial to approximate the integral of f(x).

Problem 1.1-4. Derive Simpson's quadrature formula. Let $x_0 = -h$, $x_1 = 0$, $x_2 = h$, then integrate $P_2(x)$ from equation (1.1-1) to obtain the approximation

$$\int_{-h}^{h} f(x)dx \cong \frac{1}{3h}\left[f_0 + 4f_1 + f_2\right]$$

1.1.3 The method of undetermined coefficients for construction of the interpolating polynomial. We can look at this method of polynomial approximation in a less direct way. This will frequently lead to an easier derivation of the formula for an approximate derivative or integral. If we construct an interpolation polynomial $P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \dots a_0$ for a function $f(x)$ based on the points $x_0, \dots x_n$, then the coefficients $a_k = a_k(f_0, \dots f_n, x_0, \dots x_n)$ are functions of the points x_j and the values f_j. These functions produce an exact fit in case the function $f(x)$ is a polynomial of degree m where $m \le n$. That is, the polynomial $P_n(x)$ is identically equal to the polynomial $f(x)$. We will not bother to prove this statement, although the proof is not difficult. The proof is based on the fact that if a polynomial of degree m is zero at n+1 points where $m \le n$, then the polynomial must be identically zero; that is, all of its coefficients are zero. We can use this fact to derive formulas based on interpolation. For example, suppose we wish to approximate the second derivative using a second degree interpolation formula (Problem 1.1-3). From the form of the interpolating polynomial (equation (1.1-1) we know our approximation will be linear in the values f_0, f_1, f_2. That is, it has the form

$$f_1^{(2)} \cong b_0 f_0 + b_1 f_1 + b_2 f_2 \tag{1.1-2}$$

We must determine the unknown coefficients b_0, b_1, b_2. We assume the points x_j are $x_0 = -h$, $x_1 = 0$, $x_2 = h$. Note that we would get the same formula for any set of equally spaced points $x_0 = x_1 - h$, $x_2 = x_1 + h$. We know our formula (1.1-2) should be exact if $f(x)$ is a polynomial of degree less than three. In particular, it is exact for $f(x) = 1$, $f(x) = x$, and $f(x) = x^2$. Therefore we have the following three equations obtained by substitution into equation (1.1-2) which is now exact for these functions $f(x)$

$$\begin{aligned} b_0 + b_1 + b_2 &= 0 \\ -b_0 h \qquad + b_2 h &= 0 \\ b_0 h^2 \qquad + b_2 h^2 &= 2 \end{aligned}$$

If we solve this set of equations, we obtain $b_0 = b_2 = 1/h^2$, $b_1 = -2/h^2$ which yields the same formula as before.

We can use the same method to obtain Simpson's quadrature formula. We let $f_0 = f(-h)$, $f = f(0)$, and $f_2 = f(h)$. We assume the following form for our quadrature formula

$$\int_{-h}^{h} f(x)dx \cong b_0 f_0 + b_1 f_1 + b_2 f_2$$

If we require this formula to be exact for the three functions $f(x) = 1$, $f(x) = x$, and $f(x) = x^2$, we then obtain the following three equations:

$$\begin{aligned} b_0 + b_1 + b_2 &= 2h \\ -hb_0 \qquad + hb_2 &= 0 \\ h^2 b_0 \qquad + h^2 b_2 &= \frac{2h^3}{3} \end{aligned}$$

The solution of this system of equations is $b_0 = h/3$, $b_1 = 4h/3$, $b_2 = h/3$ which yields Simpson's rule.

Problem 1.1-5. Find a three-point interpolation formula for the first derivative (a one-sided difference approximation). That is, determine the b_j in the approximation

$$f^{(1)}(x_0) = b_0 f(x_0) + b_1 f(x_0+h) + b_2 f(x_0+2h)$$

Problem 1.1-6. Find a five-point interpolation formula for the second derivative $f^{(2)}(x)$; that is, determine the constants b_j in the approximation

$$f^{(2)}(x_0) \cong b_0 f(x_0-2h) + b_1 f(x_0-h) + b_2 f(x_0) + b_3 f(x_0+h) + b_4 f(x_0+2h)$$

Problem 1.1-7. Find a four-point quadrature formula; that is, find the b_j in the approximation

$$\int_0^h f(x)dx = b_0 f(-h) + b_1 f(0) + b_2 f(h) + b_3 f(2h)$$

We can derive error estimates for these formulas based on a Taylor series expansion. First we will consider Simpson's quadrature formula.

$$\int_{-h}^{h} f(x)dx \cong \frac{h}{3}\left[f(-h) + 4f(0) + f(h)\right]$$

The Taylor series expansion out to fourth order is (we assume $f^{(4)}(x)$ is continuous)

$$f(x) = f(0) + xf^{(1)}(0) + \frac{x^2}{2} f^{(2)}(0) + \frac{x^3}{6} f^{(3)}(0) + \frac{x^4}{24} f^{(4)}(\xi)$$

where $|\xi| < |x|$. We let $Q_2(x)$ denote the polynomial $Q_2(x) = f_0 + xf_0^{(1)} + x^2 f_0^{(2)}/2$ where $f_0 = f(0)$. We know our quadrature formula is exact for polynomials of degree no greater than two. We derived it in such a way that it is exact for the polynomials $f(x) = 1$, $f(x) = x$, and $f(x) = x^2$. Therefore, it is exact for any linear combination of these polynomials, hence for any second degree polynomial. For each x, we can choose $\xi = \xi(x)$ such that

$$f(x) = Q_2(x) + \frac{f_0^{(3)}}{6} x^3 + \frac{x^4}{24} f^{(4)}(\xi(x)) \qquad (1.1\text{-}3)$$

Since $f^{(4)}(\xi(x)) = 24[f(x) - Q_2(x) - x^3 f_0^{(3)}/6]/x^4$, it is clear that $f^{(4)}(\xi(x))$ is a continuous function of x for $x > 0$. We will now assume that this fourth derivative is bounded; that is, $|f^{(4)}(\xi)| \leq M$ if $|\xi| < h$. By integration of equation (1.1-3) we obtain

(note that $\int_{-h}^{h} x^3 dx = 0$)

$$\int_{-h}^{h} f(x)dx = \int_{-h}^{h} Q_2(x) + \int_{-h}^{h} e(x)dx \qquad \text{where } e(x) = \frac{x^4}{24} f^{(4)}(\xi(x)) \qquad (1.1\text{-}4)$$

Since our quadrature formula is exact for second degree polynomials we have

$$\int_{-h}^{h} Q_2(x)dx = \frac{h}{3}\left[Q_2(-h) + 4Q_2(0) + Q_2(h)\right]$$

Also $f(x) = Q_2(x) + x^3 f_0^{(3)}/6 + e(x)$. Therefore, if we use this relation to evaluate f(h), f(0), f(-h), and use the above relation for

$$\int_{-h}^{h} Q_2(x)dx$$

$$\frac{h}{3}[f(-h) + 4f(0) + f(h)] = \int_{-h}^{h} Q_2(x)dx + \frac{h}{3}[e(-h) + 4e(0) + e(h)]$$

Now by substitution into equations (1.1-4) we obtain

$$\int_{-h}^{h} f(x)dx = \frac{h}{3}[f(-h) + 4f(0) + f(h)] - \frac{h}{3}[e(-h) + 4e(0) + e(h)] + \int_{-h}^{h} e(x)dx$$

The error $E = \int_{-h}^{h} e(x)dx - \frac{h}{3}[e(-h) + 4e(0) + e(h)]$ can now be bounded. We have $|e(x)| \le |x|^4 M/24$ and therefore

$$\left|\int_{-h}^{h} e(x)dx\right| \le \int_{-h}^{h} |e(x)|dx \le \frac{M}{24}\int_{-h}^{h} |x^4|dx = \frac{2M}{24}\int_{0}^{h} x^4 dx = \frac{Mh^5}{60}$$

If we note that $e(0) = 0$ and $|x| < h$, we obtain

$$(h/3)|e(-h) + 4e(0) + e(h)| \le \frac{h}{3}\left(|e(-h)| + |e(h)|\right) \le \frac{h}{3}\left(\frac{2h^4 M}{24}\right) = h^5 M/36$$

Therefore our final bound is $E \le \frac{Mh^5}{60} + \frac{Mh^5}{36} = 4Mh^5/90$. If one is willing to work harder, a better bound can be obtained, namely $E \le h^5 M/90$. See the book by Isaacson and Keller for a derivation of this error bound.

Problem 1.1-8. Show that the five-point interpolation formula for the first derivative is

$$f^{(1)}(x_0) \cong \frac{h}{12}\left[f(x_0-2h) - 8f(x_0-h) + 8f(x_0+h) - f(x_0+2h)\right]$$

Obtain an error estimate for this formula by the method used above.

1.1.4 Fourier series. Next we will consider some properties of Fourier series. We will consider functions $f(x)$ for $-1 \le x \le 1$. If we have a function defined on the interval $A \le x \le B$, then we can use the transformation $y = -1 + 2(x-A)/(B-A)$ to reduce the problem to the interval $-1 \le y \le 1$. We will assume that $f(x)$ is a complex valued function of the real variable x, $f(x) = f_r(x) + if_i(x)$, where $f_r(x)$ and $f_i(x)$ are real (of course we may have $f(x)$ real; that is, $f_i(x) \equiv 0$). We will look for a Fourier series representation of $f(x)$; that is

$$f(x) = \sum_{k=-\infty}^{\infty} a_k e^{ik\pi x}$$

The above statement can be written

$$f(x) = \lim_{K\to\infty} \sum_{k=-K}^{k=K} a_k e^{ik\pi x} \qquad -1 \le x \le 1. \qquad (1.1\text{-}5)$$

We need to compute the coefficient a_k in terms of $f(x)$ so that this expression holds for reasonable functions $f(x)$. Note that the following formula holds for all integers m and k.

$$\int_{-1}^{1} e^{-im\pi x} e^{ik\pi x} = \begin{cases} 2 & \text{if } m = k \\ 0 & \text{if } m \ne k \end{cases}$$

Suppose the above series can be integrated termwise, then

$$\tfrac{1}{2}\int_{-1}^{1} f(x)e^{-im\pi x}\,dx = \tfrac{1}{2}\int_{-1}^{1} \sum_{-\infty}^{\infty} a_k\, e^{i(k-m)\pi x} = \tfrac{1}{2}\sum_{-\infty}^{\infty}\int_{-1}^{1} a_k\, e^{i(k-m)\pi x} = a_m$$

or $$a_m = \tfrac{1}{2}\int_{-1}^{1} f(x)\, e^{-im\pi x} \qquad (1.1\text{-}6)$$

If the series is uniformly convergent, then the termwise integration (that is, $\int_{-1}^{1} \sum_{-\infty}^{\infty} = \sum_{-\infty}^{\infty} \int_{-1}^{1}$) is permissible. Thus we have an expression for the coefficients a_m. The coefficients a_m are called Fourier coefficients. Suppose we are given a function f(x) such that the above Fourier coefficients exist; that is, the integrals 1.1-6 exist. Under what conditions does the Fourier series converge to f(x); when does equation (1.1-5) hold? The Fourier series will converge if the derivative $f'(x)$ is piecewise continuous. Weaker conditions are also sufficient, but we will not need them. We say f(x) is piecewise continuous on [-1,1] if there are a finite set of points ξ_j, $1 \le j \le n$, $-1 \le \xi_1 < \xi_2 < \ldots < \xi_h \le 1$, such that f(x) is continuous except at the points ξ_j and the following limits exist:

$$\lim_{\substack{h\to 0\\ h>0}} f(\xi_j+h) = f^+(\xi_j) \qquad \text{if } \xi_j < 1$$

$$\lim_{\substack{h\to 0\\ h>0}} f(\xi_j-h) = f^-(\xi_j) \qquad \text{if } \xi_j > -1$$

If $f'(x)$ is piecewise continuous, then the Fourier series converges to f(x) except at the points ξ_j where the series converges to $(f^+(\xi_j) + f^-(\xi_j))/2$.

There are several properties of Fourier series which we will note, although we may not use all of them. If f(x) is a real valued function, then the Fourier coefficients satisfy the condition

$$a_{-k} = \bar{a}_k$$

where $\bar{a}_k$ denotes the complex conjugate of a_k.

If f(x) is continuous and periodic and f′(x) is piecewise continuous, then

$$a_k = 0\left(\frac{1}{k^2}\right)$$

By periodic we mean f(1) = f(-1). We say a_k is of order $1/k^2$, written $a_k = 0(1/k^2)$, if there is a constant M such that $|a_k| \le M/k^2$ for all non-zero integers k. This can be generalized. If f(x), $f^{(1)}(x), \ldots, f^{(S-1)}(x)$ are continuous and periodic, and $f^{(S)}(x)$ piecewise continuous, then $a_k = 0\left(\frac{1}{k^{S+1}}\right)$. This statement helps to decide how fast a Fourier series converges; that is, the rate at which the coefficients a_k approach zero. Thus if f(x) = x for $-1 \le x \le 1$, the Fourier coefficients satisfy $a_k = 0(1/k)$. If $f(x) = |x|$, then $a_k = 0(1/k^2)$.

Problem 1.1-9. Prove the statements in the above paragraph.

The Parseval relation

$$\tfrac{1}{2}\int_{-1}^{1} f^2(x)\,dx = \sum_{-\infty}^{\infty} |a_k|^2$$

relates the L_2 norm of f to the Fourier coefficients. We will refer the reader to a text on Fourier series for the proof.

In certain cases the complex form of the Fourier expansion can be simplified. For example, if f(x) is real valued, then

$$f(x) = \tfrac{1}{2} A_0 + \sum_{k=1}^{\infty} A_k \cos k\pi x + B_k \sin k\pi x$$

where

$$A_k = \int_{-1}^{1} f(x) \cos k\pi x dx, \qquad B_k = \int_{-1}^{1} f(x) \sin k\pi x dx$$

If f(x) is a real valued odd function, $f(-x) = -f(x)$, then

$$f(x) = \sum_{k=1}^{\infty} C_k \sin k\pi x$$

$$C_k = 2 \int_0^1 f(x) \sin k\pi x dx$$

If f(x) is a real valued even function, $f(-x) = f(x)$, then

$$f(x) = \tfrac{1}{2} C_0 + \sum_{k=1}^{\infty} C_k \cos k\pi x$$

$$C_k = 2 \int_0^1 f(x) \cos k\pi x dx$$

We might remind the reader of the following fact which we will sometimes use. If f(x) is defined by a convergent series $f(x) = \sum_{k=1}^{\infty} a_k(x)$, the derivatives $a_k'(x)$ are continuous, and the derived series $\sum_{k=1}^{\infty} a_k'(x)$ is uniformly convergent, then $f'(x) = \sum_1^{\infty} a_k'(x)$. Also, a uniformly convergent series whose terms are continuous can be integrated termwise. If $f(x) = \sum_1^{\infty} a_k(x)$ uniformly, then

$$\int f(x)dx = \int \left[\sum_1^{\infty} a_k(x) \right] dx = \sum_1^{\infty} \int a_k(x) dx$$

1.2 Vectors and Matrices

We will review some aspects of matrix theory. We will give the definitions for the general n-dimensional case, but most of the explanation will be for n=3.

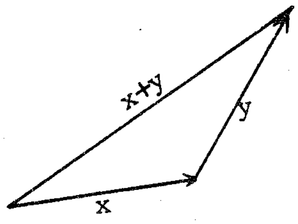
1.2.1 Some fundamental definitions.

The vector space E_n (Euclidean n-space) is the set of all ordered n-tuples $(x_1,x_2,...,x_n)$ of real numbers. The space C_n is the same except we use complex numbers. By a scaler we mean a real number if we are working with E_n and a complex number if we are in C_n. We use ordered n-tuples to insure that the vector (1,2,3) is not the same as (2,1,3). In the case of n=3 we may regard the three numbers (x_1,x_2,x_3) as the Cartesian coordinates of a point (x,y,z) in space. We may also think of a vector as a directed line segement from the origin to the point (x,y,z). Much of the intuition and nomenclature for vector spaces derives from the familiar 3-dimensional case. We define the sum of two vectors and the product of a vector by a scaler by the following relations:

$$x+y = (x_1+y_1,\ x_2+y_2,\ \ldots,\ x_n+y_n)$$

$$\alpha x = (\alpha x_1,\ \alpha x_2,\ \ldots,\ \alpha x_n)$$

We simply perform the operations on the components. In three dimensions, addition is the familiar parallelogram rule as shown below. Multiplication by a scaler may change the length of the vector and possibly reverse its direction.



We define the scaler product of two vectors in the real case by

$$x \cdot y = \sum_{j=1}^{n} x_j \, y_j$$

and in the complex case by

$$x \cdot y = \sum_{j=1}^{n} x_j \, \overline{y}_j$$

where $\overline{y}_j$ is the complex conjugate of y_j. Note that $x \cdot y \neq y \cdot x$ in the complex case. The Euclidean norm (or length) of a vector is defined by

$$|x| = \sqrt{x \cdot x}$$

Note that $x \cdot x \geq 0$ for all x, and thus $|x|$ is real and is taken non-negative. We summarize some important properties of these operations below. The reader may wish to prove these relations. Here x and y are vectors (in E_n or C_n) and α and β are scalers (real or complex).

1) $|x| > 0$ unless $x = 0$ ($x = 0$ if $x = (0,0,\ldots,0)$)

2) $|\alpha x| = |\alpha| \, |x|$

3) $|x+y| \leq |x| + |y|$

We say two vectors are orthogonal if $x \cdot y = 0$. The reader can verify that this definition agrees with the usual one for E_2 or E_3.

We will frequently use the unit vectors defined by

$$e_j^k = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k . \end{cases}$$

In E_3 we have $e^1 = (1,0,0)$, $e^2 = (0,1,0)$, $e^3 = (0,0,1)$.

If a vector w is given by $w = \alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_m x^m$ where the x^j are vectors and the α_i scalers, then we say w is a linear combination of the vectors x^j. We say a set of vectors x^j, $1 \le j \le m$ is linearly independent if there is no nontrivial linear relation among the vectors. In other words, if $\alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_m x^m = 0$, then $\alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$.

Problem 1.2-1. If the set of nonzero vectors x^j are orthogonal, then show they are linearly independent. By orthogonal we mean

$$x^j \cdot x^k = \begin{cases} 0 & j \neq k \\ \neq 0 & j = k \end{cases}$$

If a set $\{x^j\}$ of vectors is linearly independent, then it is possible to form an orthogonal set $\{y^j\}$ by using linear combinations of the x^j. This process is called the Gram-Schmidt orthogonalization. We let $y^1 = x^1$ and $y^2 = x^2 - \frac{(x^2 \cdot y^1) y^1}{y^1 \cdot y^1}$. Clearly $y^1 \cdot y^2 = 0$. If $y^2 = 0$, then x^1 and x^2 are not linearly independent. Therefore $y^2 \neq 0$. We now define y^3 by

$$y^3 = x^3 - \frac{(x^3 \cdot y^1)y^1}{(y^1 \cdot y^1)} - \frac{(x^3 \cdot y^2)y^2}{(y^2 \cdot y^2)}$$

Clearly $y^1 \cdot y^3 = y^2 \cdot y^3 = 0$ since $y^1 \cdot y^2 = 0$. If $y^3 = 0$, then x^1, x^2, and x^3 would be dependent since the equation

$$x^3 - \frac{(x^3 \cdot y^1)x^1}{(y^1 \cdot y^1)} - \frac{(x^3 \cdot y^2)}{(y^2 \cdot y^2)} \left(x^2 - \frac{(x^2 \cdot y^1)x^1}{(y^1 \cdot y^1)} \right) = 0$$

is a nontrivial relation among x^1, x^2, and x^3. Therefore $y^3 \neq 0$ and we can continue this process to produce an orthogonal set y^j.

Problem 1.2-2. Show that if the vectors y^j in the Gram-Schmidt process do not vanish, then the original vectors x^j are linearly independent.

Problem 1.2-3. Are the vectors $x^1 = (1,2,3,0)$ and $x^2 = (2,1,0,1)$ linearly independent? What about the set x^1,x^2,x^3 where $x^3 = (1,0,3,5)$? Use the Gram-Schmidt process and show that the y^j do not vanish.

Problem 1.2-4. Show that the unit vectors e^j are linearly independent.

We say a set of vectors $\{v^1,v^2,\ldots,v^m\}$ spans the space E_n if any vector x in E_n can be written as a linear combination of the v^j; that is, $x = \alpha_1 v^1 + \alpha_2 v^2 + \ldots + \alpha_m v^m$. A linearly independent set of vectors v^j which spans E_n is called a basis for E_n. Clearly the set e^j, $1 \leq j \leq n$ is a basis for E_n since any x can be represented by $x = x_1 e^1 + x_2 e^2 + \ldots + x_n e^n$.

Problem 1.2-5. Show that if $\{v^1,\ldots,v^m\}$ and $\{w^1,\ldots,w^p\}$ are bases for E_n, then m=p. Since $\{e^1,\ldots,e^n\}$ is a basis, we must have m=p=n.

Hint: Suppose $p > m$. The set $\{v^1,\ldots,v^m\}$ spans the space, therefore so does the set $\{w^1,v^1,\ldots,v^m\}$. We have $w_1 = \alpha_1 v^1 + \ldots + \alpha_m v^m$. At least one of the scalers α_i is not zero. We will assume without loss of generality that it is α_1. Then show $\{w^1,v^2,\ldots,v^m\}$ span the space. Now consider $\{w^1,w^2,v^2,\ldots,v^m\}$ and continue the argument to show that $\{w^1,w^2,\ldots,w^m\}$ span the space. But this contradicts $p > m$, since w^{m+1} is then a linear combination of the set $\{w^1,\ldots,w^m\}$.

Problem 1.2-6. Show that if $\{v^1,\ldots,v^n\}$ is an orthogonal basis, then for any x we have $x = \alpha_1 v^1 + \ldots + \alpha_n v^n$ where $\alpha_k = (x \cdot v^k)/(v^k \cdot v^k)$. If $v^k \cdot v^k = 1$, then show $x \cdot x = \alpha_1 \bar{\alpha}_1 + \ldots + \alpha_n \bar{\alpha}_n$. An orthonormal basis is an orthogonal basis for which $(v^j \cdot v^j) = 1$. For an orthonormal basis the above formula is simplified

$$x = (x \cdot v^1)v^1 + \ldots + (x \cdot v^n)v^n \ .$$

Problem 1.2-7. Show that any independent set of vectors $\{v^1,\ldots,v^m\}$ can be augmented by $\{w^1,\ldots,w^k\}$ so that $\{v^1,\ldots v^m,w^1,\ldots w^k\}$ is a basis for E_n. Show that no linearly independent set of vectors in E_n can contain more than n vectors. Also show that if we have n independent vectors $\{v^1,\ldots,v^n\}$, then any vector can be written as a linear combination of v^j; $x = \sum_{j=1}^{n} \alpha_j v^j$ for all x.

1.2.2 Linear transformations and matrices. Now we are ready to review the concept of a linear transformation. Let T be a transformation of E_n into itself; that is, for each vector x in E_n, T defines another vector T(x) in E_n. We say the transformation T is linear if $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$ for all vectors x and y and all scalers α and β.

Problem 1.2-8. Consider transformations T defined on E_3. Which of them are linear?

1) $T(x) = (x_3, 10x_2, 0)$

2) $T(x) = (x_1, x_1+x_2, |x_3|)$

3) $T(x) = (x_2, x_1, x_1^2)$

4) $T(x) = (x_3, x_3+x_1, x_3+x_2)$

5) $T(x) = (\sin x_1, 0, 0)$

Problem 1.2-9. Define a linear transformation T on E_3 such that $\{T(e^1), T(e^2), T(e^3)\}$ do not span E_3, but do span a 2-dimensional subspace (we omit the definition of a subspace).

We define a matrix to be a rectangular array of scalers; (a_{ij}), $1 \le i \le n$, $1 \le j \le m$. That is, an n×m matrix. An n×n matrix is said to be of order n. We usually write out the matrix so that the first index, i, denotes the rows.

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

A matrix $A = (a_{ij})$ of order n defines a linear transformation by the following rule. Given any vector x in E_n, then define y by $y_i = \sum_{j=1}^{n} a_{ij}x_j$, $1 \le i \le n$. We use the notation $y = Ax$. Conversely, given any linear transformation T on E_n, there is a matrix A of order n such that $T(x) = Ax$ for all x in E_n.

Problem 1.2-10. If T is a linear transformation on E_n, denote the vector $T(e^j)$ by $T(e^j) = (a_{1j}, a_{2j}, \ldots, a_{nj})$, $1 \le i \le n$. Let the matrix A be defined by these elements, $A = (a_{ij})$. Show $T(x) = Ax$ for all x.

Thus we see that a linear transformation and a matrix are essentially the same thing. We have used the basis $\{e^1, \ldots, e^n\}$ to define the relationship between a linear transformation and a matrix. We could have used another basis which would have produced a different relationship. We leave this point to the texts on linear algebra.

Now we will state some definitions.

The sum of two $(n \times m)$ matrices A and B is an $(n \times m)$ matrix C defined by $C = A+B$, $C_{ij} = a_{ij} + b_{ij}$ for $1 \le i \le n$, $1 \le j \le m$.

The product of a scaler and a matrix is $B = \alpha A$, $b_{ij} = \alpha a_{ij}$.

The transpose of an $(n \times m)$ matrix A is the $(m \times n)$ matrix B denoted by A^T, where $b_{ij} = a_{ji}$. The conjugate transpose for an $(n \times m)$ complex matrix A is an $(m \times n)$ matrix B denoted by A^* where $b_{ij} = \bar{a}_{ji}$.

The identity matrix of order n is denoted by I and defined by

$$I_{ij} = \begin{Bmatrix} 0 & i \neq j \\ 1 & i = j \end{Bmatrix}.$$

The product of an $(n \times s)$ matrix A and an $(s \times m)$ matrix B is an $(n \times m)$ matrix C defined by

$$c_{ij} = \sum_{k=1}^{s} a_{ik} b_{kj}$$

We can regard a vector x as a column vector (an $(n \times 1)$ matrix) or as a row vector (a $(1 \times n)$ matrix). That is

$$x = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{vmatrix} \qquad \text{or } x = (x_1, x_2, \ldots, x_n)$$

If we regard x as a column vector, then the linear transformation defined by the matrix A is obtained from the matrix product Ax.

Problem 1.2-11. Let A and B be two matrices of order n. Let T be the composite linear transformation defined by $T(x) = T_A(T_B(x))$ where $T_A(x) = Ax$ and $T_B(x) = Bx$. Show that the matrix C corresponding to T is the product matrix $C = AB$.

Problem 1.2-12. Give an example to show that the matrix product is not commutative. Find two matrices of order 2 such that $AB \neq BA$.

Problem 1.2-13. Given vectors x and y, show that $|x \cdot y| \leq |x||y|$. Hint: First assume x and y are real. Then $(x+\lambda y) \cdot (x+\lambda y) = |x+\lambda y|^2 = |x|^2 + 2\lambda x \cdot y + \lambda^2 |y|^2 \geq 0$ for all real λ. Show this quadratic in λ can have at most one real root. Then show $|x \cdot y|^2 \leq |x|^2 |y|^2$. The complex case can be reduced to real case since $|\sum_{i=1}^{n} x_i \bar{y}_i| \leq \sum_{i=1}^{n} |x_i||y_i|$.

Problem 1.2-14. Let A be a matrix of order n. Let α be the maximum of the lengths of the rows of A; that is,

$$\alpha = \underset{1 \leq i \leq n}{\text{Max}} \sqrt{\sum_{j=1}^{n} |a_{ij}|^2} .$$

If $y = Ax$, then show $|y| \leq \sqrt{n}\alpha |x|$. Hint: Note that the components of y are given by the inner product of the rows of A with x. Then use problem 1.2-13.

We say a matrix of order n is singular if there is a nonzero vector x such that $Ax = 0$, otherwise A is said to be nonsingular.

Problem 1.2-15. Show that if a matrix A of order n is nonsingular, then for any $y \in E_n$, there is a unique vector x such that $y = Ax$. Hint: if the vectors Ae^j, $1 \leq j \leq n$ are not independent, then $\sum_{i=1}^{n} \alpha_i A e^i = 0$ for some α_i. Now use the linearity of A and the result of problem 1.2-7.

For a nonsingular matrix A there is for each y a unique x such that $y = Ax$. We can define a transformation T by $T(y) = x$. We can show this transformation is linear. If $y^1 = Ax^1$ and $y^2 = Ax^2$, then $\alpha_1 y^1 + \alpha_2 y^2 = A(\alpha_1 x^1 + \alpha_2 x^2)$; therefore $T(\alpha_1 y^1 + \alpha_2 y^2) = \alpha_1 x^1 + \alpha_2 x^2 = \alpha_1 T(y^1) + \alpha_2 T(y^2)$. We denote the matrix of this transformation T by A^{-1}. It is called the inverse of A since $A^{-1}Ax = AA^{-1}x = x$ for all x. To summarize: for any nonsingular matrix A we define the inverse matrix A^{-1} to be the matrix such that $A^{-1}A = AA^{-1} = I$, where I is the identity matrix. We have just shown that such a matrix exists. It is easy to show that there is only one such matrix for the given matrix A. Note that if a matrix has an inverse it must be nonsingular.

We define an orthogonal matrix to be a real matrix whose inverse is equal to its transpose; that is, $A^T A = AA^T = I$. We define a unitary matrix to be a complex matrix whose conjugate transpose is equal to its inverse; that is, $A^* A = AA^* = I$.

Problem 1.2-16. Show that an orthogonal transformation preserves length. In other words, if A is an orthogonal matrix and $y = Ax$, then $|y| = |x|$. Hint: Show that $|y| = \sqrt{y^T y} = \sqrt{x^T A^T Ax}$. Show that the product of orthogonal matrices is orthogonal.

Problem 1.2-17. Suppose we are given rows r and s of any matrix A. Show that any element a_{sk}, $1 \le k \le n$, can be zeroed out by premultiplication by an orthogonal matrix U of the following form. The elements $u_{ii} = 1$ if $i \neq r$ or s, $u_{rr} = \cos\theta$, $u_{rs} = \sin\theta$, $u_{sr} = -\sin\theta$, $u_{ss} = \cos\theta$, $u_{ij} = 0$ otherwise. Use this to show there is an orthogonal matrix U such that $UA = T$ is upper triangular ($t_{ij} = 0$ if $i > j$).

1.2.3 The definition of the determinant. We will now define the determinant of a matrix of order n. First we need to define the set of permutations. We let N be the first n positive integers, $N = \{1,2,3,\ldots,n\}$. A permuation π of N is a one-to-one mapping of N onto itself; that is, a reordering of N. For example, if $n = 4$, then (2,1,4,3) defines the permutation $\pi(1) = 2$, $\pi(2) = 1$, $\pi(3) = 4$, $\pi(4) = 3$. We let $k(\pi)$ denote the number of pairs (i,j) of elements of N for which $i < j$ and $\pi(i) > \pi(j)$. Then π is said to be even or odd if $k(\pi)$ is even or odd. The permutation defined by (2,1,4,3) is even; that defined by (2,4,1,3) is odd. Note that there are n! permutations of N.

We define $\text{sgn}(\pi) = (-1)^{k(\pi)}$ to be +1 if π is even and -1 if π is odd. We define the determinant of A, denoted by $|A|$ or det(A), as the scaler evaluated by the formula below.

$$|A| = \sum_{\pi} \text{sgn}(\pi)\ a_{1\pi(1)}\ a_{2\pi(2)} \cdots a_{n\pi(n)}$$

Note that this is a sum of n! terms. In the case of a 2×2 matrix, there are only two permutations (1,2) and (2,1), thus

$$|A| = a_{11}\ a_{22} - a_{12}\ a_{21}$$

Problem 1.2-18. Using the above formula, write out the determinant for a matrix of order 3.

We will refer the reader to the book by Nering for proofs of the following statements. They follow rather easily from the definitions of a determinant.

A matrix A and its transpose have the same determinant, that is $|A| = |A^T|$.

If B is a matrix obtained from A by multiplication of a single row (or column) of A by a scaler α, then $|B| = \alpha|A|$.

If B is obtained from A by the interchange of two rows (or columns), then $|B| = -|A|$.

If two rows (or columns) of A are identical, then $|A| = 0$.

If B is obtained from A by adding a multiple of one row (or column) of A to a different row (or column), then $|B| = |A|$.

The determinant of the product is the product of the determinants. In other words, if A and B are matrices of order n, then $|AB| = |A||B|$. This statement is somewhat more difficult to prove than the previous ones.

To wind up our discussion of the determinant we will give an alternative method to compute it, again without proof. Given a matrix A of order n, we form a submatrix of order n-1 by deletion of the i^{th} row and the j^{th} column. We denote by A_{ij} the determinant of this submatrix times $(-1)^{i+j}$. The scaler A_{ij} is called the cofactor of a_{ij}. It is possible to show that

$$|A| = \sum_{j=1}^{n} a_{ij} A_{ij}$$

We can choose any row, that is any value of i, in this expansion. We can also expand on any column; that is,

$$|A| = \sum_{i=1}^{n} a_{ij} A_{ij}$$

As an example consider the expansion on the first row for a (3X3) matrix.

$$|A| = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

The above methods for evaluation of the determinant are used mainly for theoretical arguments. If we actually want the value of a determinant of order greater than 3, we would normally use Gauss elimination to obtain it.

1.2.4 Gauss elimination. Next we will describe the Gauss elimination method for solving a linear system. Suppose we have a (3×3) system of equations. In addition, suppose the system is upper triangular; that is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{33}x_3 &= b_3 \end{aligned}$$

Obviously, we can solve this system of equations by a "backward substitution", first solving for x_3, then x_2 and x_1

$$x_3 = b_3/a_{33}$$

$$x_2 = (b_2 - a_{23}x_3)/a_{22}$$

$$x_1 = (b_1 - a_{13}x_3 - a_{12}x_2)/a_{11}$$

Now suppose we consider a general 3×3 system. If we can reduce it to upper triangular form, then we can complete the solution by backward substitution. Consider the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

If we multiply the first equation by the appropriate factor m_{21} and add to the second, then multiply the first by m_{31} and add to the third, we will zero out the subdiagonal elements in the first column. The modified system is the following where $a_{ij}^{(1)} = a_{ij} + m_{i1} a_{1j}$, $b_i^{(1)} = b_i + m_{i1} b_1$, $m_{11} = 1$, $m_{i1} = -a_{i1}/a_{11}$ for $2 \leq i \leq 3$. Note that the modified system has the same solution as the original.

$$
\begin{aligned}
a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 &= b_1^{(1)} \\
a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 &= b_2^{(1)} \\
a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 &= b_3^{(1)}
\end{aligned}
$$

Using the same method we can zero out the subdiagonal element in the second column, namely $a_{32}^{(1)}$. This leaves us with an upper triangular matrix.

Before elimination of the subdiagonal elements in a given column, we normally interchange the rows of the system so that the multipliers m_{ik} will not exceed unity in magnitude. That is, working on the k^{th} column, we choose the integer s, $k \leq s \leq n$, so that $|a_{sk}^{(k-1)}| = \max_{k \leq i \leq n} |a_{ik}^{(k-1)}|$. Then we interchange the k^{th} and s^{th} rows of the system to form a new system. The multipliers $m_{ik} = -a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ will now have magnitude bounded by unity. If we attempt to solve the system below, then we must interchange rows, since a_{11} is zero and we cannot divide by it.

$$
\begin{vmatrix} 0 & 1 \\ 2 & 3 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} 1 \\ 2 \end{vmatrix}
$$

It should be clear that this procedure can be generalized to a matrix of arbitrary order n. If all the elements $a_{ik}^{(k-1)}$ for $k \leq i \leq n$ vanish, then the procedure will fail. However, in this case, the matrix is singular and we might expect trouble (see problem 1.2-19). If we wish to solve the system for two right-hand sides, then we simply perform the reductions on both vectors simultaneously. For example, suppose we wish to solve the system $Ax = b$ for the two vectors $b = (1,0)^T$ and $b = (0,1)^T$ with the same matrix

$$A = \begin{vmatrix} 0.1 & 1.2 \\ 1 & 2 \end{vmatrix} .$$

We can write this problem in the form $AX = B$ where X is a (2×2) matrix and $B = I$ is the identity matrix of order 2. Since $AX = I$, we see that the solution matrix X is the inverse of A, which we denote by A^{-1}. The Gauss elimination procedure for the above matrix is

$$\begin{vmatrix} 0.1 & 1.2 \\ 1 & 2 \end{vmatrix} X = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$

$$\begin{vmatrix} 1 & 2 \\ 0.1 & 1.2 \end{vmatrix} X = \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}$$

$$\begin{vmatrix} 1 & 2 \\ 0 & 1 \end{vmatrix} X = \begin{vmatrix} 0 & 1 \\ 1 & -0.1 \end{vmatrix}$$

$$X = \begin{vmatrix} -2 & 1.2 \\ 1 & -0.1 \end{vmatrix}$$

Problem 1.2-19. Use Gauss elimination to solve the following (4×4) system.

$$\begin{aligned} 0.001x_1 + 2.0002x_2 \qquad\qquad\qquad &= 2 \\ 5x_1 + \quad x_2 + \quad x_3 + \quad 2x_4 &= 1 \\ x_1 + \quad 1.2x_2 + 4.2x_3 + \quad .2x_4 &= 5 \\ x_1 + \quad 2.2x_2 + 8.2x_3 + 20.2x_4 &= 10 \end{aligned}$$

Use Gauss elimination to find the inverse of the following matrix:

$$A = \begin{vmatrix} 3 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 1 & 5 \end{vmatrix}$$

Problem 1.2-20. During the Gauss elimination process the original matrix is transformed into upper triangular form. Show that the determinant of this upper triangular matrix is the same as A except possibly for a difference in sign. This follows from the properties of the determinant given previously.

Problem 1.2-21. Show that the Gauss elimination process will fail due to a zero diagonal element only if $|A| = 0$. Show that a matrix is nonsingular if and only if its determinant is nonzero.

Problem 1.2-22. Write a computer program to evaluate the determinant of a matrix of order n. Use Gauss elimination.

Problem 1.2-23. Show that a matrix A is nonsingular if and only if its rows (or columns) are linearly independent when regarded as vectors.

1.2.5 Eigenvalues and eigenvectors of matrices. These concepts along with the norm of a matrix will be quite important in providing a better understanding of numerical methods for partial differential equations. We say that a vector x and a scaler λ are an eigenvector and corresponding eigenvalue of the matrix A if $Ax = \lambda x$ $(x \neq 0)$. Note that if x is an eigenvector, then so is βx for all nonzero scalers β. For any matrix B we know that $Bx = 0$ has a nontrivial solution x if and only if $|B| = 0$. Therefore we see that λ is an eigenvalue of A if and only if λ is a root of the determinental equation $|A - \lambda I| = 0$. This determinant is a polynomial in λ of degree n called the characteristic polynomial of A. This should be clear from the definition of a determinant since

$$|A - \lambda I| = \sum_{\pi} \operatorname{sgn}(\pi) \left(a_{1\pi(1)} - \lambda\delta_{1\pi(1)}\right) \left(a_{2\pi(2)} - \lambda\delta_{2\pi(2)}\right) \cdots \left(a_{n\pi(n)} - \lambda\delta_{n\pi(n)}\right)$$

Here δ_{ij} is the Kronecker delta defined by

$$\delta_{ij} = \begin{Bmatrix} 1 & i=j \\ 0 & i\neq j \end{Bmatrix}$$

Since a polynomial of degree n has at most n distinct roots, the matrix A has at most n distinct eigenvalues.

The following two examples should prove illuminating. If A is the 2×2 matrix

$$A = \begin{vmatrix} 2 & 1 \\ 1 & 3 \end{vmatrix}$$

then the characteristic polynomial is

$$\begin{vmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (2-\lambda)(3-\lambda) - 1 = \lambda^2 - 5\lambda + 5$$

The eigenvalues are thus $\lambda = \frac{5 \pm \sqrt{5}}{2}$. The eigenvector corresponding to $\lambda = \frac{5 + \sqrt{5}}{2}$ is the solution of

$$\begin{vmatrix} -\frac{1}{2}-\frac{\sqrt{5}}{2} & 1 \\ 1 & \frac{1}{2}-\frac{\sqrt{5}}{2} \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \end{vmatrix}$$

Thus (x_1, x_2) must be orthogonal to the vector $(-\frac{1}{2}-\frac{\sqrt{5}}{2}, 1)$ and thus $(x_1, x_2) = (1, \frac{1}{2}+\frac{\sqrt{5}}{2})$. The reader can verify that this is the solution. The vector corresponding to $\lambda = (5 - \sqrt{5})/2$ can be found in a similar fashion. The second example is the matrix

$$A = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}$$

The eigenvalues are both equal to 1 since the characteristic polynomial is $(1-\lambda)^2 = 0$. Any eigenvector must satisfy the equations

$$x_2 = \lambda x_2 = x_2$$

$$x_1 + x_2 = \lambda x_1 = x_1$$

Therefore $x_2 = 0$ and x_1 is arbitrary. Therefore all eigenvectors are a scaler multiple of (1,0). This set of eigenvectors does not form a basis for the space. We cannot write all vectors as a linear combination of eigenvectors. The subspace spanned by the eigenvectors is 1-dimensional.

Problem 1.2-24. Find the eigenvalues and eigenvectors of the matrices

$$\begin{vmatrix} 2 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & -2 & 1 \end{vmatrix} \qquad \begin{vmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 4 \end{vmatrix}$$

Problem 1.2-25. Show that the eigenvalues of an orthogonal matrix of order n must have absolute value unity.

If P is a nonsingular matrix and $B = P^{-1}AP$, then we say that the matrix B is similar to A. The transformation $P^{-1}AP$ is called a similarity transformation. If A and B are similar, then the eigenvalues of A and B are the same. In fact, if (λ,x) is an eigenvalue-eigenvector pair for A, then $(\lambda,P^{-1}x)$ is a pair for B.

If there are n independent eigenvectors of A, then A is similar to a diagonal matrix whose diagonal elements are the eigenvalues of A. (A matrix A is diagonal if $a_{ij} = 0$ for $i \neq j$.) To prove the above statement, let P be the matrix whose columns are the eigenvectors of A; $P = (p^{(1)},\ldots p^{(n)})$ where $Ap^{(i)} = \lambda_i p^{(i)}$. If D is the diagonal matrix $D = \text{diag}(\lambda_i)$, then $AP = PD$, and thus $P^{-1}AP = D$.

Conversely, if A is similar to a diagonal matrix, then A has n independent eigenvectors. In this case, the eigenvectors span the space.

Problem 1.2-26. Find a matrix A which is not similar to a diagonal matrix.

Problem 1.2-27. Given an arbitrary set of scalers λ_i, $1 \le i \le n$, and a set of n independent vectors x^i, define a matrix A with eigenvalues λ_i and eigenvectors x^i.

We will now show that given any matrix A, it is possible to find a unitary matrix U such that U^*AU is upper triangular. Since U is orthogonal $U^* = U^{-1}$, and therefore U^*AU is a similarity transform. (A matrix A is upper triangular if $a_{ij} = 0$ for $i > j$). We will first prove this for a matrix of order 2. Let λ_1 be an eigenvalue and x^1 the corresponding eigenvector. Assume that x^1 is normalized so that $x^1 \cdot x^1 = 1$. Choose x^2 such that $x^1 \cdot x^2 = 0$ and $x^2 \cdot x^2 = 1$. Define the matrix U so that its columns are the vectors x^1 and x^2. Then $U^*U = I$ and we have our reduction to triangular form since

$$U^*AU = \begin{vmatrix} \lambda_1 x^1 \cdot x^1 & x^1 \cdot Ax^2 \\ \lambda_1 x^2 \cdot x^1 & x^2 \cdot Ax^2 \end{vmatrix} = \begin{vmatrix} \lambda_1 & x^1 \cdot Ax^2 \\ 0 & x^2 \cdot Ax^2 \end{vmatrix}$$

In the n-dimensional case we choose the eigenvalue λ_1 and normalized eigenvector x^1, then form an orthonormal basis $\{x^1, x^2, \ldots, x^n\}$ whose first member is x^1. If we define U_1 so that its columns are the vectors x^j, then $U_1^* U_1 = I$. Note that the rows of U_1^* are just the vectors $\bar{x}^j$. To form $U_1^* U_1$, we take the product of the rows of U_1^* with the columns of U_1. Just as before

$$U_1^* A U_1 = \begin{vmatrix} \lambda_1 & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & & a_{2n}^{(1)} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{vmatrix} = A^{(1)}$$

The first column is thus in the desired form. Now let $\hat{A}^{(1)}$ denote the matrix of order n-1 formed by the lower right corner of $A^{(1)}$, $a_{ij}^{(1)}$, $2 \le i \le n$, $2 \le j \le n$. Define the matrix $\hat{U}_2$ of order n-1 in the same way that U_1 was defined. Then

$$\hat{U}_2^* \hat{A}^{(1)} \hat{U}_2 = \begin{vmatrix} \lambda_2 & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{33}^{(2)} & & \\ \vdots & & & \\ 0 & a_{3n}^{(2)} & \cdots & a_{nn}^{(2)} \end{vmatrix}$$

We define a matrix U_2 of order n by $U_{2ij} = \hat{U}_{2ij}$ $i \ge 2$, $j > 2$, and $U_{211} = 1$, $U_{21j} = 0$, $U_{2i1} = 0$ for $2 \le i$, $2 \le j$, that is

$$U_2 = \left| \begin{array}{c:c} 1 & 0 \; \cdots \cdots \; 0 \\ \hdashline 0 & \\ \vdots & \hat{U}_2 \\ 0 & \end{array} \right|$$

Then careful inspection will show the following to be true

$$U_2^* U_1^* A U_1 U_2 = \begin{vmatrix} \lambda_1 & a_{12}^{(2)} & a_{13}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & \lambda_2 & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & & & & \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{vmatrix}$$

If we continue this procedure, we will obtain the desired result.

Problem 1.2-28. A matrix is said to be symmetric if $A^T = A$, and Hermitian if $A^* = A$. Show that a symmetric (or Hermitian) matrix must have real eigenvalues. Hint: Consider an eigenvalue-eigenvector pair λ, x. Note that $x^*Ax = \lambda x^*x$ and now take the conjugate transpose of both sides of this equation to obtain $\bar{\lambda} = \lambda$.

Problem 1.2-29. Show that any symmetric matrix A can be reduced to diagonal form by an orthogonal similarity transformation. There is a matrix U such that $U^TU = I$, $U^TAU = D$ where D is diagonal. Stated otherwise, the normalized eigenvectors of a symmetric matrix form an orthonormal basis in E_n. Hint: This can be proved in the same way that we proved an arbitrary matrix can be reduced to upper triangular form by a unitary similarity transformation.

This property of symmetric matrices is very important. It makes them particularly easy to deal with.

Problem 1.2-30. Assume a matrix A of order n has n independent eigenvectors x^j and the eigenvalues satisfy the condition $|\lambda_1| > |\lambda_j|$ for $2 \le j \le n$. Any vector y can be represented in the form $y = \alpha_1 x^1 + \ldots + \alpha_n x^n$. Assume $\alpha_1 \neq 0$. Let $w^0 = y$, $v^{\nu+1} = Aw^\nu$, $w^{\nu+1} = v^{\nu+1}/|v^{\nu+1}|$ for $0 \le \nu$. Show that $\lim\limits_{\nu\to\infty} w^\nu = \beta x^1$, for some scaler β, and $\lim\limits_{\nu\to\infty} \left(\max\limits_j v_j^{\nu+1}/w_j^\nu\right) = \lambda$. This is the power method for finding the largest eigenvalue and corresponding eigenvector.

If A has eigenvalues λ_j, $1 \le j \le n$, then we define the spectral radius of A, denoted by $\sigma(A)$, to be $\sigma(A) = \max\limits_{1\le j\le n} |\lambda_j|$. Suppose we have a sequence of matrices A_ν, all of order n, $1 \le \nu < \infty$. For example, we might have

$$A_1 = \begin{vmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{vmatrix} \quad A_2 = \begin{vmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{vmatrix} \quad A_3 = \begin{vmatrix} 1/4 & 1/8 \\ 1/8 & 1/4 \end{vmatrix} \quad A_\nu = \begin{vmatrix} \frac{1}{2^{\nu-1}} & \frac{1}{2^\nu} \\ \frac{1}{2^\nu} & \frac{1}{2^{\nu-1}} \end{vmatrix}$$

We say $\lim_{\nu\to\infty} A_\nu = 0$ if all the elements of the matrices A_ν approach zero as $\nu \to \infty$.

A fundamental fact about the spectral radius is the following. For any matrix A, $\lim_{n\to\infty} A^n = 0$ if and only if the spectral radius $\sigma(A)$ is less than one ($\sigma(A) < 1$). The book of Isaacson and Keller contains a proof of this statement. It is easy to show that if the limit is zero, we must have $\sigma(A) < 1$. For suppose there is an eigenvalue λ and eigenvector x with $|\lambda| \ge 1$. Then $A^n x = \lambda^n x$. But $\lim_{n\to\infty} A^n = 0$, therefore $\lim_{n\to\infty} A^n x = 0 = \lim_{n\to\infty} \lambda^n x$. However, if $|\lambda| \ge 1$, this is clearly impossible. Therefore we must have $\sigma(A) < 1$. We will omit the proof that $\lim A^n = 0$ if $\sigma(A) < 1$.

Problem 1.2-31. If A has n independent eigenvectors and $\sigma(A) < 1$, then show $\lim_{n\to\infty} A^n = 0$. Hint: Show that $\lim A^n x = 0$ for all x, therefore $\lim A^n = 0$.

1.2.6 Matrix norms. Next we will introduce the concept of the norm of a matrix. This norm is useful in the study of the stability of finite difference schemes. First we will state the requirements which the norm of a vector must satisfy. A vector norm is a mapping which associates a non-negative real number with each vector. It is denoted by $\|x\|$. It must satisfy the conditions

1) $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.

2) For any scaler α, $\|\alpha x\| = |\alpha| \ \|x\|$.

3) For all vectors x and y $\|x+y\| \leq \|x\| + \|y\|$.

First we observe that the length of a vector is a vector norm, usually called the Euclidean or L_2 norm. We have denoted the length of a vector x by $|x|$. Another frequently used notation is $\|x\|_2$, that is

$$\|x\|_2 = |x| = \sqrt{\sum_{j=1}^{n} |x_j|^2}$$

The subscript 2 is used because we have the so-called L_2 norm. The L_p norm (p, a positive integer) is defined by

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}$$

The maximum norm, sometimes called the L_∞ norm, is defined by

$$\|x\|_\infty = \max_{1 \leq j \leq n} |x_j|$$

These norms all provide a way to measure the "size" of a vector x -- a different measure for each norm. The most commonly used norms are the L_2 and L_∞ norms. To gain some idea of how these norms measure the size of a vector, we might look at those vectors in E_2 for which $\|x\| = 1$. We will do this for four norms.

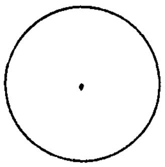
1) Consider $\|x\|_2 = \sqrt{x_1^2 + x_2^2} = 1$. The set of points is a circle of radius 1.

2) For the maximum norm we have $\|x\|_\infty = \text{Max}\{|x_1|, |x_2|\} = 1$. This set is a square of side length 2 centered about the origin.

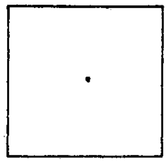
3) For the L_1 norm $\|x\|_1 = |x_1| + |x_2| = 1$. This set is also a square of side length $\sqrt{2}$ centered about the origin and rotated 45 degrees.

4) We could define a norm by $\|x\| = \sqrt{\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2}}$. We will show later that this definition does satisfy the conditions for a norm. The set is an ellipse whose axes have length a and b.

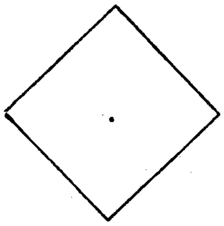
The figures below show the set $\|x\| = 1$ for the four cases.



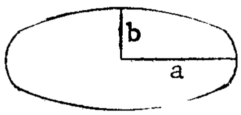
1



2



3



4

Problem 1.2-32. Show that $\|x\|_1$ and $\|x\|_\infty$ as defined above satisfy the conditions for a norm. Let A be a symmetric matrix all of whose eigenvalues are greater than zero--then A is said to be positive definite. Show that the relation $\|x\|_A = \sqrt{x^T Ax}$ defines a norm.

The reader might well ask if we are simply playing a mathematical game. Why have so many different norms? As we will see, it sometimes happens that we can prove a result using one norm but not another, or the proof may be much easier using a particular norm. It is sometimes desirable to define a special norm of the form $\|x\| = \sqrt{x^T Ax}$ in order to prove that a finite difference scheme for a differential equation is stable.

Next we will pass on to the idea of a matrix norm. We will consider only matrix norms which are induced by vector norms. Suppose $\|x\|$ is a given vector norm. Then the corresponding matrix norm (defined for any square matrix A) is defined by

$$\|A\| = \operatorname*{Max}_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Problem 1.2-33. Show that the above matrix norm satisfies the following conditions:

1) $\|A\| = \operatorname*{Max}_{\|x\|=1} \|Ax\|$

2) $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = 0$

3) $\|\alpha A\| = |\alpha| \, \|A\|$

4) $\|A+B\| \le \|A\| + \|B\|$

5) $\|AB\| \le \|A\| \, \|B\|$

Problem 1.2-34. Let $\|A\|_2$, $\|A\|_1$ and $\|A\|_\infty$ be matrix norms induced by the corresponding vector norms. Show that

1) $\|A\|_2 = \sqrt{\sigma(A^T A)}$ where $\sigma(A^T A)$ is the spectral radius

2) $\|A\|_1 = \underset{k}{\text{Max}} \sum_{j=1}^{n} |a_{jk}|$

3) $\|A\|_\infty = \underset{j}{\text{Max}} \sum_{k=1}^{n} |a_{jk}|$

Note that it is usually rather difficult to determine the spectral radius of a matrix. This usually requires a computer. However, we can quite easily evaluate the norms $\|A\|_1$ and $\|A\|_\infty$ when given the elements of the matrix A.

Problem 1.2-35. Show that for any matrix norm and any matrix A $\sigma(A) \le \|A\|$. The spectral radius can not be greater than the norm. Give an example of a matrix A where $\sigma(A) = 0$ but $\|A\|$ can be arbitrarily large depending on A.

We will illustrate the use of the matrix norm by a "perturbation analysis". Consider the linear system $Ax = b$. Suppose the matrix A is perturbed to form a new matrix $A + \delta A$ and the resulting system is solved $(A + \delta A)y = b$. We wish to estimate the difference between y and x. The perturbation in A might be caused by errors in some physical measurement and we would be interested in the resulting perturbation in the solution. Suppose we let $y = x + \delta x$. We wish to estimate the magnitude of δx. First we prove a preliminary result. If $\|B\| < 1$, then $I+B$ is nonsingular and $\|(I+B)^{-1}\| \leq 1/(1 - \|B\|)$. Since $\|B\| < 1$, we know the spectral radius of B is less than 1. Therefore B cannot have an eigenvalue equal to -1 and thus $I+B$ cannot have an eigenvalue equal to zero. Therefore, $I+B$ is nonsingular. We have $I = I + B - B$, therefore $(I+B)^{-1} = I - B(I+B)^{-1}$. Taking the norm of both sides we have $\|(I+B)^{-1}\| \leq 1 + \|B\| \, \|(I+B)^{-1}\|$ or $\|(I+B)^{-1}\| \, (1 - \|B\|) \leq 1$ or $\|(I+B)^{-1}\| \leq 1/(1 - \|B\|)$ which is the desired result.

Now we are ready to prove the main result. We have by assumption $Ax = b$, $(A + \delta A)(x + \delta x) = b$. Therefore $(A + \delta A)\delta x = - \delta Ax$. If we multiply by A^{-1} we obtain $(I + A^{-1}\delta A)\delta x = - A^{-1}\delta Ax$. Let $B = A^{-1}\delta A$ and assume $\|B\| < 1$. Then we have $(I+B)\delta x = -Bx$ and $(I+B)^{-1}$ exists. Therefore $\delta x = - (I+B)^{-1}Bx$ and $\|\delta x\| \leq \|(I+B)^{-1}\| \, \|B\| \, \|x\|$. Using the result above we obtain

$$\|\delta x\| \leq \frac{\|B\|}{1 - \|B\|} \|x\|$$

and the relative error $\|\delta x\|/\|x\|$ is therefore bounded by $\|B\|/(1 - \|B\|)$. Note that $\|B\|$ measures the size of the perturbation δA, since $B = A^{-1}\delta A$ or $AB = \delta A$ and $\|A\| \, \|B\| \geq \|\delta A\|$. Therefore $\|B\|$ is a bound for the relative perturbation in A since $\|\delta A\|/\|A\| \leq \|B\|$. However, we could have $\|\delta A\|/\|A\|$ quite small but A^{-1} fairly large and hence B might be fairly large. If A^{-1} is large we might expect $\|\delta x\|$ to be large compared with $\|\delta A\|/\|A\|$. Therefore we cannot expect a bound for

$\|\delta x\|$ in terms of $\|\delta A\|/\|A\|$ alone. Suppose we assume δx and δA are both quite small. In the equation $(I + A^{-1}\delta A)\delta x = - A^{-1}\delta Ax$ we might then ignore the second order term $A^{-1}\delta A\delta x$ to obtain $\delta x = -A^{-1}\delta Ax$. This would lead to the approximate bound $\frac{\|\delta x\|}{\|x\|} \leq \|A^{-1}\| \|\delta A\|$. Note that the correct bound is quite similar to this, namely

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \|\delta A\|}{1 - \|A^{-1}\delta A\|} .$$

1.2.7 Discrete Fourier analysis. We will briefly describe the finite Fourier representation of a vector in C_n (C_n is the complex n dimensional space). We know that if f(x), $-1 \le x \le 1$ is reasonable, then it can be represented by a Fourier series.

$$f(x) = \sum_{-\infty}^{\infty} a_k e^{ik\pi x}$$

where $a_k = \frac{1}{2}\int_{-1}^{1} f(x)e^{-ik\pi x}$. Suppose we consider the values of f on the discrete set $\{x_j\}$, $-J \le j \le J$, $x_j = j/J$. We will assume that f is periodic so that $f(-1) = f(1)$. Let f_j denote $f(x_j)$. Then we can prove that the following finite Fourier representation is valid.

$$f_j = \sum_{k=-J}^{J-1} a_k e^{ik\pi x_j} \qquad -J \le j < J \tag{1.2-1}$$

where $a_k = \frac{1}{2J} \sum_{j=-J}^{J-1} f_j e^{-ik\pi x_j}$.

We will omit the proof. Note that the formula for a_k is an approximation to the integral

$$\frac{1}{2}\int_{-1}^{1} f(x)e^{-ik\pi x}\, dx$$

If we then define vectors φ^k in C_{2J} by the formula $\varphi_j^k = e^{ik\pi j/J}$, $-J \le j < J$, $-J \le k < J$, then these vectors are orthogonal

$$\varphi^k \cdot \varphi^m = \begin{Bmatrix} 2J & \text{if } m=k \\ 0 & \text{if } m \ne k \end{Bmatrix} \qquad -J \le k,\ m < J$$

To simplify the notation we ran the index j which denotes the components of φ^k through the range $-J \le j < J$. It is more common to start the first component with index $j=1$, in which case the expression for the vector φ^k becomes

$$\varphi_j^k = e^{\frac{ik\pi j}{J} - \frac{ik\pi(J+1)}{J}} = e^{\frac{ik\pi(j-1)}{J} - ik\pi}, \qquad 1 \le j \le 2J$$

Since $\{\varphi^k\}$ is an orthogonal basis, we know that any vector v in C_{2J} can be represented in the form

$$v = \alpha_1\varphi^1 + \dots + \alpha_{2J}\,\varphi^{2J}$$

where $\alpha_j = v\cdot\varphi^j/\varphi^j\cdot\varphi^j$.

This is simply another way of writing the expansion given in equation (1.2-1). The proof of this expansion thus reduces to a proof that the vectors φ^k are orthogonal. We leave this proof to a problem in chapter 2.

An orthogonal basis for E_N can be constructed in a similar fashion. Define the vectors φ^k by

$$\varphi_j^k = \sin kj\pi/(N+1) \qquad 1 \le j \le N, \quad 1 \le k \le N$$

Problem 1.2-36. Show that the above vectors φ^k are orthogonal. First prove the formula

$$\sum_{j=1}^{J} \cos j\theta = \tfrac{1}{2}\left[-1 + \frac{\sin(J+\frac{1}{2})\theta}{\sin\frac{\theta}{2}}\right] \qquad \text{if } \theta \ne 2m\pi$$

Note that $\sum_{j=0}^{J} r^j = \frac{1-r^{J+1}}{1-r}$ and use $\cos j\theta = \frac{e^{j\theta} + e^{-j\theta}}{2}$. Now use $\sin A \sin B = \frac{1}{2}[\cos(A-B) + \cos(A+B)]$ to prove orthogonality. What is the value of $\varphi^k \cdot \varphi^k$?

1.3 Some General Comments on Partial Differential Equations.

We will attempt to classify partial differential equations (PDE) into three types--elliptic, parabolic, and hyperbolic. The diversity of PDE is such that it is not possible to neatly classify these equations into three groups. However, the basic approach used for the numerical solution of elliptic equations is quite different from that used for the other two types. As we will see, elliptic equations are "pure boundary value" problems whereas parabolic and hyperbolic equations are "initial value" problems. Therefore it is important for the numerical analyst to have some feeling for the nature of these equations, even if the PDE problems frequently fail to fit nicely into one of these three categories.

1.3.1 A classification of linear second-order partial differential equations--elliptic, hyperbolic and parabolic.

An explanation of the classification of PDE can be based on the following equation

$$au_{xx} + 2bu_{xy} + cu_{yy} + 2du_x + 2eu_y + fu = h(x,y) \qquad (1.3\text{-}1)$$

Here a,b,...,f are assumed to be real constants, h is a known function, $u = u(x,y)$, $u_x = \partial u/\partial x$, and similarly for the other derivatives. We let λ_1 and λ_2 be the roots of the characteristic equation

$$a\lambda^2 - 2b\lambda + c = 0 \ .$$

If $b^2-ac > 0$, then these roots are real and distinct. If we introduce the coordinate transformation $y-\lambda_1 x = \xi+\eta$ and $y-\lambda_2 x = \xi-\eta$, then $u = u(\xi,\eta)$ and equation (1.3-1) becomes

$$u_{\xi\xi} - u_{\eta\eta} + 2Du_{\xi} + 2Eu_{\eta} + Fu = H(\xi,\eta) \tag{1.3-2}$$

If $b^2-ac > 0$, then we say our original equation (1.3-1) is hyperbolic and we regard the above equation (1.3-2) as the canonical form for a hyperbolic equation.

By use of the following transformation of the dependent variable $u = ve^{-D\xi-E\eta}$ we can reduce the hyperbolic equation to

$$v_{\xi\xi} - v_{\eta\eta} + kv = f(\xi,\eta) \tag{1.3-3}$$

Problem 1.3-1. Show that the transformations described above produce equation (1.3-3).

If $b^2-ac < 0$, then we say the equation is elliptic. In this case we use the transformation $y-\lambda_1 x = \xi+i\eta$, $y-\lambda_2 x = \xi-i\eta$ and our canonical form is

$$v_{\xi\xi} + v_{\eta\eta} + kv = f(\xi,\eta)$$

If $b^2-ac = 0$, then we have the parabolic case. The canonical form below is obtained from the transformation $y-\lambda = \eta$, $\alpha y + \beta x = \xi$ where α and β are suitably chosen. The parabolic equation can be reduced to

$$v_{\xi\xi} - v_{\eta} = f(\xi,\eta) .$$

Most problems which arise in practice cannot be reduced to one of these simple forms. However a study of these simple equations is essential because it gives us some idea of how to proceed with a numerical solution of the more complicated equations. (These three equations are really not so simple--much deep mathematics has been created in an attempt to understand these equations.)

1.3.2 An elliptic equation - Laplaces equation. Solution by separation of variables. We will first consider Laplace's equation $u_{xx} + u_{yy} = 0$. This is an elliptic equation. The separation of variables technique will yield a certain family of solutions to the problem. The basic assumption involved in the separation of variables is the existence of functions F(x) and G(y) such that u(x,y) = F(x)G(y). Substitution into Laplace's equation yields

$$F''(x)G(y) + G''(y)F(x) = 0 .$$

This can be written

$$\frac{F''(x)}{F(x)} = - \frac{G''(y)}{G(y)} .$$

The only way a function of x can equal a function of y (x and y are independent) is to have both functions equal to a constant. If this constant is positive, then we have

$$\frac{F''(x)}{F(x)} = \lambda^2 \qquad \frac{G''(y)}{G(y)} = - \lambda^2$$

The general solution of these equations yields

$$F(x) = A_1 e^{\lambda x} + A_2 e^{-\lambda x}$$

$$G(y) = B_1 \sin\lambda x + B_2 \cos\lambda x$$

If the constant is negative, then the trigonometric functions appear in the solution for F and the exponentials in the solution for G.

Note that Laplace's equation is symmetric in x and y, so we might expect such an interchange. Obviously many functions satisfy Laplace's equation $u_{xx} + u_{yy} = 0$. In order to get a unique solution we must impose some boundary conditions. Physical insight is frequently a great help in setting up proper boundary conditions. We might look at Laplace's equation from a physical point of view. The steady-state temperature distribution $T(x,y)$ on a flat plate satisfies Laplace's equation $T_{xx} + T_{yy}$, at least approximately. Suppose we have a square plate given by $0 \le x \le 1$, $0 \le y \le 1$. We would expect the temperature distribution in the interior of the plate to depend on the boundary conditions. If the side at $y = 0$ is insulated, we would have $T_y(x,0) = 0$ for $0 \le x \le 1$; that is, the normal temperature gradient would vanish. Perfect insulation implies zero heat flux which in turn requires zero temperature gradient. If the side at $y = 0$ is held in a bath of boiling water, then we would have $T(x,0) = 100^{\circ}C$. We must specify a boundary condition on each side of the square in order to obtain a unique solution for Laplace's equation. This is a fundamental property of elliptic equations. In order to obtain a solution, we must specify a boundary condition at all points of the boundary.

Suppose we attempt to solve the following boundary value problem for Laplace's equation.

$$u_{xx} + u_{yy} = 0 \qquad 0 \le x \le 1, \quad 0 \le y \le 1$$

$$u(x,0) = f(x)$$

$$u(x,1) = 0$$

$$u(y,0) = u(y,1) = 0$$

We assume that f(x) is smooth enough to have a Fourier expansion

$$f(x) = \sum_{k=1}^{\infty} a_k \sin k\pi x \quad \text{where} \quad \sum_{k=1}^{\infty} k^2 |a_k| < \infty.$$

Problem 1.3-2. Verify that the series below is a solution of the above problem.

$$u(x,y) = \sum_{k=1}^{\infty} a_k \frac{\sinh k\pi(1-y)}{\sinh k\pi} \sin \pi kx$$

where $\sinh z = \frac{1}{2}(e^z - e^{-z})$.

1.3.3 <u>A hyperbolic equation - the wave equation</u>. Next we will consider a hyperbolic equation. Suppose we have a string stretched between two points on the x-axis. If we pluck the string, it will vibrate. The displacement $u(x,t)$ of the string from its undisturbed position along the x-axis will then be a function of position x and time t. This displacement will satisfy the hyperbolic equation below (c is a constant) which is called the wave equation (a derivation of the vibrating string equation can be found in many places).

$$u_{tt} = c^2 u_{xx} .$$

In order to obtain a unique solution we need the boundary conditions which state that the string is held fixed at $x = 0$ and $x = 1$.

$$u(0,t) = u(1,t) = 0$$

We also need the initial conditions

$$u(x,t) = f(x)$$

$$u_t(x,t) = g(x) \qquad 0 \leq x \leq 1$$

These are also really boundary conditions. As we will see, the solution of the wave equation can be obtained by "marching forward" in time. Hence we call this an initial value problem, and these conditions are called initial conditions. How do we know these are proper initial-boundary conditions for the wave equation? We obtain them from physical insight based on a derivation of the differential equation. We might then assume the mathematician's role and prove that there is a unique solution of the wave equation which satisfies these conditions. If our initial-boundary conditions are not correct, we may find that there is no solution or there may be more than one solution. For the complicated nonlinear PDE problems which arise in practice, we may not be successful in our mathematician's role. We then have to depend solely on physical insight and analogy with simpler problems for which proper boundary conditions are known.

Instead of the initial-boundary problem given above, we will consider the pure initial value problem for the wave equation.

$$u_{tt} = c^2 u_{xx}$$

$$u(x,0) = f(x) \qquad -\infty < x < \infty$$

$$u_t(x,0) = g(x)$$

That is, we require the initial conditions to hold for all x. Our interval has no boundary, and thus we have no boundary conditions. If we use the following change of variables $\xi = x+ct$, $\eta = x-ct$, then the wave equation becomes

$$u_{\xi\eta} = 0$$

If we integrate this equation with respect to η we obtain

$$u_\xi = F_0(\xi)$$

If we integrate with respect to ξ we obtain

$$u(\xi,\eta) = F_1(\xi) + F_2(\eta)$$

and thus

$$u(x,t) = F_1(x+ct) + F_2(x-ct)$$

Now we must relate the functions $F_1(\xi)$ and $F_2(\eta)$ to the initial functions $f(x)$ and $g(x)$.

Problem 1.3-3. Show that the above solution can be written in the form

$$u(x,t) = \tfrac{1}{2}[f(x+ct) + f(x-ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\tau)d\tau$$

Note that we can write this solution in the form

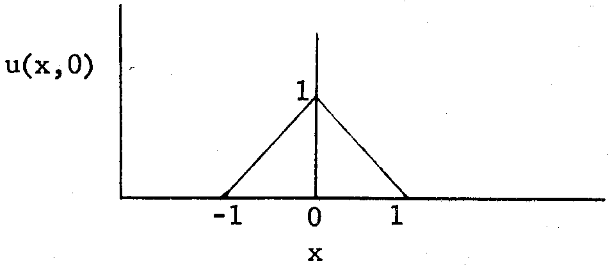
$$u(x,t) = \tfrac{1}{2}[f(x+ct) + f(x-ct)] + \tfrac{1}{2}[G(x+ct) - G(x-ct)]$$

where $G(x) = \frac{1}{c}\int_0^x g(\tau)d\tau$.

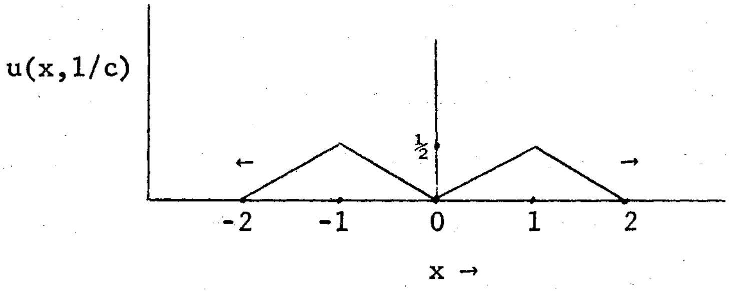
This solution shows wave propagation. Suppose we have $g(x) \equiv 0$ and

$$f(x) = \begin{cases} 1+x & -1 \le x \le 0 \\ 1-x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

This is a tent shaped curve as shown in the figure below



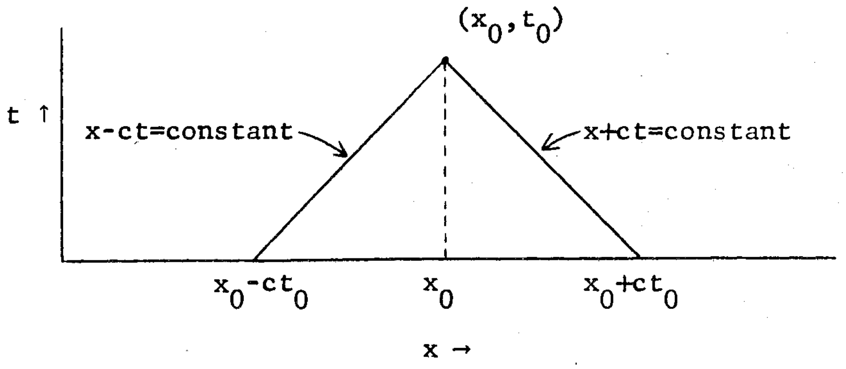
The term $\frac{1}{2}f(x+ct)$ represents a wave propagating to the left with velocity $-c$, the term $\frac{1}{2}f(x-ct)$ propagates to the right with velocity c. After time $t = 1/c$ the solution would have the form shown in the next figure.



Problem 1.3-4. Choose a reasonable function for $g(x)$, set $f(x) \equiv 0$, then describe the wave which is the resulting solution of the wave equation.

A fundamental property of hyperbolic equations is the tendency to propagate "disturbances" in the initial conditions much as the wave equation does in the example above. We will have more to say about this in a later section.

Also note that this is an initial value problem. We need only to specify the functions $u(x,0)$ and $u_t(x,0)$ at time $t = 0$. Also we have a finite "interval of dependence." If we take a point (x_0,t_0), then the value $u(x_0,t_0)$ depends on the values of the functions $f(x)$ and $g(x)$ only in the interval between x_0-ct_0 and x_0+ct_0. The figure below illustrates this interval of dependence.



We will show later that the elliptic equation $u_{tt} + u_{xx} = 0$ cannot be treated as an initial value problem. In order to solve an elliptic problem in the region $0 \le t \le T$ we would have to specify boundary conditions on the upper line $t = T$. To solve an elliptic problem in a region bounded by a closed curve, it is necessary to specify boundary conditions along the entire closed curve. The hyperbolic problem is an initial value problem. We "march" the solution ahead from $t = 0$ to $t = T$, using only the initial values at $t = 0$. In general a hyperbolic problem may have boundary conditions along "sides" such as the lines $x = 0$ and $x = 1$, as well as initial conditions at $t = 0$, but we do not need boundary conditions at the "top" $t = T$. We will give an example of such a mixed initial-boundary value problem for the wave equation in a later section.

1.3.4 <u>A parabolic equation - the heat equation</u>. Next we will consider an equation of parabolic type. This is the one-dimensional heat equation $u_t = \sigma u_{xx}$, $u = u(x,t)$. If we let u denote the temperature of a rod, then u would satisfy (approximately) this equation. Then x would denote distance along the rod and t the time. In order to obtain a unique solution we would have to specify the initial temperature $u(x,0)$. We also need some boundary conditions. If the ends of the rod are at $x = 0$ and $x = 1$, then we could specify the temperature at each end $u(0,t)$ and $u(1,t)$. Instead of the temperature we could specify the heat flux at one or both ends, $u_x(0,t)$ and $u_x(1,t)$. Suppose we consider the following problem for the heat equation

$$u_t = \sigma u_{xx}$$

$$u(x,0) = f(x)$$

$$u(0,t) = u(1,t) = 0$$

We can obtain a solution for this problem by the separation of variables technique.

Problem 1.3-5. We will assume that f(x) can be represented by a Fourier expansion

$$f(x) = \sum_{k=1}^{\infty} a_k \sin k\pi x \qquad \text{where} \qquad \sum_{k=1}^{\infty} |a_k| < \infty .$$

Now assume $u(x,t) = F(x)G(t)$. Using this assumption show that the solution of the above problem is

$$u(x,t) = \sum_{k=1}^{\infty} a_k e^{-\sigma k^2 \pi^2 t} \sin k\pi x$$

Now we will look at some properties of this solution. First of all, the solution decays with time,

$$\lim_{t\to\infty} u(x,t) = 0$$

The higher frequency modes (that is, $a_k e^{-\sigma k^2\pi^2 t}$ sin $k\pi x$ for large values of k) decay more rapidly than the lower frequency modes. Since corners (that is, discontinuities in the derivative) require greater amplitudes at the higher frequencies, we might expect the solution of the heat equation to become smoother with time. This is illustrated in figure 1.3-1 shown below. The initial function $f(x) = u(x,0)$ is given by

$$f(x) = \begin{cases} 1-x & 0 \le x \le 1 \\ 1+x & -1 \le x \le 0 \\ 0 & \text{otherwise} \end{cases}$$

The curves of $u(x,t)$ for fixed t are plotted for the values of t shown on the figure. This curve was obtained from a numerical solution of the heat equation using the method of finite differences. The curve was plotted by the computer on a graphic display device: Note how the curve becomes smoother and also spreads out with time.

The solution of the heat equation decays with time. Therefore we say the term u_{xx} is a dissipative term. We will define the energy $E(t)$ by the relation

$$E(t) = \tfrac{1}{2}\int_0^1 u^2(x,t)\,dx$$

Then from the equation $u_t = \sigma u_{xx}$ we have $uu_t = \sigma uu_{xx}$ or

$$\tfrac{1}{2}(u^2)_t = \sigma\left(uu_x\right)_x - \sigma\left(u_x\right)^2$$

Therefore if we integrate with respect to x and use the boundary conditions $u(0,t) = u(1,t) = 0$ we obtain

$$\tfrac{1}{2}\int_0^1 (u^2)_t dx = \frac{\partial E}{\partial t} = \sigma\int_0^1 \left(uu_x\right)_x dx - \sigma\int_0^1 \left(u_x\right)^2 dx$$

or $$\frac{\partial E}{\partial t} = -\sigma\int_0^1 \left(u_x\right)^2 dx$$

If we now integrate with respect to t, we obtain

$$\int_0^t \frac{\partial E}{\partial t} dt = -\sigma\int_0^t\int_0^1 \left(u_x\right)^2 dxdt$$

or $$E(t) - E(0) = -\sigma\int_0^t\int_0^1 [u_x(c,t)]^2 dxdt$$

therefore $E(t) \le E(0)$ and we know that the energy is a nonincreasing function. In fact $E(t_2) < E(t_1)$ if $t_2 > t_1$, unless $u_x(x,t) \equiv 0$ for $t_1 \le t \le t_2$. This integration by parts argument should make it clear why the u_{xx} term is an energy reducing term. In general the energy relation for the heat equation is

$$\frac{\partial E}{\partial t} = \sigma\left[u(1,t)u_x(1,t) - u(0,t)u_x(0,t)\right] - \sigma\int_0^1 \left[u_x(x,t)\right]^2 dx$$

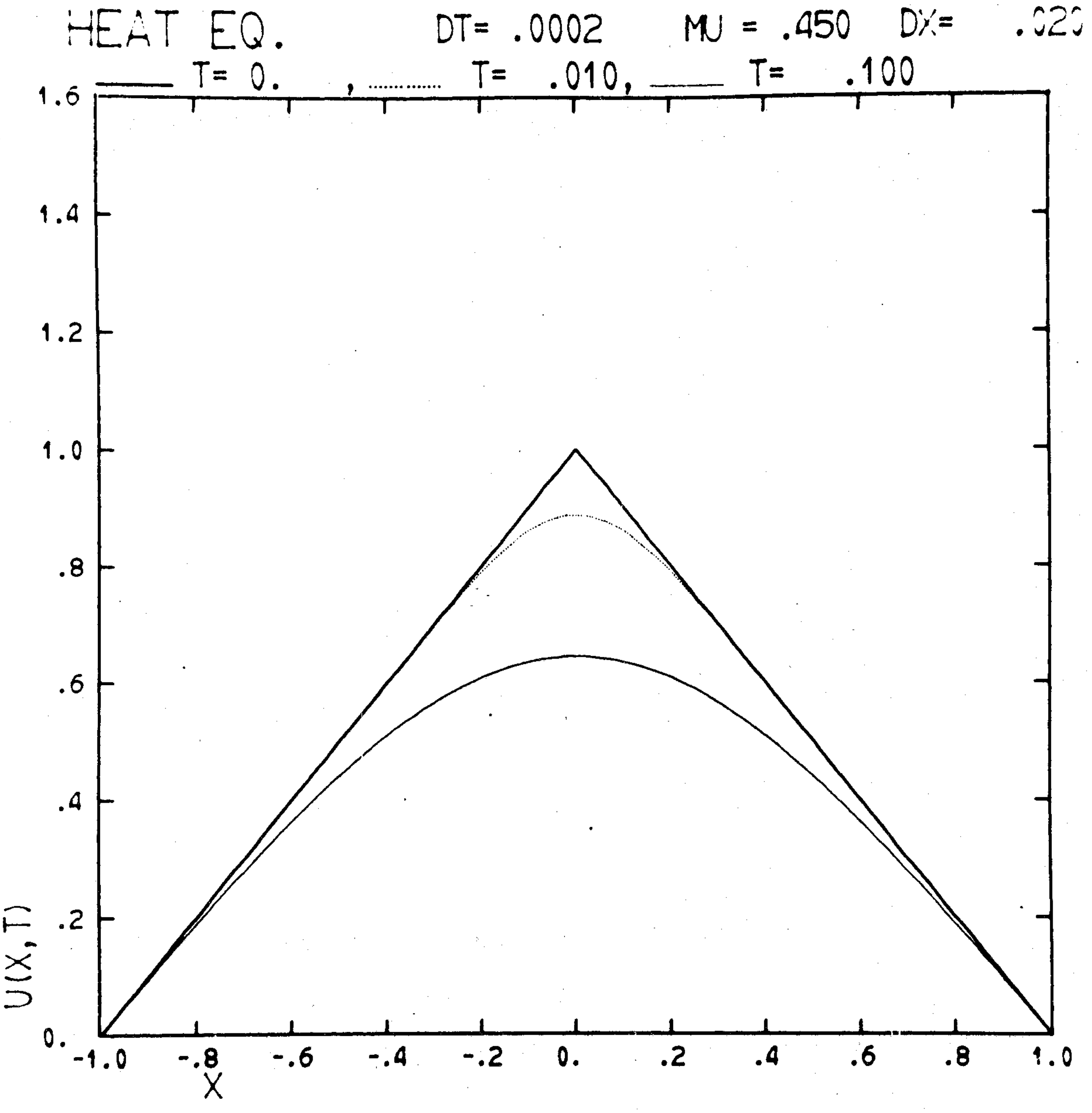

HEAT EQ. DT= .0002 MU = .450 DX= .02
T= 0. , T= .010, T= .100
1.6
1.4
1.2
1.0
.8
.6
.4
.2
0.
U(X,T)
-1.0
-.8
-.6
-.4
-.2
0.
.2
.4
.6
.8
1.0
X

Figure 1.3-1

Therefore we would expect a contribution to the energy unless u or u_x vanish at the boundary. Note that we must specify some boundary conditions in order to obtain a unique solution for the heat equation. If the boundary conditions are $u(0,t) = u(1,t) = 0$ then

$$\lim_{t\to\infty} u(x,t) = 0 \qquad 0 \le x \le 1$$

If $u(0,t) = 0$, $u(1,t) = 1$, then

$$\lim_{t\to\infty} u(x,t) = x$$

1.3.5 Properly posed problems - Hadamard's example. In order that a PDE problem be properly posed there should be a unique solution for any admissible initial-boundary conditions. However, an additional restriction is usually imposed. We require that the solution depend continuously on the data which defines the solution. By this we mean that a small perturbation in the initial-boundary conditions should produce a small perturbation in the solution. For example, we will consider the heat equation. Suppose we restrict the discussion to homogeneous boundary conditions $u(0,t) = u(1,t) = 0$. Suppose we have two solutions $u_1(x,t)$ and $u_2(x,t)$ corresponding to the initial functions $f_1(x)$ and $f_2(x)$. That is

$$u_{1t} = \sigma u_{1xx} \qquad\qquad u_{2t} = u_{2xx}$$

$$u_1(x,0) = f_1(x) \qquad\qquad u_2(x,0) = f_2(x) \ .$$

We wish to show that if the functions f_1 and f_2 are "close," then the solutions u_1 and u_2 are also close. We will say that f_1 is close to f_2 if the difference $f_1 - f_2$ is "small." We must now decide on a measure to define the size of a function $f(x)$. We will use a norm to define the size of our functions. A norm for these functions is a rule which assigns a non-negative real number to each function. Such a norm must satisfy the same conditions as the vector norms we defined in section 1.2. However, we will not tarry on the technical aspects of such norms. One such norm is the maximum norm defined by

$$\|f\|_\infty = \max_{0\le x\le 1} |f(x)|$$

A more convenient norm at this point is the L_2 norm defined by

$$\|f\|_2 = \sqrt{\int_0^1 f^2(x)dx}$$

Of course we must restrict the class of functions to those for which this integral exists.

To return to our problem, let $w = u_1 - u_2$. Then w is a solution of the heat equation with the initial value $w(x,0) = f_1(x) - f_2(x)$. If we let $E_w(t) = \frac{1}{2} \int_0^1 w^2(x,t)dx$, then we know from our previous discussion that

$$E_w(t) \le E_w(0) \text{ for } t \ge 0.$$

If we rewrite this inequality we obtain

$$\int_0^1 [u_1(x,t) - u_2(x,t)]^2 \, dx \le \int_0^1 [f_1(x) - f_2(x)]^2 \, dx$$

or $\|u_1(t) - u_2(t)\| \le \|f_1 - f_2\|$ where $u_1(t)$ denotes the function $u_1(x,t)$ for fixed t. This is the result we wanted. It shows us that the size of the perturbation in the solution is bounded by the size of the perturbation in the initial function. In other words, the solution of the heat equations depends continuously on the initial function f(x).

Suppose we next consider Laplace's equation. We look for a solution u(x,y) defined over the unit disk $x^2 + y^2 \le 1$. We assume the values of u are given on the circumference of the disk $x^2 + y^2 = 1$. Then the problem is to solve the equation

$$u_{xx} + u_{yy} = 0 \qquad x^2+y^2 < 1$$

$$u(\cos\theta,\sin\theta) = f(\theta) \qquad 0 \le \theta \le 2\pi$$

$$x = \cos\theta, \quad y = \sin\theta$$

It turns out that the solution can be written in terms of the "Poisson integral formula" (see Garabedian's book, p. 249).

$$u(r\cos\theta,\ r\sin\theta) = \frac{1}{2\pi}\int_0^{2\pi} \frac{(1-r^2)\,f(\varphi)\,d\varphi}{1-2r\cos(\varphi-\theta)+r^2} \qquad \text{for } r < 1 .$$

The boundary data for this problem is the initial function $f(\varphi)$. We could use this formula to show that u depends continuously on the initial data $f(\varphi)$. Or we could use the maximum principle which states that u cannot take on its maximum (or minimum) in the interior of the disk. In any case we have

$$\max_{\substack{0\le r<1\\ 0\le\theta\le 2\pi}} \left|u(r\cos\theta,\ r\sin\theta)\right| \le \max_{0\le\theta\le 2\pi} \left|f(\theta)\right|$$

From this relation we see that u depends continuously on f provided we use the maximum norm.

Now we come to the point of this discussion which is an example due to Hadamard (Garabedian, p. 108). Consider the pure initial value problem for Laplace's equation

$$u_{tt} + u_{xx} = 0 \qquad u = u(x,t)$$

$$u(x,0) = f(x)$$

$$u_t(x,0) = g(x) \qquad -\infty < x < \infty$$

$$0 \le t$$

Here we are treating Laplace's equation as if it were the wave equation. It can be shown that, if the functions f(x) and g(x) are analytic, then a unique solution u(x,t) exists which is also analytic. However, suppose we set the following initial conditions

$$u(x,0) = 0$$

$$u_t(x,0) = g_n(x) = \frac{1}{n}\sin(nx)$$

where n is a positive integer. It is easy to verify that the solution is given by

$$u(x,t) = \frac{1}{n^2}\sin(nx)\sinh(nt)$$

where $\sinh nt = (e^{nt} - e^{-nt})/2$.

It is clear that the initial data can be made as small as desired (uniformly in x) merely by choice of a large n. However, if we take a fixed region ($0 \le t \le 1$, for example) it is clear that u(x,t) can be made as large as desired by choice of a large n. That is

$$\lim_{n\to\infty} g_n(x) = 0 \qquad \text{for } -\infty < x < \infty$$

$$\lim_{n\to\infty} \max_{\substack{-\infty<x<\infty \\ 0\le t\le 1}} u_n(x,t) = \lim_{n\to\infty} \max_{\substack{-\infty<x<\infty \\ 0\le t\le 1}} \frac{1}{n^2}\sin(nx)\sinh(nt) = \infty$$

In other words an arbitrarily small initial function can lead to an arbitrarily large solution. Clearly the solution does not depend continuously on the initial data. Therefore, the above initial value problem for Laplace's equation is not sensible physically. It is not a properly posed problem. If we attempted to solve such a problem on a computer, we would expect any small errors in the data to produce an explosive growth of the solution.

We will give another example of the care required in formulating the proper boundary conditions for a partial differential equation. The numerical solution of this problem is discussed at length in the book by Greenspan [Greenspan, 1965]. If we write Laplace's equation in cylindrical coordinates (r,θ,z) and assume that there is no θ dependence, we obtain

$$\frac{\partial^2 u}{\partial r^2} + \frac{\partial^2 u}{\partial z^2} + \frac{1}{r}\frac{\partial u}{\partial r} = 0$$

We suppose that a solution of this equation is desired for the cylinder $0 \le z \le 1$, $0 \le r \le 1$. This solution might be the steady-state temperature in the cylinder under the assumption that the temperature on the boundary is specified; that is

$$u(r,0) = f_1(r) \qquad 0 \le r \le 1$$

$$u(r,1) = f_2(r) \qquad 0 \le r \le 1$$

$$u(1,z) = f_3(z) \qquad 0 \le z \le 1$$

Note that our region is the (r,z) square $0 \le r \le 1$, $0 \le z \le 1$. The boundary segment $(0,z)$ does not represent a physical boundary. Instead we have a coordinate system singularity at $r = 0$. This is reflected in the term $(1/r)\partial u/\partial r$ in the differential equation. We would not expect to have to specify the value of u along the segment $(0,z)$ for $0 \le z \le 1$.

Indeed it is possible to show that the above problem is properly posed without the specification of $u(0,z)$. However, if we consider the differential equation

$$\frac{\partial^2 u}{\partial r^2} + \frac{\partial^2 u}{\partial z^2} + \frac{K}{r}\frac{\partial u}{\partial r} = 0$$

with $K < 1$, then we must specify $u(0,z)$ in order to obtain a unique solution. If $K \geq 1$, then we obtain a unique solution without specification of $u(0,z)$. We refer the reader to Greenspan [1965] for further information which includes a careful treatment of a numerical approximation for the solution of this equation. The point of this example is to show that the proper boundary conditions for a problem are not always obvious. Unless we know the proper boundary conditions, we are not likely to get an accurate numerical approximation.

1.3.6 The method of characteristics applied to a simple hyperbolic equation.

We have given the wave equation as a simple example of a hyperbolic equation. To illustrate numerical methods we prefer to use another less complicated hyperbolic equation, namely

$$u_t + cu_x = 0 \qquad\qquad u = u(x,t)$$

The pure initial value problem is properly posed for this equation. If the initial condition is $u(x,0) = f(x)$ for $-\infty < x < \infty$, then the solution is $u(x,t) = f(x-ct)$ as the reader can easily verify. It is clear that this solution represents a wave which propagates to the right (if $c > 0$) and is unchanged in form as it moves. We can show that "energy" is conserved in the solution of this equation. This is obvious for this simple equation since $u(x,t) = f(x-ct)$. However, we will use a different method to show energy conservation since this method can be applied to more general cases. We will define the energy by

$$E(t) = \tfrac{1}{2} \int_{-\infty}^{\infty} u^2(x,t)\,dx \ .$$

We assume the solution is such that this integral exists. For example, the solution might vanish outside some interval--the interval may depend on t. We can then use the same method that we used for the heat equation to obtain

$$uu_t = - cuu_x$$

$$\tfrac{1}{2} \int_{-\infty}^{\infty} (u^2)_t dx = - \frac{c}{2} \int_{-\infty}^{\infty} (u^2)_x$$

$$\frac{\partial E}{\partial t} = -\frac{c}{2} \lim_{R\to\infty} [u^2(R,t) - u^2(-R,t)] = 0$$

We have assumed our solution is smooth enough so that $\lim_{R\to\infty} u(\pm R,t) = 0$. Remember that $\int_{-\infty}^{\infty} u^2 dx$ exists so this is not an unreasonable additional restriction. Thus we have

$$\frac{\partial E}{\partial t} = 0$$

or $\quad E(t) = E(0) = \frac{1}{2} \int_{-\infty}^{\infty} f^2(x)dx$

We will next return to the equation $u_t + cu_x = 0$ and define the characteristic curves for this equation. These are straight lines $x-ct = K$ (K is a constant). Along these lines u is a function of t alone, $u(t) = u(K+ct, t)$. If we take the total derivative of u with respect to t and use the fact that u is a solution of the hyperbolic equation we obtain

$$\frac{du}{dt} = u_t + \frac{dx}{dt} u_x = u_t + cu_x = 0$$

Therefore u is a constant along these characteristics. This is the reason we defined the characteristics to be lines with slope $\frac{dx}{dt} = c$. If we consider the point (x_0,t_0), then we see that the characteristic through this point has the equation $x-ct = K = x_0 - ct_0$. The point $(x_1,0)$ where $x_1 = x_0 - ct_0$ lies on this characteristic, therefore $u(x_0,t_0) = u(x_1,0) = f(x_1) = f(x_0 - ct_0)$. The method of characteristics has thus yielded the solution of the initial value problem $u(x,t) = f(x-ct)$.

This equation is too trivial to illustrate the power of the method of characteristics. If we consider a more general equation, $u_t + xu_x = 0$, we can still use the method of characteristics even though the solution is no longer obvious.

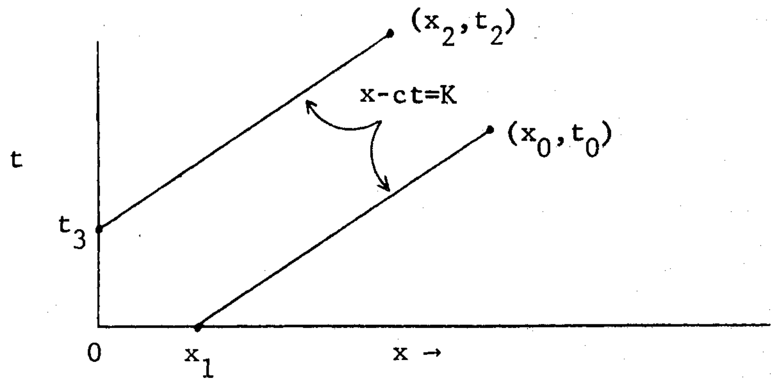
Perhaps more important the method will yield the solution of the following initial-boundary value problem.

$$u_t + cu_x = 0 \qquad c > 0, \quad x \geq 0, \quad t \geq 0$$

$$u(x,0) = f(x)$$

$$u(0,t) = g(t) \ .$$

If we take any point (x_0,t_0) with $t > 0$ and draw the characteristic $x-ct = K = x_0 - ct_0$ through the point back toward the initial line $t = 0$, then this characteristic may strike the line $t = 0$ or it may first strike the left boundary line $x = 0$. The figure below illustrates the situation.



In the case of (x_0,t_0) we have $u(x_0,t_0) = u(x_0-ct_0, 0) = f(x_0-ct_0)$ as before. In the other case we have $u(x_2,t_2) = u(0,t_3) = g(t_3) = g(t_2-x_2/c)$. If the point $(0,t_3)$ lies on the characteristic through

(x_2, t_2), we have $x_2 - ct_2 = -ct_3$ or $t_3 = t_2 - x_2/c$. Therefore the solution is

$$u(x,t) = \begin{cases} f(x-ct) & x \geq ct \\ g(t-x/c) & x < ct \end{cases}$$

Problem 1.3-6. Given the problem

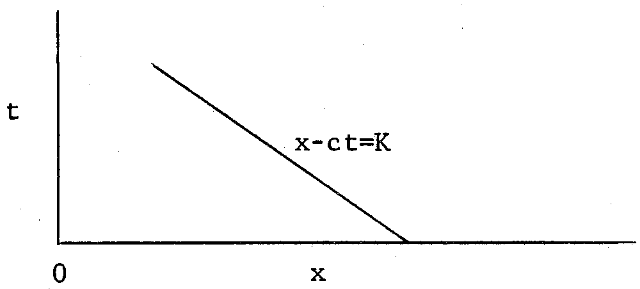
$$u_t + cu_x = 0$$

$$u(x,0) = f(x)$$

$$u(0,t) = g(t)$$

what conditions should be placed on f and g to insure that u_t and u_x exist and satisfy the equation along the line $x-ct = 0$.

Now suppose we have to solve the equation $u_t + cu_x = 0$ in the region $t \geq 0$, $x \geq 0$ with $c < 0$. What are the proper boundary conditions. The characteristics now slope down to the right as the figure below shows.



Therefore to obtain a solution we need specify only the initial condition $u(x,0) = f(x)$. We do not need any boundary condition along the line $x = 0$, and in fact we cannot impose a condition $u(0,t) = g(t)$ and expect to get a solution (why?).

Problem 1.3-7. Suppose we wish to solve the hyperbolic equation with $c > 0$ $u_t + cu_x = 0$ in the region $t \geq 0$, $0 \leq x \leq 1$. What are the proper initial-boundary conditions? Write out the solution in terms of these conditions. Do the same for the case $c < 0$.

Problem 1.3-8. Consider the hyperbolic equation $u_t + xu_x = 0$ in the region $-1 \leq x \leq 1$, $t \geq 0$. The characteristics for this problem are curves with slope $dx/dt = x$. These are curves on which u is a constant. Find the equation of these characteristics. Find the proper initial-boundary conditions for this problem. Write the solution in terms of these conditions. Consider the initial-boundary conditions $u(x,0) = 1 - x^2$, $u(-1,t) = u(1,t) = 0$ (perhaps the boundary conditions are not used). What is the nature of the solution $u(x,t)$, regarded as a function of x for fixed large t, for these initial-boundary conditions. Answer the same questions for the equation $u_t - xu_x = 0$.

1.3.7 Further remarks on the classification of partial differential equations. Given a system of partial differential equations, how can we determine whether the system is elliptic, parabolic, or hyperbolic. Actually a system may not fit into any of these categories. We will next consider a few examples. Suppose we have a system of equations for the vector unknown $u(x,t) = (u_1(x,t), \ldots, u_N(x,t))^T$

$$\frac{\partial u_1}{\partial t} = a_{11}\frac{\partial u_1}{\partial x} + a_{12}\frac{\partial u_2}{\partial x} + \ldots + a_{1N}\frac{\partial u_N}{\partial x}$$

$$\vdots$$

$$\frac{\partial u_N}{\partial t} = a_{N1}\frac{\partial u_1}{\partial x} + a_{N2}\frac{\partial u_2}{\partial x} + \ldots + a_{NN}\frac{\partial u_N}{\partial x}$$

It is quite profitable to write this equation in matrix form where we regard u as a column vector.

$$\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x}$$

Suppose A has N distinct real eigenvalues $\{\lambda_1, \ldots, \lambda_N\}$. Then there is a nonsingular matrix P of eigenvectors such that PAP^{-1} is a diagonal matrix whose diagonal elements are the eigenvalues. We will assume that the coefficients a_{ij} are constants. If we define a vector w by $w = Pu$, then the differential equation transforms into diagonal form

$$Pu_t = PAu_x$$

$$w_t = PAP^{-1}Pu_x = Dw_x$$

or $$\frac{\partial w_1}{\partial t} = \lambda_1 \frac{\partial w_1}{\partial x}$$

$$\vdots$$

$$\frac{\partial w_N}{\partial t} = \lambda_N \frac{\partial w_N}{\partial x}$$

These equations are no longer coupled to one another, except possibly through the boundary conditions, and each is a simple hyperbolic equation which we can solve by the method of characteristics. Therefore it is reasonable to call an equation $u_t = Au_x$ hyperbolic if the matrix A has real distinct eigenvalues, or if it has a linearly independent set of N real eigenvectors. If some of the eigenvalues of A are complex, then A is not hyperbolic. Note that if we have a solution u(x,t) of the wave equation $u_{tt} = c^2 u_{xx}$, and we define v by

$$v(x,t) = \frac{1}{c} \int_0^x u_t(\xi,t)d\xi + c \int_0^t u_x(0,\tau)d\tau$$

then $u_t = cv_x$ and $v_t = cu_x$. Therefore the wave equation is equivalent to the system of equations $w_t = Aw_x$, where $w = (u,v)^T$,

$$A = \begin{vmatrix} 0 & c \\ c & 0 \end{vmatrix}$$

The eigenvalues of A are ±c. The elements of A might be functions of x and t. We would still have a hyperbolic system if the eigenvalues of A were real and distinct for all x and t and also sufficiently smooth functions of x and t. We might also have a nonlinear hyperbolic system such as $u_t + uu_x = 0$. This equation has the character of a hyperbolic equation but the nonlinearity creates many interesting effects. We will discuss this in a later chapter.

Suppose we look at examples of elliptic systems. Laplace's equation in three dimensions where $u = u(x,y,z)$ is certainly elliptic.

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$

We could also consider Laplace's equation in n dimensions for $u(x_1,\ldots,x_n)$; that is

$$\frac{\partial^2 u}{\partial x_1^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2} = 0$$

We could have an elliptic equation with variable coefficients

$$a(x,y)\,\frac{\partial^2 u}{\partial x^2} + b(x,y)\,\frac{\partial^2 u}{\partial y^2} = 0$$

where $a > 0$ and $b > 0$ for all points (x,y) of the region in which the equation is defined.

The equation below for $u = u(x_1,\ldots,x_n)$ where the matrix of coefficients $A = (a_{ij})$ is symmetric and positive definite is elliptic.

$$a_{ij}\,\frac{\partial^2 u}{\partial x_i\,\partial x_j} = 0$$

Problem 1.3-9. Let Q be the orthogonal matrix whose columns are the eigenvectors of A; that is, $QAQ^T = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ where the λ_i are the eigenvalues of A. Define the transformation $(x_1,\ldots,x_n) \to (y_1,\ldots,y_n)$ by $y = Qx$. Show that the equation then assumes the following form

$$\lambda_1 \frac{\partial^2 u}{\partial y_1^2} + \lambda_2 \frac{\partial^2 u}{\partial y_2^2} + \ldots + \lambda_n \frac{\partial^2 u}{\partial y_n^2} = 0$$

Note that the λ_i are all positive and therefore this is an elliptic equation. How does this tie in with the statement that the equation $au_{xx} + 2bu_{xy} + cu_{yy} = 0$ is elliptic if and only if $b^2 - ac < 0$?

The biharmonic equation

$$\frac{\partial^4 u}{\partial x^2} + 2 \frac{\partial^4 u}{\partial x^2\, \partial y^2} + \frac{\partial^4 u}{\partial y^2} = 0$$

is also an elliptic equation. Note that this equation can be written as

$$\nabla^2(\nabla^2 u) = 0$$

where ∇^2 is the Laplacian operator

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

If we impose the boundary conditions $\nabla^2 u = f_1(x,y)$ and $u = f_2(x,y)$ on the boundary of our domain, then the biharmonic equation can be solved in two steps:

$$\nabla^2\varphi = 0$$

$$\varphi = f_1 \quad \text{on the boundary}$$

$$\nabla^2 u = \varphi$$

$$u = f_2 \quad \text{on the boundary}$$

An elliptic problem is a boundary value problem rather than a "marching problem." This fact is of fundamental importance for the numerical solution of these equations.

Lastly we will look at a linearized version of Burger's equation, namely

$$u_t + cu_x = \sigma u_{xx} \tag{1.3-4}$$

$$u(x,0) = f(x) \qquad -\infty < x < \infty \ .$$

This is a pure initial value problem--no boundary conditions. Since this problem is defined over the entire real line we will use the Fourier integral transform to solve it. If f(x) is a continuous function such that the following integral exists,

$$\int_{-\infty}^{\infty} |f(x)|dx$$

then

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\alpha x} F(\alpha)d\alpha$$

where $F(\alpha)$ is the Fourier integral transform defined by

$$F(\alpha) = \int_{-\infty}^{\infty} f(x) e^{-i\alpha x} dx$$

The Fourier transform of the derivative (under suitable conditions) is given by $i\alpha F(\alpha)$. If we take the Fourier transform of $u(x,t)$ with respect to x, then we obtain

$$U(\alpha,t) = \int_{-\infty}^{\infty} u(x,t) e^{-i\alpha x} dx$$

If we take the transform of the differential equation (1.3-4), then we obtain

$$U_t + i\alpha c U = -\alpha^2 \sigma U$$

$$U(\alpha,0) = \int_{-\infty}^{\infty} f(x) e^{-i\alpha x} dx = F(\alpha)$$

The solution of this equation is

$$U(\alpha,t) = F(\alpha) e^{(-i\alpha c - \alpha^2 \sigma) t}$$

Therefore the function $u(x,t)$ is given by

$$u(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} U(\alpha,t) e^{i\alpha x} d\alpha = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\alpha) e^{(-i\alpha c - \alpha^2 \sigma) t + i\alpha x} d\alpha$$

The reader should verify that if $\sigma = 0$, then the above integral yields

$$u(x,t) = f(x-ct)$$

This is the solution that we have already obtained for the equation $u_t + cu_x = 0$. If we let g(x,t) denote the solution of the heat equation obtained by setting c = 0, then

$$g(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\alpha)e^{-\alpha^2\sigma t}\ e^{i\alpha x}\ d\alpha$$

By substitution of x-ct for x in the above equation we obtain

$$u(x,t) = g(x-ct,\ t)$$

The function g(x,t) represents a diffusion or dissipation with time of the function f(x). If σ is small, then the diffusion will be slow. The function g(x,t) is simply a solution of the heat equation $u_t = \sigma u_{xx}$. When we add the term cu_x to the equation, we cause this diffused solution to propagate with velocity dx/dt = c. Therefore this equation has a mixture of hyperbolic and parabolic properties. However, strictly speaking, it is a parabolic equation as long as $\sigma > 0$. But the numerical analyst will have to take the hyperbolic nature of the equation into account particularly if σ is much smaller than c.

2. AN INTRODUCTION TO DIFFERENCE SCHEMES FOR INITIAL VALUE PROBLEMS. THE CONCEPTS OF STABILITY AND CONVERGENCE.

Here we will be concerned with initial value problems such as the heat equation, rather than pure boundary value problems such as Laplace's equation. In chapter 1 we tried to point out the fundamental difference between these two types of problems. An initial value problem is a "marching problem". We solve it by marching forth with a finite difference scheme, starting with the given initial values of the solution. In chapter 2, we will deal with the very fundamental concept of convergence - does the finite difference solution converge to the solution of the differential equation. It will turn out that convergence is closely related to stability. Roughly speaking, a difference scheme is stable if a small perturbation in the initial values produces a correspondingly small perturbation in the solution of the difference equation. We will illustrate these ideas by looking at difference schemes for the simple heat equation

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2} \quad \text{or} \quad \frac{\partial u}{\partial t} = \sigma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

or the simple hyperbolic equation

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0.$$

Of course, difference schemes are seldom used for such simple equations. However, these problems provide us the intuition necessary to set up difference schemes for more complicated problems. For such problems it is always important to have a good knowledge of the physics from whence the problems came. However, it is also important to understand the basic nature of difference schemes and that is our concern here.

2.1 A Finite Difference Scheme for the Heat Equation - the Concept of Convergence.

We will describe a method for solving the following problem:

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2} \qquad \begin{array}{l} u = u(x,t) \\ 0 \le x \le 1 \quad 0 \le t \end{array} \qquad (2.1\text{-}1)$$

$$u(x,0) = f(x) \qquad u(0,t) \equiv u(1,t) \equiv 0$$

We assume $f(x) = \sum_{r=1}^{\infty} a_r \sin r\pi x$ where the series $\sum_{r=1}^{\infty} |a_r|$ converges.

If we are going to solve this problem on a digital computer, we must reduce it to the consideration of a finite set of numbers. The strip $0 \le x \le 1$, $0 \le t$ clearly contains an infinite number of points. Thus, our initial step is to make the problem discrete in the x-direction. That is, we will compute the solution only at the finite set of points $\{x_j\}$ where $0 \le j \le J$, $x_0 = 0$, $x_J = 1$, $x_j < x_{j+1}$. We will take the points to be equally spaced; that is, $x_{j+1} - x_j = \Delta x$ where $\Delta x = 1/J$.

We thus need to restate the problem in terms of this discrete set of x values. For the time being, we will not make the problem discrete in time. It will remain on a continuum in the t-direction. We will approximate the second derivative $\partial^2 u/\partial x^2$ by the centered three point formula

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{\Delta x^2}$$

Problem 2.1-1. Show that

$$\frac{\partial^2 u}{\partial x^2}(x,t) = \frac{(u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t))}{\Delta x^2} - \Delta x^2 \, \tau(x,t)$$

where $\tau(x,t) = u_{x^4}(\eta,t)/12$, where $u_{x^4} = \frac{\partial^4 u}{\partial x^4}$, $|x-\eta| < \Delta x$.

Hint: Use the Taylor series expansion in x of u(x,t). Assume u is sufficiently differentiable.

We will next use a notation which has become quite standard for this type of problem. We let $u_j(t) = u(x_j,t)$. Then we can write our approximation as follows:

$$\frac{\partial u_j}{\partial t} = \sigma \left(\frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} \right) \qquad 1 \le j \le J-1 \qquad (2.1\text{-}2)$$

$$u_0(t) = u_J(t) = 0 \qquad 0 \le t$$

$$u_j(0) = f(x_j) = f_j \qquad 1 \le j \le J-1.$$

We thus have a coupled set of ordinary differential equations for the unknown functions $u_j(t)$ $1 \le j \le J-1$.

Problem 2.1-2. Solve the system of equations 2.1-2. Assume $f(x) = \sum_{r=1}^{\infty} a_r \sin r\pi x$.

Hint: Try $u_j(t) = \sum_{r=1}^{\infty} A_r(t) \sin r\pi x_j$ and solve for $A_r(t)$.

We must now make the problem discrete in time in order to solve these equations on a digital computer. One approach would be the use of a standard integration scheme, such as "Runge-Kutta" or the predictor-corrector "Adams-Moulton". This is usually not done for reasons we will discuss later. Instead, we use a less sophisticated method for solving this system of equations.

We illustrate the method of solution of the initial value problem for the ordinary differential equation $y' = f(y,t)$ $y(0) = y_0$. Perhaps the simplest way to solve an ordinary differential equation is the "Euler-Cauchy" scheme defined as follows. We approximate the time derivative by

$$y'(t) = \frac{y(t + \Delta t) - y(t)}{\Delta t}$$

where Δt is the "step size". Then if we substitute into the differential equation $y' = f(y,t)$ we have

$$y^{(n+1)} = y^{(n)} + \Delta t\ f(y^{(n)}, t_n)$$

$$y^{(0)} = y_0$$

Here we have used the notation

$$y^{(n)} = y(t_n) \text{ where } t_n = n\Delta t \text{ with } n \text{ a non-negative integer.}$$

The equations clearly define the sequence of values $y^{(n)}$ by a "marching procedure", starting with the given initial value $y^{(0)} = y_0$.

Now we will use the same "Euler-Cauchy" method on the system of equations (2.1-2). We use the notation which is quite standard

$$u_j^{(n)} = u(x_j, t_n) \qquad x_j = j\Delta x \qquad t_n = n\Delta t$$

where Δx is the step size in the x-direction and Δt is the increment in the t-direction. We have now made the problem discrete in both space and time. We will hereafter write u_j^n instead of $u_j^{(n)}$, since it will be reasonably clear that we do not mean the n^{th} power of u_j. Hereafter, we will usually use a capitol U_j^n to denote the solution of the difference equations and a lower case $u(x,t)$ to denote the solution of the differential equation. The marching scheme is now defined by the following equations

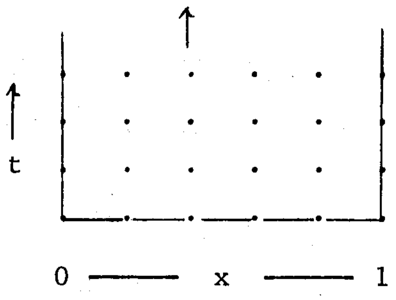
$$U_j^{n+1} = U_j^n + \frac{\sigma\Delta t}{\Delta x^2} (U_{j+1}^n - 2U_j^n + U_{j-1}^n) \qquad 1 \le j \le J-1$$

$$U_j^0 = f(x_j) = f_j \qquad 1 \le j \le J-1 \qquad (2.1\text{-}3)$$

$$U_0^n = U_J^n = 0$$

Note that the term $\sigma(U_{j+1}^n - 2U_j^n + U_{j-1}^n)/\Delta x^2$ takes the place of $f(y^n, t_n)$ in the above Euler-Cauchy formula.

We start with the given values U_j^0 and march ahead to obtain U_j^n. We will thus obtain an approximation to the values of u(x,t) on the discrete mesh (x_j, t_n) pictured below



The fundamental question is the accuracy of the approximation. Given values of the mesh spacing Δx and Δt and knowledge of the initial value f(x), then what is an upper bound for the error in the finite difference approximation. Since we will not know the solution u(x,t), we would like the error bound to depend only on the "data" for the problem, namely, f(x), Δt, and Δx. Usually, it is impossible to obtain such an "a priori" error bound. Instead, it may be possible to show that the error approaches zero as the mesh spacing goes to zero.

To simplify the description which follows, we will denote the vector u_j^n $0 \le j \le J$ by u^n, that is we simply suppress the subscript j. Perhaps this is not really necessary, since it is usually clear when we are talking about the component u_j^n and when we mean the vector u^n.

What we wish to compute is the error vector $e^n = u^n - U^n$. The thing of paramount interest to the numerical analyst is the accuracy of his approximate solution, which in our case is the solution U^n of equation (2.1-3). Usually the error is initially zero since $u_j^0 = U_j^0 = f(x_j)$ for $0 \le j \le J$; that is, $e^0 = u^0 - U^0 = 0$. To compute U^n, we "march" forward computing U^{n+1} from U^n. Each step in this marching procedure adds a little error to the result. What we are interested in is the accumulated error e^n.

It is rather easy to compute the error in U^{n+1} under the assumption that there is no error in U^n, that is $U_j^n = u_j^n = u(x_j, t_n)$ for $0 \leq j \leq J$. We call this the truncation error and denote it by $\Delta t \tau_j^n$. The factor Δt is used to normalize the definition. Its use avoids an awkward final expression for the accumulated error e^n. If $U^n = u^n$, then from equation (2.1-3)

$$U_j^{n+1} = u_j^n + \mu(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \quad 0 < j < J$$

where $\mu = \sigma\Delta t/\Delta x^2$.

Thus $\Delta t \tau_j^n = e_j^{n+1} = u_j^{n+1} - U_j^{n+1} = u_j^{n+1} - u_j^n - \mu(u_{j+1}^n - 2u_j^n + u_{j-1}^n)$ for $0 < j < J$.

Note that the relation $e_0^n = e_J^n = 0$ follows from the boundary conditions.

Problem 2.1-3. Let $u(x,t)$ be a solution of equations (2.1-1). Assume the following bounds

$$\left|\frac{\partial^2 u}{\partial t^2}\right| \leq M_1 \qquad \left|\frac{\partial^4 u}{\partial x^4}\right| \leq M_2.$$

Then obtain the following estimate for the truncation error

$$|\tau_j^n| \leq \frac{\Delta t}{2} M_1 + \frac{\sigma\Delta x^2}{12} M_2.$$

Hint: To obtain this result, use a Taylor series expansion with remainder. An expansion in t for fixed x yields

$$u_j(t_n + \Delta t) = u_j(t_n) + \Delta t \frac{\partial u_j}{\partial t}(t_n) + \frac{\Delta t^2}{2} \frac{\partial^2 u_j}{\partial t^2}(\xi)$$

where $t_n \le \xi \le t_n + \Delta t$. We used the notation $u_j(t) = u(x_j, t)$. Similarly we can obtain an expansion for the space difference.

$$\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = \frac{\partial^2 u^n}{\partial x^2}(x_j) + \frac{\Delta x^2}{24}\left(\frac{\partial^4 u^n}{\partial x^4}(\eta_1) + \frac{\partial^4 u}{\partial x^4}(\eta_2)\right)$$

where $x_j \le \eta_1, \eta_2 \le x_{j+1}$. If we use the fact that u(x,t) is a solution of the differential equation, we obtain the result.

Note that we were able to obtain an expression for the truncation error in problem 2.1-3 rather easily. To evaluate this expression, we must know the unknown solution u(x,t) which seems to lead to a circular argument. After all, if we knew u(x,t), we would not be trying to compute it. However, the bounds M_1 and M_2 depend only on the solution u(x,t) and not on the mesh. From such an expression for the truncation error, we will be able to prove that the finite difference solution converges to the solution u(x,t) of equation (2.1-1). We will obtain an expression for the accumulated error e^n in terms of M_1 and M_2. We can show that e^n approaches zero as the mesh spacing goes to zero. However, we cannot evaluate the error for any particular mesh without knowledge of M_1 and M_2.

We may write our definition for the truncation error in the more common form

$$u_j^{n+1} = u_j^n + \mu(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t \tau_j^n \qquad (2.1\text{-}4)$$

Note that its evaluation depends upon a knowledge of the unknown solution $u(x,t)$. Also, note that the error $\Delta t \tau^n$ is not the error produced by stepping ahead from the n^{th} to the $(n+1)^{st}$ time level, unless $U^n = u^n$. The difference solution must be exact on the n^{th} level.

The semantics of our definition leads us to the following specious argument. Since the truncation error is the error in each step, then the accumulated error is simply the sum

$$e_j^n = \sum_{k=0}^{n-1} \Delta t \tau_j^k \qquad 0 < j < J.$$

If we have a bound τ for the truncation error $(|\tau_j^k| \le \tau)$, then $|e_j^n| \le T\tau$.

Note that $n\Delta t \tau = t_n \tau \le T\tau$.

The bound in problem 2.1-3 then yields

$$|e_j^n| \le T\left(\frac{\mu}{2\sigma} M_1 + \frac{\sigma M_2}{12}\right) \Delta x^2 \qquad 0 < j < J.$$

Thus, we have convergence since the error e^n approaches zero as Δx goes to zero. However, the first step in this argument is not correct (why?).

We proceed to obtain an estimate for the error $e^n = u^n - U^n$. By substraction of equation (2.1-3) which defines the finite difference scheme and equation (2.1-4) which defines the truncation error, we have an expression for e^n.

$$U_j^{n+1} = U_j^n + \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \qquad (2.1\text{-}3)$$

$$u_j^{n+1} = u_j^n + \mu\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + \Delta t \tau_j^n \qquad (2.1\text{-}4)$$

$$e_j^{n+1} = e_j^n + \mu\left(e_{j+1}^n - 2e_j^n + e_{j-1}^n\right) + \Delta t \tau_j^n \qquad 0 < j < J.$$

Now we use the maximum principle to bound the error. Taking the absolute value of both sides of the last equation, we obtain an inequality

$$|e_j^{n+1}| \le |1-2\mu|\,|e_j^n| + \mu|e_{j+1}^n| + \mu|e_{j-1}^n| + \Delta t|\tau_j^n|$$

Now we let ϵ^n denote the value of the largest component of the vector e^n, that is

$$\epsilon^n = \max_{0<j<J} |e_j^n|$$

We now require that $\mu < \frac{1}{2}$. We will discuss the significance of this requirement later. For the present, we simply observe that we cannot use the maximum principle without this requirement. Taking the maximum of both sides of the above equation over the mesh index j, we obtain

$$\epsilon^{n-1} \le (1-2\mu)\,\epsilon^n + \mu\epsilon^n + \mu\epsilon^n + \Delta t \max_j |\tau_j^n|$$

Observe that $\max_j |u_j + w_j| \le \max_j |u_j| + \max_j |w_j|$ and $|1-2\mu| = 1-2\mu$ since $\mu < \frac{1}{2}$.

Now we assume that we do not integrate past $t = T$, that is $n\Delta t \leq T$. We let

$$\tau = \max_{\substack{0<j<J \\ n\Delta t<T}} |\tau_j^n|$$

A bound for τ, in terms of the unknown function $u(x,t)$, is provided by problem 2.1-3. The above inequality thus becomes

$$\epsilon^{n+1} \leq \epsilon^n + \Delta t\tau$$

By induction we have

$$\epsilon^n \leq \epsilon^{n-1} + \Delta t\tau \leq \epsilon^{n-2} + 2\Delta t\tau \leq \ldots \leq \epsilon^0 + n\Delta t\tau.$$

But $\epsilon^0 = 0$ since $U^0 = u^0$, thus

$$\epsilon^n = \max_{0<j<J} |u_j^n - U_J^n| \leqq T\tau$$

We have assumed $\mu = \sigma\Delta t/\Delta x^2$ to be less than $\frac{1}{2}$. We will also assume that μ is a constant independent of Δx. Then if we fix the number of mesh points J, Δt is determined by $\Delta t = \mu\Delta x^2/\sigma = \mu/(\sigma J^2)$. Thus, we can speak of a limit with respect to Δx; we need not consider Δt separately since Δt is determined by Δx. Of course, not all values of Δx are allowed since $\Delta x = 1/J$. We should really speak of our mesh as being a function of J rather than Δx.

In problem 2.1-3, we obtained a bound for τ, namely

$$\tau \leq \frac{\Delta t}{2} M_1 + \frac{\sigma\Delta x^2 M_2}{12}$$

Using $\mu = \sigma\Delta t/\Delta x^2$ we have

$$\tau \leq \left(\frac{\mu}{2\sigma}M_1 + \frac{\sigma M_2}{12}\right)\Delta x^2$$

Therefore

$$\underset{\substack{0<j<J \\ n\Delta t<T}}{\text{Max}} |u_j^n - U_j^n| \leq C\,\Delta x^2$$

where $\quad C = T\left(\frac{\mu}{2\sigma}M_1 + \frac{\sigma M_2}{12}\right)$.

We have no way of evaluating C unless we know the unknown solution u(x,t) since the constants M_1 and M_2 are bounds for the derivatives of u(x,t). However, the value of C is independent of the mesh and of the finite difference approximation. It depends only on the solution of the equation and not on our scheme for approximating the differential equation (except for the constant μ). This permits us to say that our scheme is convergent, that is

$$\lim_{\substack{\Delta x\to 0 \\ n\Delta t\to t \\ j\Delta x\to x}} U_j^n = u(x,t). \qquad (2.1\text{-}5)$$

The existence of the above limit means that given x, t and $\varepsilon > 0$, we can find a $\delta > 0$ such that if $\Delta x < \delta$, $|n\Delta t-t| < \delta$, and $|j\Delta x-x| < \delta$, then $|U_j^n - u_j^n| < \epsilon$. In this case, δ is independent of x and t, δ depends only on ε. Therefore, we have uniform convergence. However, we do not have, strictly speaking, an error estimate. Even though $|e_j^n| \leq C\Delta x^2$, $0 < j < J$, $n\Delta t < T$, we are unable to evaluate C without knowledge of the solution u(x,t). What we have is an asympotic error estimate.

That is, we have

$$U_j^n = u(x_j,t_n) + O(\Delta x^2) \qquad 0 < j < J, \quad n\Delta t \le T.$$

Note: we write $f(x) = O(x)$ if there is a constant C such that $|f(x)| \le C|x|$ for sufficiently small x; that is, we say f(x) is of order x at zero.

We have demonstrated the convergence of our difference scheme by use of a maximum principle argument (which works only because the coefficients in our finite difference scheme are positive). Next, we will prove convergence by use of the Fourier representation of the solution of the difference scheme. This will give us considerably more insight into the behavior of the difference scheme.

In chapter 1, we obtained a solution of the heat equation (2.1-1) by separation of variables. We might briefly review this. If $u(x,0) = f(x) = a_k \sin \pi kx$ then we look for a solution in the form

$$u(x,t) = A_k(t) \sin \pi kx.$$

By substitution into the heat equation, we obtain an ordinary differential equation for $A_k(t)$, namely

$$\frac{dA_k}{dt} = -\sigma\pi^2 k^2 A_k$$

$$A_k(0) = a_k$$

The solution of this equation is

$$A_k(t) = a_k e^{-\sigma\pi^2 k^2 t}$$

and thus

$$u(x,t) = a_k e^{-\sigma\pi^2 k^2 t} \sin \pi kx.$$

The heat equation is linear. Thus, if v(x,t) and w(x,t) are solutions, then u = Av + Bw is also a solution for arbitrary constants A and B. Therefore, if

$$u(x,0) = f(x) = \sum_{k=1}^{m} a_k \sin \pi kx \qquad \text{then}$$

$$u(x,t) = \sum_{k=1}^{m} a_k e^{-\sigma\pi^2 k^2 t} \sin \pi kx \qquad \text{is}$$

the desired solution. If we make the above sum infinite, then we must be a little careful, since we must be able to interchange the summation and differentiation, that is

$$\frac{\partial}{\partial t} \sum_{k=1}^{\infty} = \sum_{k=1}^{\infty} \frac{\partial}{\partial t} .$$

This is certainly possible if $t > 0$. Thus, we have a solution for our problem in terms of a Fourier expansion.

We will now obtain a solution of the difference equation (2.1-3) by the same technique. Suppose $U_j^0 = a_k \sin \pi kx$. Assume $U_j^n = A_k(n) \sin \pi kx$, then we must solve for the sequence of values $A_k(n)$. Note that

$$\sin \pi k x_{j+1} - 2 \sin \pi k x_j + \sin \pi k x_{j-1} = -4 \sin^2 \frac{\pi k}{2J} \sin \pi k x_j$$

Therefore, if we substitute our expression for U_j^n into the difference scheme (2.1-3) we obtain

$$A_k(n+1) = \left(1 - 4\mu \sin^2 \frac{k\pi}{2J}\right) A_k(n)$$

$$A_k(0) = a_k$$

If we make the definition

$$M(k) = 1 - 4\mu \sin^2 \frac{k\pi}{2J}$$

then the solution of the above equation for $A_k(n)$ is

$$A_k(n) = [M(k)]^n a_k .$$

That is, raise the constant $M(k)$ to the n^{th} power to obtain the solution on the n^{th} time level. Note that $A_k(n) = M(k)A_k(n-1) = [M(k)]^2 A_k(n-2) \ldots = M(k)^n A_k(0)$. Therefore, the solution of the difference scheme is (for $U_j^0 = f(x_j) = a_k \sin \pi k x_j$)

$$U_j^n = a_k [M(k)]^n \sin \pi k x_j$$

Thus, we may regard M(k) as an amplification factor. It depends on the frequency k of our Fourier mode. It also depends on the mesh ratio $\mu = \sigma\Delta t/\Delta x^2$ and on the mesh spacing $\Delta x = 1/J$.

If $f(x) = \sum_{k=1}^{\infty} a_k \sin \pi kx$ then

we can use the linearity of our finite difference scheme to write the solution as

$$U_j^n = \sum_{k=1}^{\infty} a_k \left[M(k)\right]^n \sin k\pi x_j$$

Since our amplification factors are uniformly bounded

$|M(k)| \le |1+4\mu|$ and since we have assumed $\sum_{k=1}^{\infty} |a_k| < \infty$ we know that this series will converge. Therefore U_j^n is a solution. Since a finite difference operator will always commute with an infinite sum, unlike a differential operator, we do not have the same concern that we had with the heat equation itself.

Now we are ready to consider the convergence of our difference scheme. We first look at the case of a single mode that is $f(x) = a_k \sin \pi kx$. We need the following result.

Problem 2.1-4. Let f(x) be a complex valued function of the real argument x such that

$$\lim_{x\to 0} f(x) = a$$

then

$$\lim_{\substack{x\to 0 \\ nx\to t}} (1+xf(x))^n = e^{at}.$$

We need to consider the behavior of the solution U_j^n as Δx approaches zero. We assume the ratio $\mu = \sigma\Delta t/\Delta x^2$ is a constant. Then Δt is determined by Δx and $\Delta t\to 0$ as $\Delta x\to 0$. Note that $\Delta x = 1/J$.

Problem 2.1-5. Show that our scheme converges for a single mode, that is

$$\lim_{\substack{\Delta x\to 0 \\ n\Delta t\to t \\ j\Delta x\to x}} U_j^n = u(x,t).$$

The meaning of this multivariable limit is hopefully clear. It was defined for equation (2.1-5). Note that

$$U_j^n = a_k \left[1 - 4\mu \sin^2 \frac{\pi k\Delta x}{2}\right]^n \sin \pi k x_j$$

$$u(x,t) = a_k e^{-\sigma\pi^2 k^2 t} \sin \pi k x.$$

Now for an interpretation of this result. The above problem shows that we have convergence regardless of the value of μ. Note that for small enough values of $k\Delta x$ we have

$$1 - 4\mu \sin^2 \frac{\pi k\Delta x}{2} \approx 1 - \sigma\pi^2 k^2 \Delta t.$$

Therefore, the amplification factor $M(k)$ is less than one, if Δx is sufficiently small (note that k is fixed). This statement holds regardless of the value of μ. Our solution has the form

$$U_j^n = a_k \left[M(k)\right]^n \sin \pi k x_j$$

and therefore $U_j^n \to 0$ as $n \to \infty$ for a fixed k and fixed Δx provided Δx is sufficiently small. Thus, the solution of the difference scheme has the right behavior as the discrete time n approaches infinity - the solution decays to zero. Now suppose $\mu > \frac{1}{2}$. Then for a fixed but small Δx we can always find a frequency k such that $|M(k)| > 1$, simply take k so that $\sin(\pi k \Delta x/2)$ is close to one. But then $\left[M(k)\right]^n$ approaches infinity as n becomes large; the mode for this value of k will grow rather than decay. If there are only a finite number of modes present in the solution, then we can always take Δx small enough so that $M(k)$ will be less than one for all these modes. Then the solution of the difference equation behaves in a reasonable way; it decays to zero as n approaches infinity. However, if $\mu > \frac{1}{2}$ and an infinite number of modes are present in the solution (that is $f(x) = \sum_1^\infty a_k \sin k\pi x$ and $a_k \neq 0$ for infinitely many k) then some of these modes will grow, no matter how small we take Δx. We can no longer base our argument on the behavior of a single mode. We cannot expect convergence in this case, since there will always be some modes which are growing at an exponential rate. Note that if $\mu \leq \frac{1}{2}$ then $|M(k)| \leq 1$ for all values of k.

We have here the notion of stability, which is basic to the use of finite difference schemes. We say a scheme is stable, if the solution U^n remains bounded independent of the mesh spacing Δx. That is, there is a constant C such that $\|U^n\| \le C\,\|U^0\|$ for $n\Delta t \le T$ independent of U^0 and Δx. Here $\|U^n\|$ denotes some suitable measure of the "size" of U^n, such as its Euclidean length $\sqrt{\sum_{j=0}^{J} |U_j^n|^2}$ or its largest component $\operatorname{Max}_j |U_j^n|$. We will have much more to say about this concept of stability later. Another way to look at this concept is to note that a small perturbation in the initial values U^0 should produce a correspondingly small perturbation in the solution U^n. For the present, we simply note that under suitable restrictions, a stable scheme is always convergent.

We will let the reader prove that our scheme is convergent if $\mu < \frac{1}{2}$ and will diverge, at least for some f(x), if $\mu > \frac{1}{2}$.

Problem 2.1-6. If $\mu = .55$, $J = 100$ and $|a_k| > 0$ for all k, then find the lowest frequency mode which will amplify under the difference scheme (2.1-3).

Problem 2.1-7. If $\mu < \frac{1}{2}$ and

$$f(x) = \sum_{k=1}^{\infty} a_k \sin \pi kx \qquad \text{where } \sum_{1}^{\infty} |a_k| < \infty$$

then prove convergence, that is

$$\lim_{\substack{\Delta x \to 0 \\ n\Delta t \to t \\ j\Delta x \to x}} U_j^n = u(x,t).$$

Hint. We have

$$U_j^n = \sum_{k=1}^{\infty} a_k \left[M(k)\right]^n \sin \pi kx \qquad M(k) = 1 - 4\mu \sin^2 \frac{\pi \Delta x k}{2}$$

Show that if $\mu < \frac{1}{2}$, then the above series converges uniformly.

Then note that

$$\lim_{\Delta x\to 0} U_j^n = \sum_{k=1}^{\infty} \lim_{\Delta x\to 0} a_k \left[M(k) \right]^n \sin \pi kx$$

$$= \sum_{k=1}^{\infty} a_k e^{-\sigma\pi^2 k^2 t} \sin \pi kx = u(x,t).$$

This problem provides an excellent illustration of the necessity for care in the use of the calculus. We really **need the uniform** convergence. If we assume that we can interchange the limit and the infinite sum $\lim_{\Delta x\to 0} \sum_{k=1}^{\infty} = \sum_{k=1}^{\infty} \lim_{\Delta x\to 0}$, then our difference scheme is convergent regardless of the value of μ. This would be an erroneous conclusion.

Problem 2.1-8. In problem 1.2-37 we have shown that for integers k and m

$$\sum_{j=0}^{J} \sin \pi kx_j \sin \pi mx_j = \begin{cases} \frac{J}{2} & \text{if } k = m+2rJ \\ & \text{where r is an integer} \\ 0 & \text{otherwise} \end{cases}$$

Remember that $x_j = 1/J$. Now assume

$$g(x) = \sum_{1}^{\infty} a_k \sin \pi kx, \quad \sum_{1}^{\infty} |a_k| < \infty$$

Then show

$$\frac{1}{J} \sum_{j=0}^{J} g^2(x_j) = \tfrac{1}{2} \sum_{k=1}^{\infty} \sum_{\substack{r=-\infty \\ k+2rJ\geq 1}}^{\infty} a_k\, a_{k+2rJ}$$

The solution of the difference scheme is given by

$$U_j^n = \sum_{k=1}^{\infty} a_k M^n(k) \sin \pi k x_j \qquad M(k) = 1 - 4\mu \sin^2 \frac{\pi k \Delta x}{2}$$

Show that

$$\frac{1}{J} \sum_{j=0}^{J} \left(U_j^n \right)^2 = \frac{1}{2} \sum_{k=1}^{\infty} \sum_{\substack{r=-\infty \\ k+2rJ\geq 1}}^{\infty} M^n(k) a_k a_{k+2rJ}$$

Problem 2.1-9. Assume that $\mu > \frac{1}{2}$. Assume that initial function f(x) is given by

$$f(x) = \sum_{k=1}^{\infty} a_k \sin k\pi x \qquad \text{where } a_k = \frac{1}{k^p}$$

with p an integer $p \geq 2$. Then show that the finite difference scheme does not converge uniformly. Hint: If U_j^n converges uniformly to u(x,t), then show that $\frac{1}{J} \sum \left(U_j^n\right)^2$ must be bounded independent of n and Δx. Then use problem 2.1-8 to show that this sum is not bounded if $\mu > \frac{1}{2}$.

To illustrate this problem of stability, we programmed the scheme of equations (2.1-3) for the computer. We used the initial function $u(x,0) = 4\,x(1-x)$ and 41 points in the mesh (J=40). We plot the vector U^n for the values of $T = n\Delta t$ shown on the graphs. The graphs were drawn by the computer (a CDC 6600) using a microfilm recorder (the CDC dd80). We made two runs, one for $\mu = .45$ and one for $\mu = .55$. The instability is quite obvious. It is also quite mysterious, if one does not go through the type of analysis that we have just completed. The instability has nothing to do with the physics of the problem from which the heat equation might have been derived. It is a property of the finite difference scheme and thus cannot be explained by looking back to the physical origin of the problem. (See Fig. 2.1-1 and 2.1-2.)

The study of this finite difference scheme gives us a good idea of the relation between stability and convergence. Let us assume that

$$f(x) = \sum_1^\infty a_k \sin \pi kx \qquad \text{where } \sum_1^\infty |a_k| < \infty .$$

We will assume that infinitely many modes are present; that is, given any k_0 we can find $k > k_0$ such that $a_k \neq 0$. If $k\Delta x$ is small, that is $k\Delta x \ll 1$, then a finite difference approximation to the second derivative of $\sin \pi kx$ will be a good approximation. Thus, for a given single mode $\sin \pi kx$, we might expect the finite difference solution to converge to the solution of the heat equation. This is exactly what happens, regardless of the value of μ, as we have shown in problem 2.1-5. However, we have infinitely many modes present. At a given Δx there are always some modes for which $a_k \neq 0$ and $k\Delta x$ is large. Since $k\Delta x$ is large, we cannot expect the finite difference approximation of the derivatives to be valid for these modes. However, the series $\sum_{k=1}^\infty |a_k|$ is convergent, therefore, $\lim_{k\to\infty} \sum_{k=K}^\infty |a_k| \to 0$. If we take Δx small enough the lower frequency modes ($k\Delta x \ll 1$) will be accurately represented by

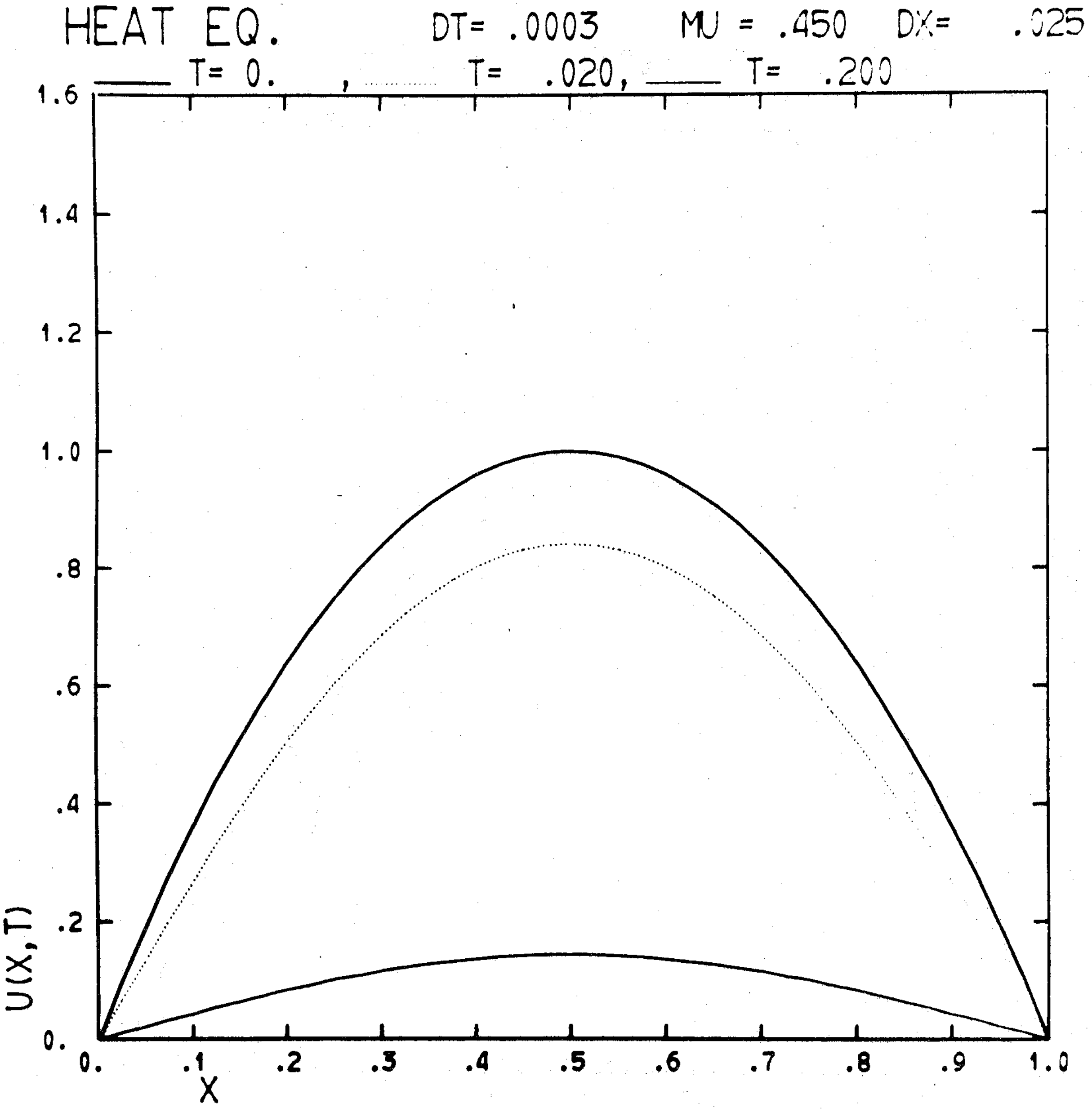

HEAT EQ.
DT= .0003
MU = .450
DX= .025
T= 0.
T= .020,
T= .200
U(X,T)
X
1.6
1.4
1.2
1.0
.8
.6
.4
.2
0.
0.
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0

Figure 2.1-1

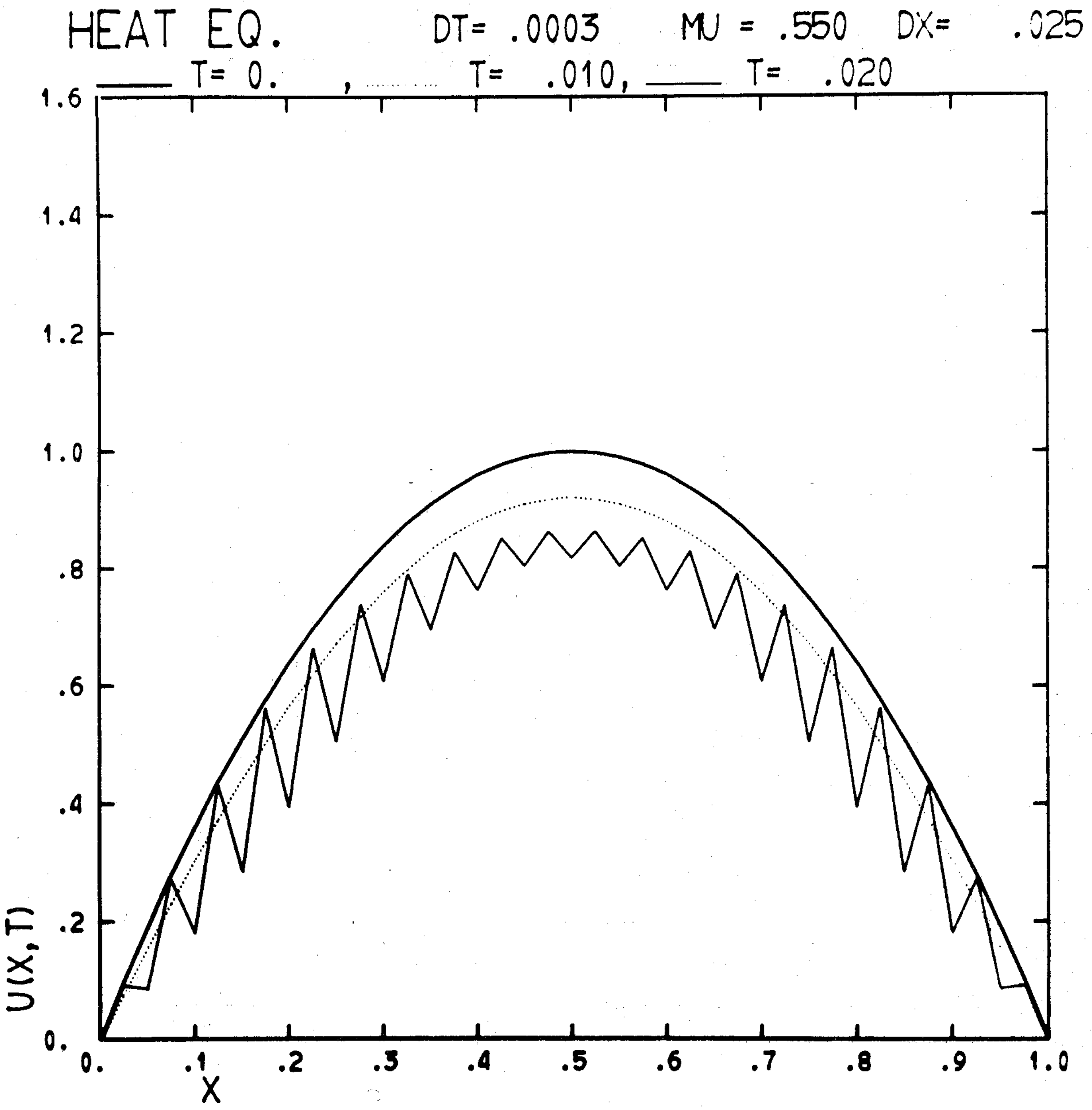

HEAT EQ.
DT= .0003
MU = .550
DX= .025
T= 0. ,
T= .010,
T= .020
1.6
1.4
1.2
1.0
.8
.6
.4
.2
0.
U(X,T)
0.
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
X

Figure 2.1-2

the difference scheme. If $\sum_{k=K}^{\infty} |a_k| \ll 1$, then the error due to inaccurate treatment of these higher modes will be small, simply because their total contribution to the solution is small. However, this is true only if the finite difference scheme does not permit these higher modes to grow. If the mode in the finite difference solution is given by $\left[M(k)\right]^n a_k$ and $|M(k)|>1$ then this mode can become large even if a_k is small. Also, this mode is not accurately treated by the difference scheme, if $k\Delta x$ is not small. Therefore, the error contribution from this mode can become large. This leads us to the requirement of stability; namely, $\|U^n\| \le M \|U^0\|$ where M is a constant independent of the mesh spacing and independent of the initial vector U^0. A stable scheme permits only a modest growth in any mode. The stable difference scheme may not accurately represent the given mode, but at least there is no exponential growth of the mode.

As we have seen, an unstable scheme may converge for some initial functions f(x) (see problem 2.1-5 and problem 2.2-8). However, an unstable scheme is useless in practice, even for these initial functions. An unstable scheme will amplify roundoff error on a computer. This error is small initially, but it will grow exponentially and eventually ruin the solution. Note that roundoff error tends to be high frequency and will therefore usually amplify rapidly with an unstable scheme.

2.2 Difference Schemes for a Hyperbolic Equation

Here we will consider the following simple hyperbolic differential equation:

$$\frac{\partial u}{\partial t} + c\,\frac{\partial u}{\partial x} = 0 \qquad u = u(x,t) \qquad (2.2\text{-}1)$$

$$-1 \le x \le 1 \qquad 0 \le t$$

$$u(x,0) = f(x)$$

$$u(-1,t) = u(1,t)$$

We have imposed periodic boundary conditions $u(-1,t) = u(1,t)$. This is usually unreasonable from a physical point of view, but it does simplify our analysis of the difference scheme. If the function $f(x)$ is defined for all $x(-\infty < x < \infty)$, periodic ($f(x+2) = f(x)$ for all x), and differentiable, then the solution of this problem is $u(x,t) = f(x-ct)$.

Now we will consider two difference schemes for this problem. The mesh is defined by the points x_j where $-J \le j \le J$, $x_j = j\Delta x$, $\Delta x = 1/J$. The values of $u(x,t)$ are sought at the points (x_j, t_n) where $t_n = n\Delta t$. The solution of the finite difference scheme is denoted by U_j^n and is an approximation to $u(x_j,t_n)$. Just as for the heat equation, we use a capital U to denote the solution of the difference equation and a lower case u to denote the solution of the differential equation.

Our experience with the heat equation suggests use of the same scheme for the hyperbolic equation; namely, a forward difference in time and a centered difference in space.

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + c\,\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = 0 \qquad -J \le j \le J \qquad (2.2\text{-}2)$$

$$U_j^0 = f(x_j)$$

$$U_{j\pm 2J}^n = U_j^n \quad \text{for } n \ge 0 \text{ and } -J \le j \le J$$

As we will now explain, the last equation is an expression of the periodic boundary conditions. If we look at the above scheme for $j = -J$, we see that the value U_{-J-1}^n is required for the approximation of the spatial derivative. This value is an approximation to $u(x_{-J-1}, t_n)$, but the point $x_{-J-1} = (-J-1)\Delta x = -1 - \Delta x$ lies outside the mesh interval. To resolve this problem we use periodicity; that is, we assume $u(x\pm 2,t) = u(x,t)$. In the finite difference scheme this becomes $U_{j\pm 2J}^n = U_j^n$, for all j. Thus we have $U_{-J-1}^n = U_{J-1}^n$ and x_{J-1} does lie in the mesh interval. Similarily, we let $U_{J+1}^n = U_{-J+1}^n$ and then the difference scheme (2.2-2) is defined for $-J \le j \le J$. Note that we need compute U_j^n only for $-J \le j < J$ since the periodicity condition gives us $U_J^n = U_{-J}^n$.

Problem 2.2-1. Assume that f(x) is periodic, then $U_{-J}^0 = U_J^0$. Suppose we compute U_j^n from equation (2.2-2) for $-J \le j \le J$ (then we do not use the condition $U_{-J}^n = U_J^n$). Next, suppose we compute U_j^n from equations (2.2-2) for $-J \le j < J$ and use the condition $U_{-J}^n = U_J^n$. Do we get the same result? Or to phrase the question differently, are we being consistent in our treatment of the boundary condition?

We will show (problem 2.4-3) that the "forward-centered" scheme given by equations (2.2-2) is not stable, and therefore it is useless. We will now describe a stable scheme for this problem.

$$\frac{U_j^{n+1} - \frac{U_{j+1}^n + U_{j-1}^n}{2}}{\Delta t} + c\,\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = 0 \qquad (2.2\text{-}3)$$

$$U_j^0 = f(x_j)$$

$$U_{j\pm 2J}^n = U_j^n \qquad -J \le j \le J$$

The only difference is in the approximation of the time derivative. Instead of using U_j^n, we have used $\frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right)$.

We may write the first scheme as

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad (2.2\text{-}4)$$

and the second as

$$U_j^{n+1} = \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad (2.2\text{-}5)$$

where $\lambda = c\Delta t/\Delta x$. We will assume that the mesh ratio λ is held constant during an integration, just as $\mu = \sigma\Delta t/(\Delta x)^2$ was held constant for the heat equation. Note that the above equations clearly show that if we are given the vector $\left\{U_j^0\right\}$, then we can obtain the vector $\left\{U_j^n\right\}$ for any n simply by marching forward starting from U^0. We will need the truncation error for these schemes. Remember that the truncation error τ_j^n is obtained by substitution of the solution $u(x,t)$ into equations (2.2-4) or (2.2-5) and equating the remainder to $\Delta t\tau_j^n$.

Problem 2.2-2. Assume that the solution of equation (2.2-1) has continuous 3rd order partial derivatives. Suppose

$$\left|\frac{\partial^2 u}{\partial t^2}\right| \le M_1, \quad \left|\frac{\partial^2 u}{\partial x^2}\right| \le M_2, \quad \left|\frac{\partial^3 u}{\partial x^3}\right| \le M_3 \text{ for } |x| \le 1, \ t \le T.$$

Then show that a bound for the truncation error for equation (2.2-4) is

$$|\tau_j^n| \le \frac{M_1}{2} \Delta t + \frac{|c| M_3}{6} \Delta x^2$$

and for equation (2.2-5)

$$|\tau_j^n| \le \frac{M_1}{2} \Delta t + \frac{|c|}{2\lambda} M_2 \Delta x + \frac{|c| M_3}{6} \Delta x^2$$

We can use the maximum principle to prove that the scheme given by equations (2.2-3) converges. If we rearrange (2.2-3) we have

$$U_j^{n+1} = \alpha U_{j-1}^n + (1-\alpha) U_{j+1}^n \qquad \alpha = \frac{1}{2} + \frac{\lambda}{2}$$

$$u_j^{n+1} = \alpha u_{j-1}^n + (1-\alpha) u_{j+1}^n + \Delta t \tau_j^n .$$

Problem 2.2-3. Assume that the truncation error τ_j^n in the above equation is bounded, $|\tau_j^n| \le \tau$ for all j and for $n\Delta t \le T$. Using the maximum principle, show that the error $\varepsilon^{(n)} = \text{Max}|u_j^n - U_j^n|$ is bounded by $\varepsilon^{(n)} \le t_n \tau$, $t_n \le T$.

Note that this result implies convergence, since problem 2.2-2 shows $\tau = 0(\Delta x)$ and thus $\lim\limits_{\Delta x \to 0} \tau = 0$.

Next we will prove convergence for the scheme by use of a Fourier expansion just as we did for the heat equation.

Given any sufficiently respectable function f(x) $-1 \le x \le 1$ we can expand it in a Fourier series

$$f(x) = \sum_{k=-\infty}^{\infty} a_k T^{ik\pi x}$$

where $a_k = \frac{1}{2} \int_{-1}^{1} f(x) e^{-ik\pi x}$

If f(x) is real valued then $a_{-k} = \bar{a}_k$ where $\bar{a}_k$ denotes the complex conjugate of a_k.

We will assume that our function f(x) is smooth enough to insure the series $\sum_{k=-\infty}^{\infty} k|a_k|$ converges (that is $\sum_{-\infty}^{\infty} k|a_k| < \infty$) although we may not always need this.

Next we assume that U^0 is given by a single Fourier mode, that is

$$U_j^0 = a_k e^{i\pi k x_j}, \; -J \le j \le J.$$

The solution for this simple case will provide considerable insight into the general case.

Problem 2.2-4. Let U_j^n be a solution of equation 2.2-3 (or equation 2.2-2) where $U_j^0 = f(x_j)$ is given by $U_j^0 = a_k e^{i\pi k x_j}$. Show that U_j^n is given by

$$U_j^n = a_k (M_k)^n \, e^{i\pi k x_j}$$

where $M_k = \cos\pi k\Delta x - i\lambda \sin\pi k\Delta x$ for equation (2.2-3) and $M_k = 1 - i\lambda \sin \pi k \Delta x$ for equation (2.2-2).

Next we will show that the scheme given by equations (2.2-3) converges.

Problem 2.2-5. Let $M_k = \cos \pi k \Delta x - i\lambda \sin \pi k \Delta x$ for the scheme (2.2-3). Assume $|\lambda| < 1$ ($\lambda = c\Delta t/\Delta x$). Then show $|M_k| \leq 1$.

Problem 2.2-6. Assume that λ is a constant. Then show that

$$\lim_{\substack{\Delta x \to 0 \\ n\Delta t \to t}} \left(M_k(\Delta x)\right)^n = e^{-i\pi kct}$$

Note that $\Delta x = 1/J$ and since J is an integer Δx cannot take on arbitrary real values in the finite difference equation. We could write the above limit in terms of J and thereby stay closer to the difference scheme. Note that if we fix Δx, then Δt is fixed by $\lambda = c\Delta t/\Delta x$. Thus the requirement $n\Delta t \to t$ merely fixes the rate at which n goes to infinity as $\Delta x \to 0$. We could simply set n equal to the integer part of $t/\Delta t = tc/(\lambda \Delta x) = Jtc/\lambda$.

Problem 2.2-7. If $|\lambda| < 1$, then show that scheme 2.2-3 is convergent, that is

$$\lim_{\substack{\Delta x \to 0 \\ n\Delta t \to t \\ x_j \to x}} U_j^n = \sum_{k=-\infty}^{\infty} a_k e^{-ik\pi(x-ct)} = u(x,t)$$

Problem 2.2-8. Let the initial function $f(x)$ be given by

$$f(x) = \sum_{-\infty}^{\infty} a_k e^{ik\pi x} \qquad a_k = 1/k!$$

and let λ have any fixed nonzero value. Show that the scheme (2.2-2) converges for this $f(x)$.

2.3 Representation of a Finite Difference Scheme by a Matrix Operator

In this section we will study finite difference operators using concepts from matrix theory, especially the norm of a matrix. To simplify our notation, we will frequently replace Δx by h and Δt by k. This is a standard practice in the literature on difference schemes.

We will first consider the following difference scheme for the heat equation which we studied in section 2.1.

$$U_j^{n+1} = U_j^n + \mu(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \quad \mu = \sigma k/h^2 \quad 0 \le j \le J$$

$$U_0^n = U_j^n = 0 \qquad U_j^0 = f(x_j)$$

If we denote the vector $\left\{U_j^n\right\}$ $1 \le j \le J-1$ by U^n, then we may write the above equations in matrix form

$$U^{n+1} = L_h U^n$$

where the matrix L_h is given by

$$L_h = \left|\begin{array}{cccccc} 1-2\mu & \mu & 0 & & & \\ \mu & 1-2\mu & \mu & & & \\ 0 & \mu & 1-2\mu & \mu & 0 & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & \cdots & \cdots & 0 & \mu & 1-2\mu \end{array}\right| \qquad (2.3\text{-}1)$$

Note that L_h is a symmetric tridiagonal matrix. We use the subscript h in the notation L_h for the finite difference matrix, because the order of the matrix depends on h ($h = 1/J$ and the order is $J-1$). For some difference schemes the terms of the matrix will depend explicitly on h; here we assume μ is a constant independent of h.

Next we might look at the two schemes we have used for the simple hyperbolic equation $U_t + CU_x = 0$. The first is equation (2.2-2).

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad \lambda = c\Delta t/\Delta x$$

$$-J \leq j < J$$

$$U_{j+2J}^n = U_j^n$$

The matrix equation is $U^{n+1} = L_h U^n$. The matrix L_h and vector U^n are given by

$$L_h = \left|\begin{array}{cccccc} 1 & -\frac{\lambda}{2} & 0 & & & \frac{\lambda}{2} \\ \frac{\lambda}{2} & 1 & -\frac{\lambda}{2} & 0 & & 0 \\ \vdots & & & & & \\ -\frac{\lambda}{2} & \cdots\cdots & & 0 & \frac{\lambda}{2} & 1 \end{array}\right| , \quad U^n = \left|\begin{array}{c} U_{-J}^n \\ U_{-J+2}^n \\ \\ \\ U_{J-1}^n \end{array}\right|$$

The other scheme for this equation is

$$U_j^{n+1} = \tfrac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right)$$

and the matrix is $(\alpha = \frac{1}{2} + \frac{\lambda}{2})$

$$L_h = \begin{vmatrix} 0 & \alpha & 0 & & & & 1-\alpha \\ 1-\alpha & 0 & \alpha & 0 & & & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \alpha & 0 & \cdots & \cdots & \cdots & 1-\alpha & 0 \end{vmatrix} , U^n = \begin{vmatrix} U^n_{-J} \\ U^n_{-J+2} \\ \\ \\ U^n_{J-1} \end{vmatrix}$$

Next we will recapitulate the properties of a matrix norm. First the vector norm. Given a vector (possibly complex) $u = \{u_j\}$ $1 \le j \le J$ we define the L_2 norm by $||u||_2 = \sqrt{\sum_{j=1}^{J} |U_j|^2}$, the maximum norm by $\max_{1\le j\le J} |u_j|$, and the L_1 norm by $||u||_1 = \sum_{j=0}^{J} |u_j|$. These norms all provide a measure of the "size" or "length" of u. Given any vector norm, we can define an induced matrix norm. If A is the matrix, then the norm of A is defined by

$$||A|| = \max_{||u||\neq 0} \frac{||Au||}{||u||} = \max_{||u||=1} ||Au||$$

The norm used on the right is the vector norm from which the matrix norm is induced. The matrix norm is a measure of the maximum expansion caused by the mapping Au of the vector space into itself. Important inequalities for the matrix norm are the following (A and B are square matrices, u a vector)

$$||A + B|| \le ||A|| + ||B||$$

$$||AB|| \le ||A|| \; ||B||$$

$$||Au|| \le ||A|| \; ||u||$$

There is an obvious relation between the eigenvalues $\lambda_j(A)$ of a matrix A and the value of its norm $\|A\|$, namely, $|\lambda_j(A)| \le \|A\|$. We define the spectral radius $\sigma(A)$ of a matrix A to be the maximum modules of its eigenvalue,

$$\sigma(A) = \max_j |\lambda_j(A)|$$

$$\text{then} \qquad \sigma(A) \le \|A\|.$$

The three matrix norms mentioned above can be characterized by

$$\|A\|_2 = \sqrt{\sigma(A^T A)} \qquad A^T \text{ denotes the transpose of } A$$

$$\|A\|_\infty = \max_i \sum_{j=1}^{J} |a_{ij}|$$

$$\|A\|_1 = \max_j \sum_{i=1}^{J} |a_{ij}|$$

For a symmetric matrix $\|A\| = \sigma(A)$.

Next we will study the matrix which defines our finite difference scheme for the heat equation. This is the symmetric, tridiagonal matrix given in equation (2.3-1). Note that this matrix L_h has order J-1. The mesh points are $x_j = jh$, $0 \le j \le J$, $h = \Delta x = 1/J$, $\mu = \sigma\Delta t/\Delta x^2 = \sigma k/h^2$.

Problem 2.3-1. Verify that the eigenvalues of L_h (equation 2.3-1) are $\lambda_r = 1-4\mu\sin^2(\pi rh/2)$ and the eigenvectors $\{U^r_j\} = \sin \pi r x_j$ where $1 \le r \le J-1$ and $1 \le j \le J-1$. Show that

$$\|A\|_2 = \begin{cases} 1-4\mu\sin^2\frac{\pi h}{2} & \mu < \frac{1}{2} \\ 4\mu\sin^2\frac{\pi(J-1)}{2J} & \mu > \frac{1}{2} \end{cases}$$

$$\|A\|_\infty = \begin{cases} 1 & \mu \le \frac{1}{2} \\ 4\mu-1 & \mu > \frac{1}{2} \end{cases}$$

Now we are ready to talk about stability. First the definition: We assume that we have a finite difference scheme represented by a matrix operator L_h. We are given the initial vector U_h^0 and succeeding vectors are defined by $U_h^{n+1} = L_h U_h^n$. Here we use the subscript h to denote the fact that the vectors and the matrix depend on the mesh spacing. As h approaches zero, the order of the matrix approaches infinity. We are not working in a vector space of fixed dimension. We say the scheme $U_h^{n+1} = L_h U_h^n$ is stable provided there is a constant M such that $\|U_h^n\| < M\,\|U_h^0\|$ for all U_h^0, all $h > 0$, and all n provided $n\Delta t < T$. The constant M must be independent of U_h^0, h and n. However, M may depend on the time limit T (we require $n\Delta t = t_n < T$). Since $U_h^n = L_h^n\, U_h^0$ our stability requirement is the same as placing a bound on the power of L_h. We could require a constant M, such that $\|L_h^n\| < M$ for $n\Delta t < T$.

Why is this concept of stability so important? One reason is the Lax-Richtmyer theorem which states that a stable scheme with sufficiently small truncation error is a convergent scheme. We will consider this theorem in section 2.5. Stability has another very important consequence .

It insures that a difference scheme is not unduly sensitive to small perturbations. For example, the effect of roundoff error is that of a small perturbation on the finite difference calculation. Suppose we let U^n denote the solution of the unperturbed equation $U^{n+1} = L_h U^n$. We let V^n denote the perturbed solution. What do we mean by a perturbation? For one thing, we could have a perturbation E^0 in the initial data, that is $V^0 = U^0 + E^0$. Let ϵ^0 denote the magnitude of the perturbation. Then $\epsilon^0 = \|U^0 - V^0\| = \|E^0\|$. We might also assume that we have a perturbation at each stage of the solution denoted by ρ^n, thus $V^{n+1} = L_h V^n + \rho^n$, $V^0 = U^0 + E^0$. For example, roundoff error creates such a perturbation as we will see in section 2.6. Suppose we let $\epsilon^n = \|V^n - U^n\|$ denote the error resulting from this perturbation. We also suppose that we know an upper bound for these perturbations; thus $\|\rho^n\| \le \rho$ for some constant ρ. We would like to obtain an estimate for the error ϵ^n in terms of ϵ^0 and ρ. We have the following equations

$$U^n = L_h U^{n-1}$$

$$V^n = L_h V^{n-1} + \rho^{n-1}$$

$$V^n - U^n = L_h (V^{n-1} - U^{n-1}) + \rho^{n-1}$$

Therefore, if we let $E^n = V^n - U^n$,

$$E^n = L_h E^{n-1} + \rho^{n-1}$$

$$E^{n-1} = L_h E^{n-2} + \rho^{n-2}$$

If we combine these equations we obtain

$$E^n = L_h^n E^{n-2} + L_h \rho^{n-2} + \rho^{n-1}$$

If we continue this process by induction we obtain

$$E^n = L_h^n E^0 + L_h^{n-1} \rho^0 + L_h^{n-2} \rho^1 + L_h^{n-3} \rho^2 + \dots + \rho^{n-1}$$

Now if we take the norm of both sides and use the fact that $\|AB\| \le \|A\| \, \|B\|$ and $\|A + B\| \le \|A\| + \|B\|$ we obtain

$$\epsilon^n \le \|L_h^n\| \epsilon^0 + \|L_h^{n-1}\| \, \|\rho^0\| + \|L_h^{n-2}\| \, \|\rho^1\| + \dots + \|\rho^{n-1}\|$$

Now suppose the scheme is stable, then $\|L_h^n\| \le M$ for all n. Also, by assumption $\|\rho^n\| \le \rho$.

Therefore: $\epsilon^n \le M\epsilon^0 + nM\rho$

This is our desired estimate of the error in terms of ϵ^0 and ρ. For this estimate to be useful, ρ must be small enough so that $n\rho$ does not become large; for example, ρ might depend on Δt such that $\rho \le k\Delta t$ for some k. Then if $n\Delta t < T$, $n\rho < KT$.

If the scheme is not stable, then $L_h^{n-1} \rho^0$ may become large. For example, if $\|L_h\| \ge 1.2$, then growth at the rate $1.2^{(n-1)}$ is possible. It is worth pointing out that n frequently exceeds 10^4 in some initial value problems and $(1.2)^{200} \sim 10^{16}$. In section 2.6, we show an example of the disastrous growth of roundoff error for a unstable scheme.

What we have demonstrated is that a stable scheme is not unduly sensitive to a small perturbation - hence, the name stable. Note our essential use of the concept of norm, not only for proofs, but for the definition of stability. This concept is a valuable tool for the understanding of stability.

We terminate this section with some problems to illustrate the concept of a matrix norm.

Problem 2.3-2. If $||L_h|| \le 1 + k\Delta t$, then $||L_h^n|| \le e^{kT}$ provided $n\Delta t \le k$.

Problem 2.3-3. Find a matrix A whose spectral radius is unity, $\sigma(A) = 1$, but such that $||A^n|| \ge n$.

2.4 Analysis of Schemes for Linear Problems with Constant Coefficients and Periodic Boundary Conditions - the Use of Finite Fourier Analysis.

We will restrict ourselves to initial value problems with a single unknown function $u(x,t)$ on the interval $-1 \le x \le 1$. The initial function is $u(x,0) = f(x)$. We assume the problem is periodic $u(x \pm 2,t) = u(x,t)$, $f(x \pm 2) = f(x)$. The mesh is the set of points x_j, $-J \le j \le J$, $x_j = jh$, $h = 1/J$. We assume the periodicity condition $U^n_{j\pm 2J} = U^n_j$. We assume the finite difference scheme can be written in the form given below.

$$U^{n+1}_j = \sum_{\nu=1}^{s} C_\nu U^n_{j+j_\nu} \qquad (2.4\text{-}1)$$

In the case of the scheme defined by equations (2.2-4); $s = 3$, $C_1 = 1$, $j_1 = 0$, $C_2 = -\lambda/2$, $j_2 = 1$, $C_3 = \lambda/2$, $j_3 = -1$. For the scheme defined by eqn(2.2-5); $s = 2$, $C_1 = 1/2 - \lambda/2$, $j_1 = 1$, $C_2 = 1/2 + \lambda/2$, $j_2 = -1$.

Some of the values $j + j_\nu$ may lie outside the allowable range $-J \le j+j_\nu < J$. We invoke the periodicity condition to bring these values back into range. For example, consider the finite difference scheme defined by equation (2.2-3).

$$U^{n+1}_j = \alpha U^n_{j-1} + (1-\alpha)U^n_{j+1} \qquad -J \le j < J.$$

When $j = -J$ the value $U^n_{j-1} = U^n_{-J-1}$ is outside the range. But

$U^n_{-J-1} = U^n_{-J-1+2J} = U^n_{J-1}$ which lies inside the range. Thus, the equation above for $j = -J$ becomes

$$U^{n+1}_{-J} = \alpha U^n_{J-1} + (1-\alpha)U^n_{-J+1}$$

We will need to use a finite Fourier series to represent our mesh functions U_j^n. Given a smooth function f(x), $-1 \le x \le 1$, it can be represented by an infinite Fourier series

$$f(x) = \sum_{-\infty}^{\infty} a_k e^{i\pi kx}$$

We are dealing with discrete functions U_j^n $-J \le j < J$ or vectors U^n. We will define a set of vectors $\varphi^m = \{\varphi_j^m\}$ $-J \le j < J$, $0 \le m < 2J$ such that given any vector $\{U_j\}$ there is a set of coefficients a_m (these may be complex) such that

$$U_j = \sum_{m=0}^{2J-1} a_m \varphi_j^m \tag{2.4-2}$$

Furthermore, the following orthogonality condition holds

$$\sum_{j=-J}^{J-1} \varphi_j^n \overline{\varphi}_j^m = \begin{cases} 0 & n \neq m \\ 2J & n = m \end{cases} \tag{2.4-3}$$

The coefficients a_m are given by

$$a_m = \frac{1}{2J} \sum_{j=-J}^{J-1} U_j \overline{\varphi}_j^m \tag{2.4-4}$$

Thus the set $\{\varphi^m\}$ of vectors forms an orthonormal basis for the 2J dimensional complex vector space in which our vectors $\{u^n\}$ lie.

Problem 2.4-1. Let φ^m be defined by

$$\varphi_j^m = e^{i\pi m x_j} \qquad x_j = j/J.$$

Show that the orthogonality condition of equation (2.4-3) holds. Show that if the coefficients a_m are given by equation (2.4-4), then equation (2.4-2) is true. Also, show that equation (2.4-2) implies equation (2.4-4).

Hint: $\sum_{j=0}^{2J-1} e^{i\pi mj/J} = \sum_{j=0}^{2J-1} Z^j = \frac{1-Z^{2J}}{1-Z} = 0$ if $Z \neq 1$.

Problem 2.4-2. Let $\|U\|$ denote the L_2 norm of the vector U

$$\|U\|^2 = \frac{1}{2J} \sum_{j=-J}^{J-1} |U_j|^2$$

Suppose the Fourier representation for U is $U = \sum_{m=0}^{2J-1} a_m \varphi^m$

Then show $\|U\|^2 = \frac{1}{2J} \sum_{m=0}^{2J-1} |a_m|^2$.

We can use the finite Fourier representation to study the stability of the difference scheme given in equation (2.4-1). We first assume that U^n is equal to one of our Fourier modes φ^m

$$U_j^n = \varphi_j^m = e^{i\pi m x_j} \qquad -J \leq j < J$$

Then U^{n+1} is equal to this same Fourier mode multiplied by a complex constant which we call an amplification factor

$$U_j^n = \left(\sum_{\nu=1}^{s} c_\nu e^{i\pi m h j_\nu} \right) e^{i\pi m x_j} = A_h(m)\, \varphi_j^m \qquad (2.4\text{-}5)$$

Note that $x_{j+j_\nu} = x_j + j_\nu h$. The reader should verify this formula by substitution into the difference scheme of equation (2.4-1).

For example, consider the difference scheme

$$U_j^{n+1} = \alpha U_{j-1}^n + (1-\alpha) U_{j+1}^n$$

The amplification factor for this scheme has already been determined in section 2.2 and is given by

$$A_h(m) = \cos\pi m h + i(1-2\alpha)\sin\pi m h.$$

Now suppose U^n is given by the sum $\sum_{m=0}^{2J-1} a_m \varphi^m$ where $\varphi_j^m = e^{i\pi m x_j}$.

Then $U_j^{n+1} = \sum_{\nu=1}^{s} c_\nu U_{j+j_\nu}^n = \sum_{\nu=1}^{s} c_\nu \sum_{m=0}^{2J-1} a_m \varphi_{j+j_\nu}^m = \sum_{m=0}^{2J-1} a_m \sum_{\nu=1}^{s} c_\nu \varphi_{j+j_\nu}^m$

$$U_j^{n+1} = \sum_{m=0}^{2J-1} a_m A_h(m) \varphi_j^m$$

Thus, to go from the n^{th} to the $n+1^{th}$ level, we simply multiply the amplitude a_m of each mode by the amplification factor for that mode. Thus, if the initial vector is given by $U^0 = \sum_{m=0}^{2J-1} a_m \varphi^m$

then the expansion for U^n is

$$U^n = \sum_{m=0}^{2J-1} a_m \left[A_h(m)\right]^n \varphi^m$$

Now suppose that the amplification factors satisfy the so-called "von Neumann" condition, that is, there is a constant k independent of h and m, such that

$$|A_h(m)| \leq 1 + k\Delta t$$

We assume that there is a constant mesh ratio such as $\mu = \sigma\Delta t/\Delta x^2$ or $\lambda = c\Delta t/\Delta x$ $(h = \Delta x)$ so that Δt is a function of h and $\lim_{h\to 0} \Delta t = 0$. Given this von Neumann condition, we know from problem 2.3-2 that

$$\left|\left[A_h(m)\right]^n\right| \leq e^{kT} \quad \text{where} \quad n\Delta t \leq T.$$

Then from problem 2.4-2 we have

$$\|U^n\|^2 = \frac{1}{2J} \sum_{m=0}^{2J-1} |a_m|^2 \left|\left[A_h(m)\right]^{2n}\right| \leq e^{kT} \frac{1}{2J} \sum_{m=0}^{2J-1} |a_m|^2 = e^{kT}\|U^0\|$$

Therfore, the von Neumann condition implies that our difference scheme is stable.

Suppose there is a real number ρ such that $\rho > 1$ and for any $h > 0$, we can find an integer m, $0 \leq m < 2J$, $J \leq \frac{1}{h}$, and $|A_h(m)| > \rho$. That is, we can always find an amplification factor greater than ρ no matter how small we choose h. Then it is clear that the scheme cannot be stable.

For given any h let m be chosen so that $|A_h(m)| > \rho$. Then let $U_j^0 = \varphi_j^m = e^{i\pi m x_j}$. Let n_h be the largest integer, such that $n_h \Delta t < 1$. Then $U^{n_h} = \left[A_h(m)\right]^{n_h} \varphi^m$ and clearly

$$||U^{n_h}|| \geq \rho^{n_h} ||U^0||.$$

But since $n_h \to \infty$ as $h \to 0$, and $\rho > 1$, $\rho^{n_h} \to \infty$ as $h \to 0$. Therefore, the scheme can not be stable since stability would imply

$$||U^n|| \leq M \, ||U^0|| \quad \text{for some } M.$$

Problem 2.4-3. Show that the difference scheme given by equation (2.2-2) is unstable. Use the Fourier analysis, or von Neumann method.

Problem 2.4-4. Show that the difference scheme given by equation (2.2-3) is stable provided $|\lambda| < 1$ $(\lambda = c\Delta t/\Delta x)$ and unstable if $|\lambda| > 1$. Note that our proof of convergence for the scheme of (2.2-3) very nearly provides the solution for this problem. We merely have to switch from the infinite Fourier series to the finite Fourier expansion.

The von Neumann method for the determination of stability is very important in the design of finite difference schemes for initial value problems. We will use it constantly in the remainder of our discussion of initial value problems. We will next give a few problems which review the methods for the analysis of stability.

The first problem uses the maximum principle (which is really an "energy method" - see chapter 3); the second is almost identical with the first except we work with the norm of the matrix operator; and the third uses the von Neumann analysis.

Problem 2.4-5. Consider the simple hyperbolic equation

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0,\ 0 \le x \le 1,\ 0 \le t,\ c > 0.$$

The initial and boundary conditions are $u(x,0) = f(x)$, $u(0,t) = g(t)$, $g(0) = f(0)$. The solution of this problem is $f(x-ct)$ for $x \ge ct$ and $g(t - x/c)$ for $x \le ct$ (see section 1.3.6). Consider the following difference scheme for this problem.

$$x_j = jh,\ 0 \le j \le J,\ h = 1/J$$

$$U_j^{n+1} = U_j^n - \lambda(U_j^n - U_{j-1}^n) \qquad \lambda = \frac{c\Delta t}{\Delta x} \quad 1 \le j \le J$$

$$U_0^n = g(t_n) \quad U_j^0 = f(x_j) \quad 0 \le j \le J.$$

By use of the maximum principle, prove that this scheme is stable provided $\lambda < 1$.

Problem 2.4-6. For the scheme given in problem 2.4-5, write out the matrix L_h defined by the scheme

$$U^{n+1} = L_h U^n \qquad\qquad U^n = \{U_j^n\} \ 1 \le j \le J.$$

Show that if $\lambda < 1$ then $\|L_h\|_\infty = 1$ where this norm is the one induced by the maximum norm. What are the eigenvalues of L_h. What can you say about $\|L_h\|_2$ (the norm induced by the L_2 vector norm). Can you prove stability using this norm.

Problem 2.4-7. Suppose we modify the above problem to have periodic boundary conditions.

$$U_j^{n+1} = U_j^n - \lambda(U_j^n - U_{j-1}^n) \qquad -J \le j < J,\ x_j = jh.$$

$$U_{j\pm 2J}^n = U_j^n$$

$$U_j^0 = f(x_j)$$

We must then solve for U_j^n, $0 \le j < J$. Use the von Neumann method to analyze the stability of this scheme. Note that we were forced to make the boundary conditions periodic in order to prove stability with the von Neumann method.

2.5 The Relation Between Stability and Convergence - the Lax-Richtmyer Theorem.

We will restrict our discussion to linear initial value problems of the form

$$\frac{\partial u}{\partial t} = L(u) + g \qquad u = u(x,t) \qquad g = g(x,t) \qquad u(x,0) = f(x)$$

We assume u is defined over some region in space. Space may be more than one dimensional, for example, a plane or a cube in which case x is a vector, $x = (x_1, x_2, x_3)$. We assume a discrete mesh is imbedded in our region. The points of this mesh are denoted by $x_j(h)$, or just x_j. The parameter h is used to denote the mesh. We will be a little loose about exactly how to define h, in some sense h must determine the mesh and as h approaches zero, the mesh spacing must also approach zero. We will let u_j^n denote the value of u at the j^{th} mesh point on the n^{th} time level. Then $\{u_j^n\}$ where n is fixed and j ranges over the mesh is a finite dimensional vector. Our finite difference scheme can be represented as a family of matrix operators

$$U^{n+1} = L_h U^n + G_h$$

where U^n and G_h are vectors. We have assumed, of course, that our finite difference scheme is linear. In many cases, perhaps most, the problems one puts on a computer are non-linear. We will assume that L_h does not depend on the time level n. This can happen only if the original differential operator L in the equation

$$\frac{\partial u}{\partial t} = Lu + g$$

is independent of the time t. Next we will give some examples to indicate the diversity of problems which can be placed in this framework.

We have already discussed the heat equation on the interval $0 \le x \le 1$, namely

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2}$$

We assume that the values of u on the boundary are given by $u(0,t) = A_0(t)$, $u(1,t) = A_1(t)$ where A_0 and A_1 are known functions. Here $x_j = jh$, $0 \le j \le J$, $h = 1/J$. The order of the matrix L_h is J-1 and we use the same finite difference scheme as in section 2.1

$$L_h = \begin{vmatrix} 1-2\mu & \mu & 0 & & & & \\ \mu & 1-2\mu & \mu & 0 & & & \\ 0 & \mu & 1-2\mu & \mu & 0 & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & \ldots\ldots\ldots & & & 0 & \mu & 1-2\mu \end{vmatrix}, U^n = \begin{vmatrix} U_1^n \\ \\ \\ \\ \\ U_{J-1}^n \end{vmatrix}, G^n = \begin{vmatrix} \mu A_0(t_n) \\ 0 \\ 0 \\ \\ 0 \\ \mu A_1(t_n) \end{vmatrix}$$

where $\mu = \sigma \Delta t/\Delta x^2$. Note that we have picked up an inhomogenous term G^n because of the boundary conditions. The difference equation centered at $j = 1$ is $U_1^{n+1} = U_1^n + \mu(U_2^n - 2U_1^n + U_0^n) = \mu U_2^n + (1-2\mu)U_1^n + \mu A_0(t_n)$ since $U_0^n = A_0(t_n)$.

The heat equation on a square leads to a similar matrix equation. The differential equation is

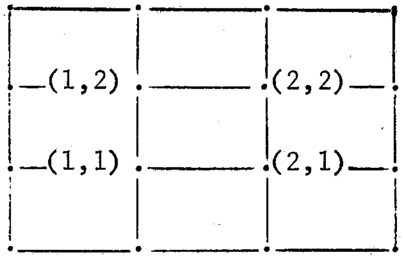
$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \qquad u = u(x,y,t) \qquad u(x,y,0) = f(x,y)$$

$$0 \le x \le 1, \; 0 \le y \le 1$$

We assume that the boundary conditions require u to vanish on the sides of the square. The finite difference scheme is basically the same as in the one dimensional case. We let $x_j = j/J$, $y_k = k/K$, $0 \le j \le J$, $0 \le k \le K$, $u^n_{jk} = u(x_j, y_k, t_n)$. Then the difference scheme is

$$U^{n+1}_{jk} = U^n_{jk} + \mu_x\left[U^n_{j+1,k} - 2U^n_{jk} + U^n_{j-1,k}\right] + \mu_y\left[U^n_{j,k+1} - 2U^n_{jk} + U^n_{j,k-1}\right]$$

where $\mu_x = \Delta t/\Delta x^2$, $\mu_y = \Delta t/\Delta y^2$. Usually we have used a single index j to label our mesh points. Here we have used two indexes, j and k. We could use a single index r to label the mesh, that is $(x,y)_r = (x_j, y_k)$ where $r = j + k(J+1) + 1$. Thus $1 \le r \le (J+1)(K+1)$ since $0 \le j \le J$ and $0 \le k \le K$. We would use a somewhat different algorithm to obtain a single index for the unknown values U_{jk}, since we know the boundary values and thus the range is $1 \le j \le J-1$, $1 \le k \le K-1$. We would define our single index r by $r = j + (k-1)(J-1)$. Therefore, $1 \le r \le (J-1)(K-1)$. The ordering of the components of the vector U^n given below is simply an ordering by increasing values of this index r. We form the vector U^n by ordering the terms U^n_{jk} varying first j then k. Consider the case $J = K = 3$ for which there are only four unknowns, the other values of u_{jk} being zero because of the boundary conditions. The mesh and difference scheme are then



$$U^n = \left| \begin{matrix} U^n_{11} \\ U^n_{21} \\ U^n_{12} \\ U^n_{22} \end{matrix} \right|, \; L_h = \left| \begin{matrix} 1-2\mu_x-2\mu_y & \mu_x & \mu_y & 0 \\ \mu_x & 1-2\mu_x-2\mu_y & 0 & \mu_y \\ \mu_y & 0 & 1-2\mu_x-2\mu_y & \mu_x \\ 0 & \mu_y & \mu_x & 1-2\mu_x-2\mu_y \end{matrix} \right|, \quad G^n = 0$$

In general L_h is a block tridiagonal matrix of order (J-1) x (K-1) of the following form

$$U^n = \left| \begin{matrix} U^n_{11} \\ U_{21} \\ U_{J-1,1} \\ U_{12} \\ U_{J-1,2} \\ U_{1K-1} \\ U_{J-1,K-1} \end{matrix} \right| \qquad L_h = \left| \begin{matrix} C & D & 0 & & & \\ D & C & D & & & \\ 0 & D & C & D & & \\ & & & & & \\ & & 0 & D & C & D \\ & & & 0 & D & C \\ & & & & & \end{matrix} \right|$$

The matrices C and D are square matrices of order J-1 given below ($\alpha = 1-2\mu_x - 2\mu_y$)

$$C = \left|\begin{array}{cccccc} \alpha & \mu_x & 0 & & & \\ \mu_x & \alpha & \mu_x & 0 & & \\ 0 & \mu_x & \alpha & \mu_x & & \\ & & & & & \\ & & 0 & \mu_x & \alpha & \mu_x \\ & & & 0 & \mu_x & \alpha \end{array}\right| , \quad D = \left|\begin{array}{cccc} \mu_y & 0 & & \\ 0 & \mu_y & 0 & \\ & & & \\ & & & \\ & & 0 & \mu_y \end{array}\right| = \mu_y I$$

Problem 2.5-1. Compute the eigenvalues, eigenvectors and the "L_2" norm of L_h, $||L_h||_2$. Also, compute the maximum norm $||L_h||_\infty$. What condition on μ_x and μ_y will insure stability of this difference scheme. Hint: Try eigenvectors of the form $W_{jk} = \sin \pi r x_j \sin \pi s y_k$, $1 \le r \le J-1$, $1 \le s \le K-1$.

A system of equations can be treated in the same fashion. For example, consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \qquad \begin{array}{l} u = u(x,t) \\ -1 \le x \le 1 \end{array} \qquad (2.5\text{-}1)$$

The initial conditions are

$$u(x,0) = f_1(x)$$

$$\frac{\partial u}{\partial t}(x,0) = f_2(x)$$

We assume periodic boundary conditions

$$u(x+2,t) = u(x,t)$$

$$f_i(x+2) = f_i(x) \qquad 1 \le i \le 2.$$

This is not in the form

$$\frac{\partial u}{\partial t} = L(u)$$

since we have a second order time derivative. However, the wave equation is equivalent to the system of equations

$$\frac{\partial v}{\partial t} = c\,\frac{\partial w}{\partial x} \qquad v(x,0) = f_1(x)$$

$$\frac{\partial w}{\partial t} = c\,\frac{\partial v}{\partial x} \qquad w(x,0) = f_3(x) = \frac{1}{c}\int_{-1}^{x} f_2(\tau)\,d\tau$$

A difference scheme for this system is

$$V_j^{n+1} = \tfrac{1}{2}(V_{j+1}^n + V_{j-1}^n) + \frac{\lambda}{2}(W_{j+1}^n - W_{j-1}^n) \qquad \Delta x = 1/J$$

$$-J \le j < J$$

$$W_j^{n+1} = \tfrac{1}{2}(W_{j+1}^n + W_{j-1}^n) + \frac{\lambda}{2}(V_{j+1}^n - V_{j-1}^n) \qquad \lambda = c\Delta t/\Delta x$$

Let the vector U^n be defined as below, then the matrix L_h for this scheme is as given below. The order of the matrix is 4J.

$$U^n = \begin{vmatrix} V^n_{-J} \\ W^n_{-J} \\ V^n_{-J+1} \\ W^n_{-J+1} \\ \vdots \\ V^n_{J-1} \\ W^n_{J-1} \end{vmatrix} \qquad L_h = \begin{vmatrix} 0 & 0 & \frac{1}{2} & \frac{\lambda}{2} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{2} & -\frac{\lambda}{2} \\ 0 & 0 & \frac{\lambda}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & & 0 & 0 & -\frac{\lambda}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{\lambda}{2} & 0 & 0 & \frac{1}{2} & \frac{\lambda}{2} & 0 & 0 & & 0 & 0 & 0 & 0 \\ -\frac{\lambda}{2} & \frac{1}{2} & 0 & 0 & \frac{\lambda}{2} & \frac{1}{2} & 0 & 0 & & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & & & & & & \vdots \\ \frac{1}{2} & \frac{\lambda}{2} & 0 & 0 & & & & & & \frac{1}{2} & -\frac{\lambda}{2} & 0 & 0 \\ \frac{\lambda}{2} & \frac{1}{2} & 0 & 0 & \cdots & & & & & -\frac{\lambda}{2} & \frac{1}{2} & 0 & 0 \end{vmatrix}$$

Problem 2.5-2. For the above difference scheme determine the eigenvectors and eigenvalues of the matrix L_h. Also, compute the maximum norm of L_h, $\|L_h\|_\infty$.

We will obtain yet another example of a finite difference scheme by dealing directly with the wave equation.

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

We will approximate the second order derivatives directly to obtain the following difference scheme

$$U_j^{n+1} - 2U_j^n + U_j^{n-1} = \lambda^2 (U_{j+1}^n - 2U_j^n + U_{j-1}^n)$$

$$-J \le j < J, \ \lambda = c\Delta t/\Delta x$$

We will assume periodic boundary conditions as before. We can write this scheme in the following matrix form

$$U^{n+1} = \hat{L}_h U^n - U^{n-1} \quad \text{where } \hat{L}_h \text{ is a matrix of order } 2J \qquad (2.5\text{-}2)$$

Note that we have 3 time levels involved here, instead of two. In order to compute U^{n+1}, we must already know U^n and U^{n-1}. Therefore, in order to start this scheme we must know the values of U^0 and U^1. These values can be obtained from the initial conditions in equations (2.5-1) as follows.

$$U_j^0 = f_1(x_j) \quad \text{since } u(x,0) = f_1(x) \qquad (2.5\text{-}3)$$

and $$U_j^1 = U_j^0 + \Delta t f_2(x_j) \quad \text{since } \frac{\partial u}{\partial t}(x,0) = f_2(x)$$

Equation (2.5-2) is not in the form $U^{n+1} = L_h U^n$. Our theory of difference scheme will apply only to schemes which involve two time levels, n+1 and n. However, we can change variables and write equations (2.5-2) in the two level form. Let W^n, $1 \leq n$, be a vector of order 4J defined as follows

$$W^n = \left| \begin{matrix} U^n_{-J} \\ U^n_{-J+1} \\ \vdots \\ U^n_{J-1} \\ U^{n-1}_{-J} \\ U^{n-1}_{-J+1} \\ \vdots \\ U^{n-1}_{J-1} \end{matrix} \right|$$

We may write W^n as the composite of two vectors U^n and U^{n-1} each of order 2J, that is

$$W^n = \left| \begin{matrix} U^n \\ \\ U^{n-1} \end{matrix} \right|$$

Then the difference scheme given by equation (2.5-2) can be written

$$W^{n+1} = \left| \begin{matrix} U^{n+1} \\ \\ U^n \end{matrix} \right| = \left| \begin{matrix} \hat{L}_h U^n - U^{n-1} \\ \\ U^n \end{matrix} \right| = \left| \begin{matrix} \hat{L}_h & -I \\ \\ I & 0 \end{matrix} \right| \left| \begin{matrix} U^n \\ \\ U^{n-1} \end{matrix} \right| = L_h W^n$$

where L_h is a matrix of order 4J formed by the submatrices $\hat{L}_h$ and I which are of order 2J.

$$L_h = \left| \begin{matrix} \hat{L}_h & -I \\ \\ I & 0 \end{matrix} \right|$$

If we write this matrix out explicitly we obtain

$$L_h = \left| \begin{array}{cccccccccccc} \alpha & \lambda^2 & 0 & \cdots & \cdots & -\lambda^2 & -1 & 0 & & & & 0 \\ -\lambda^2 & \alpha & \lambda^2 & 0 & \cdots & 0 & 0 & -1 & 0 & & & 0 \\ 0 & -\lambda^2 & \alpha & \lambda^2 & \cdots & 0 & 0 & 0 & -1 & 0 & \cdots & 0 \\ \vdots & & & & & & & & & & & \\ \lambda^2 & 0 & \cdots & \cdots & 0 & -\lambda^2 & \alpha & 0 & \cdots & \cdots & \cdots & -1 \\ 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & & & \\ \vdots & & & & & & & & & & & \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 1 & 0 & \cdots & \cdots & \cdots & 0 \end{array} \right|$$

Note that W^1 is known since U^1 and U^0 can be determined from the boundary conditions, as shown in equation (2.5-3). Thus, we have a marching scheme

$$W^{n+1} = L_h W^n \qquad 1 \le n$$

which falls within our standard format, except the vector W now approximates the solution u(x,t) at two time levels.

Problem 2.5-3. Find the eigenvectors and eigenvalues of the matrix $\hat{L}_h$ given in equation (2.5-2). Show that these eigenvalues λ lie in the range $-2 \le \lambda \le 2$.

Problem 2.5-4. Let λ_k $1 \le k \le 2J$ be the eigenvalues of the matrix $\hat{L}_h$ of equation (2.5-2). Let L_h be the composite matrix defined above, that is

$$L_h = \begin{vmatrix} \hat{L}_h & -I \\ I & 0 \end{vmatrix} .$$

Show that the eigenvalues of L_h are α_+ and α_-, each repeated 2J times where α_+ and α_- are the roots of $\alpha^2 - \lambda_k + 1 = 0$,

or $\alpha_\pm = \frac{\lambda}{2} \pm \sqrt{\frac{\lambda^2-4}{2}}$. Is the scheme $W^{n+1} = L_h W^n$ stable?

Hint: The eigenvalues λ_k of $\hat{L}_h$ are distinct and thus have linearly independent eigenvector $U^{(k)}$. Show that the eigenvectors of L_h are

$$W_{\pm}^{(k)} = \begin{vmatrix} \alpha_{\pm} U^{(k)} \\ \\ U^{(k)} \end{vmatrix} \qquad \alpha_{\pm} = \frac{\lambda}{2} \pm \frac{\sqrt{\lambda^2-4}}{2}$$

The purpose of the above examples is to show that a wide variety of finite difference schemes fit into the format that we have been discussing. Now we are ready to move on to a slightly more formal definition of the concepts of truncation error, stability and convergence.

We assume that we have an initial value problem represented by a partial differential equation defined over some spatial domain. We assume (without loss of generality) that the initial conditions are given at time $t = 0$. Usually, we will require our initial value problem to be of the form (first order in time)

$$\frac{\partial u}{\partial t} = L(u) + g$$

$$u(x,0) = f(x)$$

Here the point x is restricted to lie in some region of space (space may be multidimensional, thus $x = (x_1, x_2, x_3, \ldots x_n)$. The operator L is formed of partial derivatives with respect to the spatial variables. We require L to be linear, that is $L(a_1u_1 + a_2u_2) = a_1L(u_1) + a_2 L(u_2)$. Many practical problems are not linear and the nonlinearity can cause considerable difficulty. In addition, there will usually be some boundary conditions imposed on the solution u. We will not attempt to give a precise definition of an initial value problem. The reader can refer to the book by Richtmyer and Morton for this. We will be a little vague about specification of boundary conditions and also the number of continuous derivatives we require of our solution u.

We will assume that a finite mesh is laid down on our domain. Actually, we have a family of such meshes, each mesh being labeled by the value of a parameter h. We compute an approximation U_j^n to our solution at these mesh points x_j. The vectors U^n are computed by a marching procedure from the relation

$$U^{n+1} = L_h U^n + G^n$$

The operator L_h is a matrix. The vectors G^n are known functions which may depend on h but not on U^n. The starting value U^0 is obtained from the initial conditions. We will assume the time step Δt is a function of the parameter h. We also assume that

$$\lim_{h \to 0} \Delta t = 0$$

Definition 2.5-1. Truncation Error The truncation error associated with a solution u of the differential equation is obtained by substitution of u into the difference scheme. We let τ_h^n denote the truncation error. We use the subscript h because τ_h^n is a family of vectors, one vector for each value of h. The components of τ_h^n are $\tau_{h,j}^n$ corresponding to the mesh points x_j. The truncation error is defined by

$$u^{n+1} = L_h u^n + G^n + \Delta t \tau_h^n$$

where $u_j^n = u(x_j, t_n)$ is the vector defined by the solution u.

Definition 2.1-2. A Consistent Difference Scheme We say our scheme is consistent if for all sufficiently smooth solutions u of the differential equation the truncation error τ_h^n approaches zero with h. By sufficiently smooth we will usually mean the solution must have all its derivatives continuous up to a certain order. The estimates of truncation error will usually use a Taylor series expansion which requires certain derivatives to be continuous. Using the maximum norm, we can state our requirement on τ as follows. Given $\varepsilon > 0$ and $T > 0$, there is a $\delta > 0$ such that $\|\tau_h^n\|_\infty < \varepsilon$ for all $h < \delta$ provided $n\Delta t < T$. This means that for each mesh point x_j $\|\tau_{h,j}^n\| < \varepsilon$. Note that we are requiring τ to approach zero uniformly over the mesh and also uniformly in time.

Definition 2.5-3. A Stable Scheme We say a scheme is stable if there is a constant M such that $\|L_h^n\| < M$ if $n\Delta t < T$. Note that this must hold for all h and n provided $n\Delta t < T$. We have not specified the norm here. Usually, we will use the "L_2" or Euclidian norm $\|L_h^n\|_2$; however, we may use any norm. For example, we might use the maximum norm $\|L_h^n\|_\infty$. A scheme may be stable in one norm and unstable in another. [Stetter]

Definition 2.5-4. A Convergent Scheme We say a finite difference scheme is convergent if for all sufficiently smooth solutions u of the differential equation the corresponding solution (one with the same initial and boundary conditions) of the finite difference scheme converges to this solution u.

That is

$$\lim_{\substack{h\to 0\\ x_j\to x\\ t_n\to t}} \|U_j^n - u(x,t)\| = 0$$

This definition also leaves open the specification of the norm. We require the limit to be uniform, relative to x and t, that is for any $\epsilon > 0$ and $T > 0$, there is a $\delta > 0$ such that

$$\|U_j^n - u(x,t)\| < \epsilon$$

provided $h < \delta, |x_j - x| < \delta, |t_n - t| < \delta, t \leq T$. This must hold independent of x and t provided $t \leq T$.

Next, we treat a fundamental result - the Lax-Richtmyer theorem. This theorem tells us that stability and convergence are really the same property.

Theorem 2.5-1. If a consistent finite difference scheme is stable, then it is convergent. The converse is also true (convergence implies stability) although we will not offer a proof (see the book by Richtmyer and Morton).

The proof goes as follows. Suppose we choose an initial function f(x) and let u(x,t) be the solution of the differential equation for this initial function ($u(x,0) = f(x)$). Let U_h^n be the finite difference scheme corresponding to f(x), thus

$$U_h^{n+1} = L_h U_h^n + G_h^n$$

$$U_h^0 = f$$

This means that for a given mesh (denoted by h) we have a vector $U_{h,j}^n$, $0 \le j \le J$, $U_{h,j}^0 = f(x_j)$. We will assume that the solution u of the differential equation is in the class for which the consistency condition holds. For u in this class the truncation error approaches zero as the mesh parameter h goes to zero. Convergence only holds for initial functions f, such that the corresponding function u lies in this class. If we let τ_h^n denote the truncation error then we have

$$U^{n+1} = L_h U_h^n + G_h^n$$

$$u^{n+1} = L_h u^n + G_h^n + \Delta t \tau_h^n$$

Now we let e_h^n denote the error on a particular mesh, that is

$$e_{h,j}^n = u_j^n - U_{h,j}^n = u(x_j, t_n) - U_{h,j}^n.$$

Then, by combining the above equations we have

$$e_h^{n+1} = L_h e_h^n + \Delta t \tau_h^n$$

Note that $u_j^0 = f(x_j) = U_{h,j}^n$, thus $e_h^0 \equiv 0$. Also, note that we have made essential use of the linearity of the matrix operator L_h in deriving the expression for the error vector e_h^n. By using the above equation recursively, starting with n = 0 we obtain

$$e_h^1 = L_h e_h^0 + \Delta t \tau_h^0$$

$$e_h^2 = L_h e_h^1 + \Delta t \tau_h^1 = L_h^2 e_h^0 + \Delta t L_h \tau_h^0 + \Delta t \tau_h^1$$

$$e_h^3 = L_h e_h^2 + \Delta t \tau_h^2 = L_h^3 e_h^0 + \Delta t \left[L_h^2 \tau_h^0 + L_h \tau_h^1 + \tau_h^2 \right]$$

It is clear that the general formula is (note that $e_h^0 \equiv 0$)

$$e_h^n = \Delta t \left[L_h^{n-1} \tau_h^0 + L_h^{n-2} \tau_h^1 + \ldots + \tau_h^{n-1} \right]$$

Now using the properties of the norm $\|AB\| \le \|A\| \, \|B\|$, $\|A+B\| \le \|A\| + \|B\|$, we obtain

$$\|e_h^n\| \le \Delta t \|L_h^{n-1}\| \, \|\tau_n^0\| + \|L_h^{n-2}\| \, \|\tau_n^1\| + \ldots + \|\tau_h^{n-1}\|$$

Since our scheme is consistent, we know that we can make $\|\tau_h^k\|$ small if we make h sufficiently small. We first choose a time limit T. Then for any $\epsilon > 0$

$$\|\tau_h^k\| < \epsilon$$

provided $h < \delta$ and $k\Delta t \le T$. Since our scheme is stable

$$\|L_h^k\| \le M \qquad \text{if } k\Delta t \le T.$$

Therefore our inequality for e_h^n becomes

$$\|e_h^n\| \le \Delta t M n \epsilon$$

But for $n\Delta t \leq T$,

$$\|e_h^n\| \leq MT\epsilon$$

We can rewrite this as

$$\|U_h^n - u(x_j,t_n)\| \leq MT\epsilon$$

provided $h < \delta$, $n\Delta t \leq T$.

Then we have

$$\|U_h^n - u(x,t)\| \leq \|U_h^n - u(x_j,t_n)\| + \|u(x_j,t_n) - u(x,t)\|$$

$$\|U_h^n - u(x,t)\| \leq MT\epsilon + \|u(x_j,t_n) - u(x,t)\|$$

This inequality makes it clear that

$$\lim_{\substack{h\to 0\\ x_j\to x\\ t_n\to t}} U_{h,j}^n = u(x,t)$$

Therefore we have convergence.

2.6 The Relation Between Stability and the Growth of Roundoff Error.

Roundoff error is caused by the finite word length on a computer (60 bits on the Control Data 6600). Truncation error occurs when derivatives are replaced by finite differences. It would be better to call this discretization error, as does Henrici [1962], since it is caused by a discrete approximation to a continuous problem. For finite difference solutions to partial differential equations, the roundoff error is usually much smaller than the truncation error, thus roundoff is usually no problem (on a machine with a 48- or 60-bit word length). In this section we will estimate the roundoff error for the finite difference approximation to the heat equation which is described in section 2.1. First we will discuss the roundoff error as it occurs in the basic arithmetic operations on a computer.

Most computers store numbers as a sequence of bits; that is, a sequence of zeros or ones. Then a number may be represented in the form

$$2^S \times \sum_{k=1}^{t} x_k \, 2^{k-1}$$

where x_k is either zero or one and S is an integer. On the Control Data 6600 $t = 48$ and $-2^{11} < S < 2^{11}$. Instead of stating our analysis for a binary machine, we will assume we have a decimal machine where the numbers are stored in the form

$$10^S \times \sum_{k=1}^{t} x_k \, 10^{(k-1)}$$

where x_k is an integer $0 \le x_k \le 9$.

The theory is the same for a binary machine, but it is more difficult to describe the binary case. We will first assume we have a 4-digit machine with a 2-digit exponent. Some sample numbers would thus be $(t = 4, \quad -99 \le S \le 99)$

1.0 = 0001(+0)
0.1 = 0001(-1)
1.25 = 0125(-2)
0.0001414 = 1414(-7)
123400 = 1234(+2) .

Now consider the error in addition, subtraction, multiplication, and division. We assume our machine has an "accumulator register" of length 2t (eight digits in our case). The arithmetic operations are to be done in this register. For example, to add a = 12.12 = 1212(-2) and b = .3456 = 3456(-4) we would first place the larger number a in the register, left adjusted so that 4 zeros are added to the right. We then have in the accumulator

12120000(-6)

We then shift the decimal point in b so that its exponent matches the accumulator, thus b = 00345600(-6). We then add these representations of a and b to obtain an 8-digit number.

12465600(-6)

In order to store this number we must reduce it to 4 digits. We denote the result of an exact addition by a+b. The result of our computer addition we denote by $f\ell(a+b)$ which stands for the floating point sum of a and b. We obtain $f\ell(a+b)$ by reducing the 8-digit number to 4 digits, rounding if necessary. In the case above we thus have $f\ell(a+b) = 1247(-2)$. We have not stated exactly how our computer performs arithmetic operations, nor do we intend to. We merely want to make the following estimates plausible. We assume that the roundoff error is such that the following relations hold. These relations will be true for any computer, although we might have to enlarge the upper bound for φ somewhat and change to a binary or hexadecimal representation (then we might have $f\ell(a\pm b) = (a\pm b)(1 + \varphi 2^{-t})$ with $|\varphi| \le \frac{1}{2}$). If our machine had an accumulator of length t, we would have to change the first relation to

$$f\ell(a\pm b) = a(1 + \varphi_1\, 10^{-t}) + b(1 + \varphi_2\, 10^{-t})$$

where $|\varphi_i| \le 5$.

Our assumed bounds for the roundoff error are the following:

$$f\ell(a\pm b) = (a\pm b)(1 + \varphi 10^{-t})$$

$$f\ell(ab) = ab(1 + \varphi 10^{-t}) \qquad |\varphi| \le 5$$

$$f\ell(a/b) = (a/b)(1 + \varphi 10^{-t})$$

For a more detailed discussion of rounding error, see the books and papers by Wilkinson [1963].

Now we are ready to consider the roundoff error in our finite difference solution of the heat equation. We let U^n denote the exact solution of the finite difference equation, starting with $U_j^0 = f(x_j)$, $1 \le j \le J-1$, then

$$U_j^{n+1} = U_j^n + \mu \left(U_{j+1}^n - 2U_j^n + U_{j-1}^n \right) \qquad 1 \le j \le J-1$$

$$U_0^n = U_J^n = 0 \ , \qquad \mu = \sigma \Delta t / \Delta x^2 \ , \qquad \Delta x = 1/J \ .$$

We let V^n be the solution obtained by using floating point arithmetic on our computer. There may be some difference between U^0 and V^0 because of errors made in the evaluation of $f(x_j)$. The value of V^n will depend on the order in which the arithmetic operations are done on the computer. However, our estimate of the error (the difference between U^n and V^n) will be independent of this order. We first look at the result of an exact computation of $D_e = V_{j+1}^n - 2V_j^n$ and the computer floating point computation $D_a = f\ell(V_{j+1}^n - 2V_j^n)$ (we really should write this $f\ell(V_{j+1}^n - f\ell(2V_j^n))$ but the latter is too clumsy). We want to estimate the difference between the exact result D_e and the approximate result D_a. Using the estimates for the error in the individual arithmetic operations, we obtain an estimate for the composite result.

$$D_a = \left(V_{j+1}^n - 2V_{j-1}^n \left(1 + \varphi_1 \ 10^{-t} \right) \right) \left(1 + \varphi_2 \ 10^{-t} \right)$$

We have $|\varphi_i| \le 5$, and we will assume $t \ge 3$ so that $|\varphi_i \ 10^{-t}| \le .01$. Then

$$D_a = V_{j+1}^n - 2V_{j-1}^n - 2V_{j-1}^n \varphi_1 \ 10^{-t} + \left(V_{j+1}^n - 2V_{j-1}^n - 2V_{j-1}^n \varphi_1 10^{-t} \right) \varphi_2 10^{-t}$$

If we let $\|V^n\|_\infty = \max\limits_{1 \le j \le J-1} |V_j^n|$, then

$$D_a = V_{j+1}^n - 2V_{j-1}^n + \eta \|V^n\|_\infty$$

where $|\eta| \le 25.1 \times 10^{-t}$.

Problem 2.6-1. Find an estimate for the roundoff error in the computation of V^{n+1} from V^n. Show that

$$f\ell\left(V_j^n + \mu\left(V_{j+1}^n - 2V_j^n + V_{j-1}^n\right)\right) = V_j^n + \mu\left(V_{j+1}^n - 2V_j^n + V_{j-1}^n\right) + \xi\|V^n\|_\infty$$

where $|\xi| \le 80 \times 10^{-t}$ (assume $0 \le \mu \le 1$).

Note that in the above problem we compute only the error caused by the arithmetic used to go from the nth to the n+1st stage. There is already some error in V^n; that is, a difference between V^n and U^n. We must now estimate the growth or accumulation in the error. We do this from a knowledge of the error committed at each time level. The method is the same as that used to prove convergence in section 2.1. There we knew the truncation error at each time level, and we wanted to compute the accumulated error. We know stability has a profound effect on this error growth. We have the following equations

$$U_j^{n+1} = U_j^n + \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right)$$

$$V_j^{n+1} = V_j^n + \mu\left(V_{j+1}^n - 2V_j^n + V_{j-1}^n\right) + \epsilon_j^n$$

$$|\epsilon_j^n| \le \xi\|V^n\|_\infty \qquad \text{where } \xi = 8 \times 10^{-t+1}$$

Now suppose we are able to compute a bound for $\|V^n\|_\infty$ if $n\Delta t \le T$, $\|V^n\|_\infty \le M$. Then $|\epsilon^n| \le \hat{\epsilon}$, where $\xi M = \hat{\epsilon}$. We will find such a bound shortly, but first we will estimate the error $E_j^n = V_j^n - U_j^n$. We will again use the maximum principle as we did in section 2.1. If we subtract the equations above we obtain

$$E_j^{n+1} = E_j^n + \mu\left(E_{j+1}^n - 2E_j^n + E_{j-1}^n\right) + \epsilon_j^n$$

or $$E_j^{n+1} = (1 - 2\mu)E_j^n + \mu E_{j+1}^n + \mu E_{j-1}^n + \epsilon_j^n$$

We now assume $\mu < \frac{1}{2}$ and take the absolute value of both sides to obtain

$$|E_j^{n+1}| \le (1 - 2\mu)|E_j^n| + \mu|E_{j+1}^n| + \mu|E_{j-1}^n| + |\epsilon_j^n|$$

If we let $e^n = \max\limits_{1\le j\le J-1} |E_j^n|$ we obtain

$$e^{n+1} \le e^n + \hat{\epsilon}$$

By induction we thus have

$$e^n \le e^0 + n\hat{\epsilon}$$

Note that $e^n = \|U^n - V^n\|_\infty$ is a measure of the error growth in terms of the initial error e^0. For most problems this number $e^0 + n\hat{\epsilon}$ is very small compared to the truncation error. On the Control Data 6600, we could take $t = 14$, thus $\hat{\epsilon} = 8 \times 10^{-13} \times M$. Therefore n could be quite large and still $n\hat{\epsilon}$ would be quite small. For this scheme one can obtain a better result; both the truncation and roundoff error are bounded by a constant (at a fixed Δx) which does not grow with n. That is, convergence is uniform in time. We need not add the restriction $n\Delta t \le T$. The reason this occurs is the fact that the norm of the difference operation L_h is bounded by $\|L_h\| \le 1 - 0(\Delta t)$. The difference operator is strongly dissipative, as is the differential equation. Hence the errors are dissipated to zero as time advances. Therefore we may run as many time steps as we wish without a disastrous accumulation of error. We will not prove this statement [see Gary, 1966]. We leave the computation of a bound for $\|V^n\|_\infty$ as a problem.

Problem 2.6-2. Assume

$$V_j^{n+1} = V_j^n + \mu\left(V_{j-1}^n - 2V_j^n + V_{j+1}^n\right) + \epsilon_j^n \qquad 1 \le j \le J-1$$

with $V_0^n = V_j^n = 0$, $\mu < \frac{1}{2}$, $|\epsilon_j^n| \le \xi\|V^n\|_\infty$, $0 \le \xi$. Show that

$$\|V^n\|_\infty \le (1+\xi)^n\|V^0\|_\infty \le e^{n\xi}\ \|V^0\|_\infty$$

Hint: Use the maximum principle to prove

$$\|V^{n+1}\|_\infty \le (1+\xi)\ \|V^n\|_\infty$$

Now proceed by induction. To show $(1+\xi)^n \le e^{n\xi}$, first show $1+\xi \le e^\xi$ (remember $0 \le \xi$).

Examples: 1) Stable run $n \to \infty$ inhomogeneous heat equation

2) $\mu = .55$, $f(x) \ge \sin \pi x$

To provide an example of the effect of roundoff error, we solved the heat equation on the Control Data 6600. We used the difference scheme given by equations (2.1-3) with the initial function $f(x) = \sin \pi x$. If we neglect the effect of roundoff error, we can then solve the difference scheme to obtain $U_j^n = M^n \sin \pi x_j$ where $M = 1-4\mu\sin^2(\pi\Delta x/2)$. We used $J = 40$, therefore $\Delta x = 0.025$ and $M = 1 - \sigma\pi^2\Delta t + O(\Delta x^3)$. We used $\mu = 0.55$ so that the difference scheme is not stable. However $0 < M < 1$, and therefore $\lim_{n\to\infty} U_j^n = 0$ if we do not consider roundoff error. The scheme will converge for $f(x) = \sin \pi x$ even if $\mu > \frac{1}{2}$. However, we have neglected the effect of roundoff error. This in effect introduces a high frequency perturbation into the equation (why is it high frequency?). This perturbation will grow since $\mu > \frac{1}{2}$. A perturbation in the initial conditions of the form $f(x_j) = \epsilon \sin \pi(J-1)x_j$ will grow to an amplitude given by $\epsilon\left[1 - 4\mu \sin^2(\pi(J-1)/2J)\right]^n = M^n_{J-1}$. In figure 2.6-1 we show the result of this computation. We plotted the solution for various values of $T = n\Delta t$. The effect of roundoff error is clearly evident for $T = 0.060$. To reach $T = 0.06$ requires about 175 time steps and the corresponding value of M^n_{J-1} is 4×10^{15}. The amplitude of the perturbation is about 0.05. An initial perturbation of the form $\epsilon \sin \pi(J-1)x_j$ with $\epsilon \cong 1. \times 10^{-17}$ would grow to this amplitude after 175 steps. The Control Data 6600 uses a 48-bit mantissa so we might expect a perturbation of $2^{-48} \cong 4 \times 10^{-15}$. Thus our growth rate is somewhat less than the predicted maximum. This we might expect due to statistical fluctuation. The numbers seem to be reasonable.

Problem 2.6-3. If you make an attempt to solve the heat equation using the scheme of equations (2.1-3) with $J = 100$, $f(x) = \sin 2\pi x$, $\mu = 0.6$, how many time steps would you expect to run before the roundoff error exceeded 10 percent?

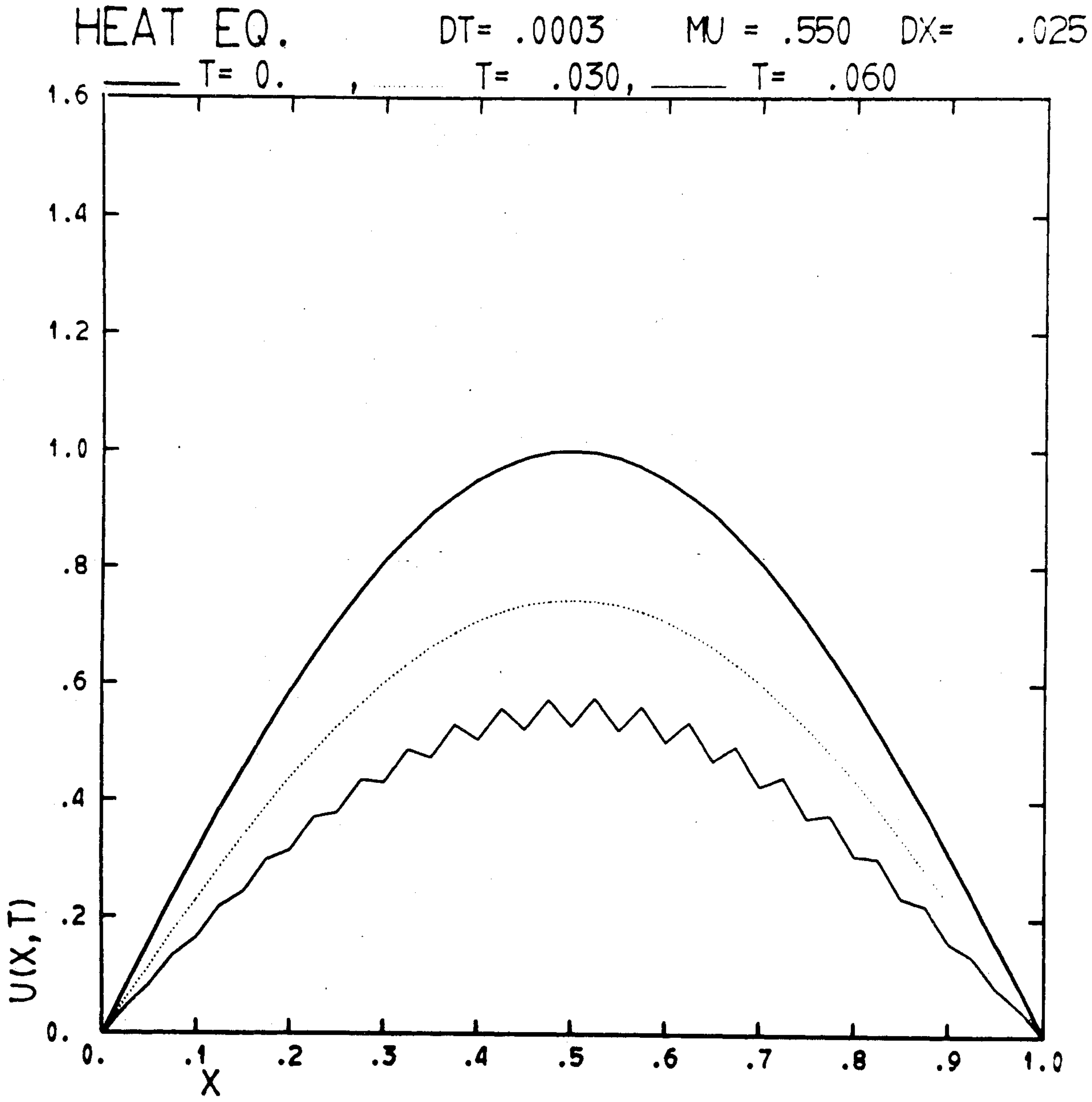

HEAT EQ.
DT= .0003
MU = .550
DX= .025
T= 0.
T= .030
T= .060
1.6
1.4
1.2
1.0
.8
.6
.4
.2
0.
U(X,T)
0.
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
X

Figure 2.6-1

3. THE CONSTRUCTION OF FINITE DIFFERENCE SCHEMES

In this chapter we will consider several of the standard methods for the construction of finite difference schemes. Most of these schemes can be explained by their application to the heat equation

$$\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2}$$

or the simple hyperbolic equation

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0$$

We will consider some schemes which apply to problems defined over multi-dimensional spaces such as the heat equation

$$\frac{\partial u}{\partial t} = \sigma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right), \qquad u = u(x,y,t)$$

We will defer to chapter 4 the treatment of systems of equations such as the wave equation

$$\frac{\partial u}{\partial t} = c \frac{\partial w}{\partial x}$$

$$\frac{\partial w}{\partial t} = c \frac{\partial u}{\partial x}$$

However, the schemes discussed in this section can, in most cases, be applied to such systems. We will also defer to later chapters complications due to boundary conditions or nonlinear terms in the differential equations.

3.1 The Leapfrog Scheme

We will illustrate this scheme as it applies to the equation

$$\frac{\partial u}{\partial t} + c\,\frac{\partial u}{\partial x} = 0.$$

We will assume periodic boundary conditions $u(x+2,t) = u(x,t)$, $-1 \le x \le 1$, with initial condition $u(x,0) = f(x)$. We will use a centered difference in space to approximate $\partial u/\partial x$ (we use the usual notation $u_t = \partial u/\partial t$ and $u_x = \partial u/\partial x$). Then we have

$$(u_x)^n_j = \frac{u^n_{j+1} - u^n_{j-1}}{2\Delta x} + \Delta x^2\,\frac{\partial^3 u}{\partial x^2}\,(\eta, t_n), \qquad x_{j-1} \le \eta \le x_{j+1}$$

This can be shown by use of a Taylor series with remainder (see problem 2.1-1). We thus have a second order truncation error; that is, $\tau = 0(\Delta x^2)$. In order to obtain a finite difference "marching scheme" we might approximate the time derivative as follows

$$(u_t)^n_j = \frac{u^{n+1}_j - u^n_j}{\Delta t} + \frac{\Delta t}{2}\,\frac{\partial^2 u}{\partial t^2}\,(x_j, \xi) \qquad t_n \le \xi \le t_{n+1}$$

If we substitute these expressions for u_t and u_x into the differential equation we obtain

$$U^{n+1}_j = U^n_j - \frac{\lambda}{2}\left(U^n_{j+1} - U^n_{j-1}\right) \qquad \lambda = c\Delta t/\Delta x\ . \qquad (3.1\text{-}1)$$

If we start with $U^0_j = f(x_j)$ we can then compute U^n_j for all values of n and j by "marching" forward. However, in section 2, we showed that this scheme is not stable; it will not work. We will have the catastrophic growth of error at high frequency which is typical of unstable schemes. Even if this scheme were stable, there would still be a disadvantage from the standpoint of accuracy. An explicit scheme for a hyperbolic equation will usually diverge unless the mesh ratio $|\lambda|$ is less than one (see the discussion of the Courant-Friedrichs-Lewy condition in section 4). This means that $\Delta t = 0(\Delta x)$ and therefore the truncation error due to the u_t term is $0(\Delta x)$ and due to the u_x term is $0(\Delta x^2)$. Of course, this imbalance is due to the use of a centered difference for the u_x term and a forward difference for the u_t term. Suppose we use a centered difference for the u_t term

$$(u_t)_j^n = \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + \frac{\Delta t^2}{6} \frac{\partial^3 u}{\partial t^3} (x_j, \xi)$$

Our finite difference scheme would then become

$$U_j^{n+1} = U_j^{n-1} - \lambda(U_{j+1}^n - U_{j-1}^n) \tag{3.1-2}$$

This is still a marching method, except we need to know the values of U_j on both the n-1st and nth time levels in order to compute U_j^{n+1}. If we knew the vectors $U^0 = \{U_j^0 | -J \leq j < J\}$ and U^1, then we could compute U^3, U^4, ... U^n ... in that order. We obtain the n+1st level if we leapfrog across the nth level from the n-1st level.

Note that we are only given $U_j^0 = f(x_j)$ from the initial conditions. In order to start this scheme, we must somehow compute the U^1. We could simply set $U_j^1 = U_j^0$. Since $U_j^1 = U_j^0 + \Delta t U_{t,j}^0 + 0(\Delta t^2)$ this would introduce an error which is first order; that is, $0(\Delta t)$. This is sometimes satisfactory because our system may be primarily dependent on an external driving force rather than the initial conditions. This means that any error in the initial data will have a small influence on the result.

Problem 3.1-1. Consider the heat equation with a driving function

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \sin t \sin \pi x$$

$$u(x,0) = f(x) = \Sigma a_k \sin \pi k x$$

Find an expression for the time T such that the effect of the initial conditions is less than ϵ for $t \geq T$. The solution will depend on the a_k.

We could use our unstable scheme to compute U^1.

$$U_j^1 = U_j^0 - \frac{\lambda}{2}(U_{j+1}^0 - U_{j-1}^0)$$

This will cause some growth in the high frequencies, but we only use this for one step so this growth will be very limited. This gives us U^1 with an accuracy $O(\Delta t^2)$ which is consistent with the accuracy of our leapfrog scheme. We know the error in the integration of a differential equation by a stable difference scheme is proportional to the truncation error τ. That is, if

$$u^{n+1} = L_h u^n + \Delta t \tau^n \qquad \|\tau^n\| \leq \tau$$

$$U^{n+1} = L_h U^n, \; U_j^0 = u_j^0 = f(x_j)$$

then $\displaystyle\max_{n\Delta t \leq T} \|U^n - u^n\| \leq M\tau$.

Thus if $\tau = O(\Delta x^2)$, then the error $\|U^n - u^n\|$ is $O(\Delta x^2)$. However, the error in a single time step is $O(\Delta t\tau)$, rather than $O(\tau)$. The following problem should illustrate this.

Problem 3.1-2. Suppose we have a stable finite difference scheme with truncation error $\tau = O(\Delta x^2)$ (that is, there is a constant M which may depend on the solution u but not on Δx such that $\|\tau^n\| \le M\Delta x^2$ if $n\Delta t \le T$). Suppose we make an error in the initial conditions so that $\|U^0 - u^0\| = O(\Delta x^2)$. Show that the error is $O(\Delta x^2)$, $\|U^n - u^n\| = O(\Delta x^2)$.

Note that if U^0 were exact, $U^0 = u^0$, then the truncation error in U^1 would be $\|U^1 - u^1\| = O(\Delta t\Delta x^2)$ rather than $O(\Delta x^2)$. That is, we can tolerate an error $O(\Delta x^2)$ in a single step, but if this error occurs in all steps, then it must be $O(\Delta t\Delta x^2)$ in order that the final error be $O(\Delta x^2)$. This is exactly what one would expect since the number of time steps is bounded by $T/\Delta t$. Usually the error is approximately proportional to the number of time steps, but not always (see problem 3.7-3).

We will now consider another method to compute U^1 in order to start our leapfrog scheme. If we have a complicated differential equation to solve, we may not care to write a separate program to compute U^1. We would like to use the same leapfrog scheme which computes U^{n+1} from U^{n-1} and U^n. This could be done as follows. Choose ρ such that $\rho = \Delta t 2^{-s}$ for some positive integer s. Define the vector $\hat{U}^0$ by $\hat{U}_j^0 = U_j^0 = f(x_j)$, $-J \le j < J$. Use the leapfrog scheme to compute $\hat{U}^\nu$ for $1 \le \nu \le s$. Start with the time increment equal to ρ and double this time increment at each step.

$$\hat{U}_j^\nu = U_j^0 - \frac{2^\nu \rho}{2\Delta x}\left(\hat{U}_{j+1}^{(\nu-1)} - \hat{U}_{j-1}^{(\nu-1)}\right) \qquad 1 \le \nu \le s$$

Then the vector $\hat{U}^\nu$ is an approximation to $u(x,\hat{t}_\nu)$ where $\hat{t}_\nu = 2^\nu \rho$. Thus $\hat{U}^s$ is an approximation to U^1. Once we have U^1, then we can use the leapfrog scheme to compute the vectors U^n, $n > 1$. Note that the error we make at the first step is $O(\rho)$ and therefore if we choose s so that $2^{-s} \cong \Delta t$ then this error is $O(\Delta t^2)$. The truncation error τ made in computing $\hat{U}^{(\nu)}$ for $\nu \geq 1$ is also $O(\Delta t^2)$, so our final error in $U^1 = \hat{U}^s$ is $O(\Delta t^2)$.

Now we are ready to consider the stability of the leapfrog scheme given by equations (3.1-2). The truncation error we leave as an easy exercise.

Problem 3.1-3. Compute the truncation error for the leapfrog scheme given by equation (3.1-2). Assume the mesh ratio is bounded by $|\lambda| < 1$. Show that the truncation error is $O(\Delta x^2)$, $|\tau| \leq M\Delta x^2$, and find an estimate for M.

We will study the stability by use of the finite Fourier analysis (note that we have assumed periodic boundary conditions). We assume U^0 and U^1 are given and U^n is computed from equation (3.1-2) for $n \geq 2$. The finite Fourier representation of U^n is

$$U_j^n = \sum_{k=-J}^{J-1} a_k^{(n)} e^{i\pi k x_j} \qquad -J \leq j < J .$$

(see section 2.4). We must compute the coefficients $a_k^{(n)}$. If we substitute the above expression for U^n into the finite difference scheme and equate terms with the same exponential factor (that is, the same value of k) we obtain

$$\sum_k a_k^{(n+1)} e^{i\pi kx} = \sum_k \left\{ a_k^{(n-1)} - \lambda\left(e^{i\pi k\Delta x} - e^{-i\pi k\Delta x}\right) a_k^{(n)} \right\} e^{i\pi kx}$$

$$a_k^{(n+1)} = a_k^{(n-1)} - 2i\lambda \sin\pi k\Delta x \; a_k^{(n)}$$

We can solve this two-term recurrence relation in the form

$$a_k^{(n)} = A_k[\beta_+(k)]^n + B_k[\beta_-(k)]^n$$

where $\beta_\pm$ are the roots of the quadratic

$$Z^2 + 2i\gamma_k Z - 1 \qquad\qquad \gamma_k = \lambda \sin k\pi\Delta x$$

Thus $\beta_\pm = -i\gamma_k \pm \sqrt{1 - \gamma_k^2}$ (see section 1.6). The values of $a_k^{(0)}$ and $a_k^{(1)}$ are known since U^0 and U^1 are given. Then A_k and B_k are determined by solving the 2×2 system of equations

$$A_k + B_k = a_k^{(0)}$$

$$A_k\beta_+(k) + B_k\beta_-(k) = a_k^{(1)}$$

If $|\gamma_k| < 1$, then $\beta_+ \neq \beta_-$ and the solution is possible. If $|\beta_\pm| \leq 1$, then

$$|a_k^{(n)}| \leq |A_k| + |B_k| \; .$$

We give the remainder of the proof of stability as an exercise.

Problem 3.1-4. Show that there is a constant M independent of n and Δx such that $|a_k^{(n)}| \le M|a_k^{(0)}|$. Assume $a_k^{(1)} = a_k^{(0)} + \Delta t b_k + O(\Delta t^2)$, $-J \le k < J$, where b_k is some complex vector. Why is this a reasonable assumption?

This completes the proof of stability under our assumption of periodic boundary conditions since

$$\|U^n\|_2^2 = \frac{1}{2J} \sum_{k=-J}^{J-1} |a_k^{(n)}|^2 \le M^2\|U^0\|_2^2$$

Note that our proof of stability breaks down if $|\lambda| > 1$ since we no longer have $|\gamma_k| < 1$. It is not difficult to show that the scheme is unstable if $|\lambda| > 1$.

We could try the leapfrog scheme on the heat equation. The scheme for the heat equation based on a forward time difference is

$$U_j^{n+1} = U_j^n + \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \qquad \mu = \sigma\Delta t/\Delta x^2$$

This scheme is stable and the truncation error is $O(\Delta t) + O(\Delta x^2)$. Since $\Delta t = O(\Delta x^2)$ we would gain little if we made the truncation error $O(\Delta t^2) + O(\Delta x^2)$ since in this case $O(\Delta t) + O(\Delta x^2) = O(\Delta t^2) + O(\Delta x^2) = O(\Delta x^2)$. In fact the leapfrog scheme is unstable for the heat equation.

Problem 3.1-5. Show that the following scheme is not stable; take $\mu = \sigma\Delta t/\Delta x^2$ to be constant.

$$U_j^{n+1} = U_j^{n-1} + 2\mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \qquad 1 \le j \le J-1$$

$$U_j^0 = f(x_j)$$

$$U_j^1 = U_j^0 + \mu\left(U_{j+1}^0 - 2U_j^0 + U_{j-1}^0\right)$$

$$U_0^n = U_J^n = 0$$

Sometimes it is useful to have a difference scheme for the heat equation which is centered in time. The reason will be apparent in section 3.7 when we discuss dissipative difference schemes. Such a centered scheme is the DuFort Frankl scheme.

$$U_j^{n+1} = U_j^{n-1} + 2\mu\left(U_{j+1}^n - U_j^{n+1} - U_j^{n-1} + U_{j-1}^n\right)$$

Problem 3.1-6. Show that the truncation error for the DuFort Frankl scheme is $0(\Delta t^2) + 0(\Delta x^2) + 0\left[\left(\frac{\Delta t}{\Delta x}\right)^2\right]$. Show that the scheme is stable for all values of μ.

3.2 Construction of a Difference Scheme by Use of a Taylor Series Expansion

We will illustrate the method by derivation of the Lax-Wendroff scheme for the simple hyperbolic equation $u_t + cu_x = 0$. If we differentiate this scheme we obtain

$$\frac{\partial^2 u}{\partial t^2} = -c\,\frac{\partial^2 u}{\partial t \partial x} = c^2\,\frac{\partial^2 u}{\partial x^2}$$

Now consider the Taylor series

$$u_j^{n+1} = u(x_j, t_n + \Delta t) = u(x_j, t_n) + \Delta t u_t(x_j, t_n) + \frac{\Delta t^2}{2}\, u_{tt}(x_j, t_n) + 0(\Delta t^3)$$

From the differential equation we have $u_t = -cu_x$, $u_{tt} = c^2 u_{xx}$. Therefore we can replace the time derivatives on the right by space derivatives

$$u_j^{n+1} = u_j^n - \Delta t c u_x(x_j, t_n) + \frac{c^2 \Delta t^2}{2}\, u_{xx}(x_j, t_n) + 0(\Delta t^3)$$

We know $u_x = (u_{j+1} - u_{j-1})/2\Delta x + 0(\Delta x^2)$ (we assume the solution $u(x,t)$ is sufficiently smooth to permit this error estimate). We have a similar expression for u_{xx}. Thus we obtain

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{\lambda^2}{2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + 0(\Delta t \Delta x^2)$$
$$+ 0(\Delta t^2 \Delta x^2) + 0(\Delta t^3) \; .$$

Using periodic boundary conditions we then obtain the following difference scheme.

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{\lambda^2}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \qquad -J \le j < J$$

$$U_j^0 = f(x_j) \qquad \lambda = c\Delta t/\Delta x \qquad (3.2\text{-}1)$$

$$U_{j\pm 2J}^n = U_j^n$$

As always, we have the fundamental question of stability before us. Since we have a linear difference scheme with constant coefficients and periodic boundary conditions, we can answer this stability question by consideration of the Fourier modes. We let U^n be represented by

$$U_j^n = [M(k)]^n e^{i\pi k x_j}$$

For notational convenience we denote $[M(k)]^n$ by M^n, here we mean the n^{th} power of M and not the value of a function at the n^{th} level, that is not $M^{(n)}$. Substituting into equation (3.2-1) and dividing out the term $M^n e^{i\pi kx}$ we obtain

$$M = 1 - i\lambda\sin\theta + \lambda^2(\cos\theta - 1) \qquad \text{where } \theta = \pi k\Delta x \qquad (3.2\text{-}2)$$

If we can show that this amplification factor M is less than 1 for all θ, then we know our difference scheme is stable. We have

$$|M|^2 = 1 + 2\lambda^2(\cos\theta-1) + \lambda^2(\cos\theta-1)^2 + \lambda^2\sin^2\theta$$

$$= 1 - \lambda^2(1-\cos\theta)^2(1-\lambda^2)$$

If $|\lambda| < 1$ then $|M|^2 \le 1$, since $\lambda^2(1-\lambda^2) \le \frac{1}{4}$ and $(1-\cos\theta)^2 \le 4$. Therefore the Lax-Wendroff scheme is stable provided $|\lambda| < 1$.

Suppose we had based our difference scheme on only the first two terms of the Taylor series expansion for u(x,t).

$$u_j^{n+1} = u_j^n + \Delta t u_t(x_j, t_n) + 0(\Delta t^2)$$

Using the relation $u_t = -cu_x$ we would obtain

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right)$$

We know this scheme is not stable. Therefore the Taylor series method may not produce a usable difference scheme.

Suppose we attempt to use a Taylor series to produce a scheme for the heat equation such that truncation error is fourth order, $\tau = 0(\Delta x^4)$. By differentiating the equation $u_t = \sigma u_{xx}$, we obtain $u_{tt} = \sigma^2 U_{x^4}$. If we substitute into the Taylor series we obtain

$$U_j^{n+1} = U_j^n + \Delta t \sigma u_{xx}(x_j, t_n) + \frac{\Delta t^2}{2}\sigma^2 u_{x^4}(x_j, t_n) + 0(\Delta t^3) \ .$$

Suppose we assume $\mu = \sigma\Delta t/\Delta x^2$ is to be held constant, then $\Delta t = 0(\Delta x^2)$. In order that the truncation error τ be fourth order $\tau = 0(\Delta x^4)$, we must approximate u_{xx} with error $0(\Delta x^4)$ and u_{x^4} with error Δx^2.

Problem 3.2-1. If u(x) is a sufficiently differentiable function, show the following difference approximations are valid.

$$u_{xx}(x_j) = \frac{-u_{j-2} + 16u_{j-1} - 30u_j + 16u_{j+1} - u_{j+2}}{12\Delta x^2} + \frac{\Delta x^4}{180} u_{x^5}(x_j) + 0(\Delta x^5)$$

$$u_{x^4}(x_j) = \frac{u_{j-2} - 4u_{j-1} + 6u_j - 4u_{j+1} + u_{j+2}}{\Delta x^4} - \frac{\Delta x^2}{144} u_{x^6}(x_j) + 0(\Delta x^3)$$

If we use the above difference approximations in the Taylor series we obtain

$$U_j^{n+1} = U_j^n + \frac{\mu}{12}\left(-U_{j-2}^n + 16U_{j-1}^n - 30U_j^n + 16U_{j+1}^n - U_{j+2}^n\right)$$

$$+ \frac{\mu^2}{2}\left(U_{j-2}^n - 4U_{j-1}^n + 6U_j^n - 4U_{j+1}^n + U_{j+2}^n\right) + 0(\Delta t \Delta x^4)$$

Again, the fundamental question is stability. Also, we should ask if it is wise to use a high order difference formula. Our error estimate ($\sigma = 0(\Delta x^4)$) is not valid unless u possesses derivatives up through the sixth order. Our solution might not be this smooth. (Our simple heat equation has an analytic solution; but if σ is no longer a constant, for example σ might be a discontinuous function of x, then $\partial u/\partial x$ might not be continuous.) In this case a high order difference scheme might do more harm than good.

Problem 3.2-2. Determine if the above fourth order difference scheme for the heat equation is stable. Once you have an expression for the amplification factor M(k), you might wish to use a computer to see if $|M(k)| \leq 1$ for all relevant k (with Δx and μ fixed).

Next we will consider the simple, nonlinear hyperbolic equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

$$u(x,0) = f(x)$$

We assume periodic boundary conditions. Since this equation is nonlinear the theory of stability and convergence which we have developed does not apply. However, it still provides a useful example for the construction of a finite difference scheme. If we are to use the Taylor series we must compute u_{tt}. This is

$$u_{tt} = -u_t u_x - u u_{xt} = u(u_x)^2 + u(u u_x)_x$$

We could also use $u_t = -\frac{1}{2}(u^2)_x$ which leads to $u_{tt} = -(u u_t)_x = (u^2 u_x)_x$.
If we substitute the first expression into the Taylor series and replace the spatial derivatives by finite differences we obtain ($\lambda = \Delta t/\Delta x$)

$$\begin{aligned} U_j^{n+1} = U_j^n &- \frac{\lambda}{2} u_j^n \left(u_{j+1}^n - u_{j-1}^n\right) + \frac{\lambda^2}{2} u_j^n \left(u_{j+1}^n - u_{j-1}^n\right)^2 \\ &+ \frac{\lambda^2}{2} u_j^n \left[\left(\frac{u_{j+1}^n + u_j^n}{2}\right)\left(u_{j+1}^n - u_j^n\right) - \left(\frac{u_j^n + u_{j-1}^n}{2}\right)\left(u_j^n - u_{j-1}^n\right)\right] \end{aligned} \qquad (3.2\text{-}3)$$

Note the method used to difference $(u u_x)_x$. We have approximated this term by

$$\frac{u(x_{j+\frac{1}{2}})\, u_x(x_{j+\frac{1}{2}}) - u(x_{j-\frac{1}{2}})\, u_x(x_{j-\frac{1}{2}})}{\Delta x}$$

If we had used

$$\frac{u(x_{j+1})\, u_x(x_{j+1}) - u(x_{j-1})\, u_x(x_{j-1})}{2\Delta x}$$

we would have obtained a difference relation involving values at the five points x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}, instead of three points, namely

$$\frac{u_{j+1}\left(\frac{u_{j+2} - u_j}{2\Delta x}\right) - u_{j-1}\left(\frac{u_j - u_{j-2}}{2\Delta x}\right)}{2\Delta x}$$

As a general rule, one uses as few points as possible in a finite difference scheme. With a larger number of points, the scheme is more likely to be unstable, especially if the boundary conditions are not of the periodic type. We will have more to say concerning nonlinear equations in a later chapter.

3.3 Predictor-Corrector (or two-step) Schemes

These schemes make an initial guess for the values of u on the $n+1^{st}$ level, and then correct this initial guess. For complex nonlinear systems of differential equations, these two-step schemes are easier to program than the schemes based on a Taylor expansion. We will illustrate the idea by use of the hyperbolic equation $u_t + cu_x = 0$ [Burstein, 1965].

Given the vector U^n we first predict an approximation for $u(x_j + \Delta x/2, t_{n+1})$. Denote this prediction by $\hat{U}^{n+1}_{j+\frac{1}{2}}$. Using this prediction we then correct it to obtain U^{n+1}_j. The definition of the scheme is given below (assume periodic boundary conditions):

$$\hat{U}^{n+1}_{j+\frac{1}{2}} = \tfrac{1}{2}\left(U^n_{j+1} + U^n_j\right) - \lambda\left(U^n_{j+1} - U^n_j\right) \qquad -J \le j < J \qquad \lambda = c\Delta t/\Delta x$$

$$U^{n+1}_j = U^n_j - \frac{\lambda}{4}\left(U^n_{j+1} - U^n_{j-1}\right) - \frac{\lambda}{2}\left(\hat{U}^{n+1}_{j+\frac{1}{2}} - \hat{U}^{n+1}_{j-\frac{1}{2}}\right) \qquad -J \le j < J \tag{3.3-1}$$

These schemes represent an attempt to center the finite differences at time $t_{n+\frac{1}{2}} = t_n + \Delta t/2$. We could write the second equation as

$$\frac{U^{n+1}_j - U^n_j}{\Delta t} = -\frac{c\Delta t}{2}\left[\frac{U^n_{j+1} - U^n_{j-1}}{2\Delta x} + \frac{\hat{U}^{n+1}_{j+\frac{1}{2}} - \hat{U}^{n+1}_{j-\frac{1}{2}}}{\Delta x}\right]$$

The time difference is certainly centered at time $t_{n+\frac{1}{2}}$. The spatial difference on the right is the average of spatial difference terms at t_n and t_{n+1}; therefore the right side is also centered at $t_{n+\frac{1}{2}}$. The error in the $\hat{U}^{n+1}$ terms is $O(\Delta t^2)$ so that the error in $\left(\hat{U}^{n+1}_{j+\frac{1}{2}} - \hat{U}^{n+1}_{j-\frac{1}{2}}\right)/\Delta x$ is also $O(\Delta t^2)$. Multiplication by $c\Delta t/2$ will produce an error $O(\Delta t^3)$ so that the truncation error is second order, $\tau = O(\Delta t^2)$.

We will next prove the stability of this method by the usual Fourier analysis. We will compute the amplification factor for the Fourier mode $e^{i\pi kx}$. Let $U_j^n = e^{i\pi kx_j}$. Then

$$\hat{U}^{n+1}_{j+\frac{1}{2}} = \left(\cos\theta/2 - 2i\lambda \sin\theta/2\right) e^{i\pi kx_{j+\frac{1}{2}}}$$

where $\lambda = c\Delta t/\Delta x$, $\theta = \pi k\Delta x$, $x_{j+\frac{1}{2}} = x_j + \Delta x/2$. Next we substitute this expression into the equation for U^{n+1} to obtain

$$U_j^{n+1} = \left[1 - \frac{\lambda}{2}\, i\sin\theta - i\lambda\sin\theta/2\left(\cos\theta/2 - 2i\lambda\sin\theta/2\right)\right] e^{i\pi kx_j}$$

$$= \left[1 - 2\lambda^2\sin^2\theta/2 - 2i\lambda\sin\theta/2\,\cos\theta/2\right] e^{i\pi kx_j}$$

$$= \left[1 + \lambda^2(\cos\theta - 1) - i\lambda\sin\theta\right] U_j^n = M(k)U_j^n$$

Therefore the amplification factor for this scheme is the same as that for the Lax-Wendroff scheme given in equation (3.2-2). Therefore this predictor-corrector scheme is stable. In fact, for this simple linear hyperbolic equation, the predictor-corrector scheme is the same as the Lax-Wendroff scheme.

Problem 3.3-1. Show that the difference scheme given by equations (3.3-1) is the same as the Lax-Wendroff scheme given by equations (3.2-1).

We can do the predictor-corrector scheme in various ways. Consider the nonlinear equation $\partial u/\partial t + u\ \partial u/\partial x = 0$ which we will use in the form $\partial u/\partial t + \frac{1}{2}\,\partial u^2/\partial x = 0$. The nonlinear equivalent of the scheme given by equations (3.3-1) is

$$\hat{U}^{n+1}_{j+\frac{1}{2}} = \tfrac{1}{2}\left(U^n_{j+1} + U^n_j\right) - \frac{\lambda}{2}\left[\left(U^n_{j+1}\right)^2 - \left(U^n_j\right)^2\right] \qquad \lambda = \Delta t/\Delta x \qquad (3.3\text{-}2)$$

$$U^{n+1}_j = U^n_j - \frac{\lambda}{4}\left[\frac{\left(U^n_{j+1}\right)^2 - \left(U^n_{j-1}\right)^2}{2} + \left(\hat{U}^{n+1}_{j+\frac{1}{2}}\right)^2 - \left(\hat{U}^{n+1}_{j-\frac{1}{2}}\right)^2\right]$$

If we lift the predictor to the $t_{n+\frac{1}{2}}$ level instead of the t_{n+1} level we obtain

$$\hat{U}^{n+\frac{1}{2}}_{j+\frac{1}{2}} = \tfrac{1}{2}\left(U^n_{j+1} + U^n_j\right) - \frac{\lambda}{4}\left[\left(U^n_{j+1}\right)^2 - \left(U^n_j\right)^2\right] \qquad \lambda = \Delta t/\Delta x \qquad (3.3\text{-}3)$$

$$U^{n+1}_j = U^n_j - \frac{\lambda}{2}\left[\left(U^{n+\frac{1}{2}}_{j+\frac{1}{2}}\right)^2 - \left(U^{n+\frac{1}{2}}_{j-\frac{1}{2}}\right)^2\right]$$

We are dealing with a very simple differential equation here. For a complex system of differential equations, the latter two-step scheme can be much simpler than the Lax-Wendroff scheme of equations (3.2-3) which is based on a Taylor series. For a complex system of PDE the computation of the second derivatives, such as u_{tt}, can involve many terms.

Another two-step version for this nonlinear equation is the following:

$$\hat{U}^{n+1}_j = \tfrac{1}{2}\left(U^n_{j+1} + U^n_{j-1}\right) - \frac{\lambda}{4}\left[\left(U^n_{j+1}\right)^2 - \left(U^n_{j-1}\right)^2\right]$$

$$U^{n+1}_j = U^n_j - \frac{\lambda}{8}\left[\left(U^n_{j+1}\right)^2 - \left(U^n_{j-1}\right)^2 + \left(\hat{U}^{n+1}_{j+1}\right)^2 - \left(\hat{U}^{n+1}_{j-1}\right)^2\right]$$

We will discuss some properties of these variations in a later chapter.

Problem 3.3-2. Devise a predictor-corrector scheme for the heat equation and determine its truncation error and stability.

Problem 3.3-3. Is the following predictor-corrector scheme for the equation $u_t + cu_x = 0$ stable?

$$\hat{U}_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad \lambda = c\Delta t/\Delta x$$

$$U_j^{n+1} = U_j^n - \frac{\lambda}{4}\left[U_{j+1}^n - U_{j-1}^n + \hat{U}_{j+1}^{n+1} - \hat{U}_{j-1}^{n+1}\right]$$

Problem 3.3-4. Suppose we try several predictor-corrector iterations. We let S be the number of iterations. If S = 2 we have the scheme described in problem 3.3-3. The scheme is the following

$$\hat{U}_j^0 = U_j^n$$

$$\hat{U}_j^\nu = U_j^n - \frac{\lambda}{4}\left[U_{j+1}^n - U_{j-1}^n + \hat{U}_{j+1}^{\nu-1} - \hat{U}_{j-1}^{\nu-1}\right] \qquad 1 \le \nu \le S$$

$$U_j^{n+1} = \hat{U}_j^S$$

Show that this scheme is stable if S = 3 provided $|\lambda| < 2$. In general, the scheme is unstable if S is even and stable for odd S provided $|\lambda| < 2$ [Gary, 1964].

3.4 Implicit Difference Schemes for One-Dimensional Problems

These schemes are obtained if we center both space and time differences midway between the nth and n+1st level. We do this without using a predictor; therefore, we obtain for the heat equation $u_t = \sigma u_{xx}$

$$U_j^{n+1} = U_j^n + \frac{\mu}{2}\left[U_{j+1}^n - 2U_j^n + U_{j-1}^n + U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}\right] \qquad (3.4\text{-}1)$$

$$U_j^0 = f(x_j) \qquad 1 \le j \le J-1 \qquad \Delta x = 1/J$$

$$U_0^n = U_J^n = 0$$

This difference scheme for the heat equations is known as the Crank-Nicholson scheme. These equations involve U_j^{n+1} on the right side in the space difference terms. They cannot be solved explicitly for the U_j^{n+1} terms; that is, we cannot obtain a simple algebraic expression for the U_j^{n+1} term. We must invert a tridiagonal matrix to find the U_j^{n+1}; the U_j^{n+1} are defined implicitly by the above equation. To see this we write equation (3.4-1) in matrix form where C is a matrix of order J-1.

$$\left(I - \frac{\mu}{2} C\right)U^{n+1} = \left(I + \frac{\mu}{2} C\right)U^n$$

$$C = \begin{vmatrix} -2 & 1 & 0 & \cdots & & & 0 \\ 1 & -2 & 1 & 0 & \cdots & & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & \cdots & & & & 1 & -2 \end{vmatrix} \qquad U^n = \begin{vmatrix} U_1^n \\ U_2^n \\ \cdot \\ \cdot \\ \cdot \\ U_{J-1}^n \end{vmatrix}$$

Note that we can write this scheme in the form

$$U^{n+1} = L_h U^n$$

where L_h is the matrix operator $\left[I - \frac{\mu}{2} C\right]^{-1} \left[I + \frac{\mu}{2} C\right]$. We must solve a system of equations whose matrix is the tridiagonal matrix $I - \frac{\mu}{2} C$ (I is the identity matrix). If we write out the matrix $I - \frac{\mu}{2} C$ we have

$$I - \frac{\mu}{2} C = B = \begin{vmatrix} 1+\mu & -\mu/2 & 0 \ldots\ldots\ldots\ldots & & & 0 \\ -\mu/2 & 1+\mu & -\mu/2 & 0 \ldots & & \\ 0 & -\mu/2 & 1+\mu & -\mu/2 & 0 \ldots & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 \ldots\ldots\ldots\ldots & & & 0 & -\mu/2 & 1+\mu \end{vmatrix}$$

This matrix is diagonally dominant. The diagonal element in each row is greater than the sum of the absolute value of the diagonal element in that row; that is

$$|b_{ii}| > \sum_{j \neq i} b_{ij} \qquad \text{for } 1 \leq i \leq n .$$

Therefore the matrix is nonsingular, and we can solve the system of equations (3.4-1). In fact, the matrix $I - \frac{\mu}{2} C$ is symmetric and positive definite. For further discussion of these matters from matrix theory see section 1.2 and section 5.2.2. We can write down the eigenvalues and eigenvectors for the matrix C and therefore for $I - \frac{\mu}{2} C$.

Problem 3.4-1. Show that the eigenvectors of C are $U_j^{(r)} = \sin\pi r x_j$ $1 \le j \le J-1$, $1 \le r \le J-1$. The eigenvalues of C are $-4\sin^2(\pi r\Delta x/2)$, $\Delta x = 1/J$. Therefore the eigenvalues of $I - \frac{\mu}{2} C$ are $1 + 2\mu\sin^2(\pi r\Delta x/2)$. This problem is very similar to problem 2.3-1.

We have to solve the system $\left(I - \frac{\mu}{2} C\right)U^{n+1} = \left(I + \frac{\mu}{2} C\right)U^n$. Since the matrix $I - \frac{\mu}{2} C$ is symmetric and positive definite we can solve this system by Gaussian elimination without interchange of rows. Normal Gaussian elimination requires the interchange of rows in order to maximize the "pivot elements." We need to solve the system Bu = f where B is a tridiagonal matrix, u is the unknown vector, and f the known vector. This is done with a forward sweep followed by a backward sweep. We start with the equations in the form

$$\begin{vmatrix} \beta_1 & \gamma_1 & & & & \\ \alpha_2 & \beta_2 & \gamma_2 & & & \\ 0 & \alpha_3 & \beta_3 & \gamma_3 & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & \cdots & \cdots & \cdots & \alpha_{J-1} & \beta_{J-1} \end{vmatrix} \begin{vmatrix} U_1 \\ \\ \\ \\ \\ \\ U_{J-1} \end{vmatrix} = \begin{vmatrix} f_1 \\ \\ \\ \\ \\ \\ f_{J-1} \end{vmatrix}$$

The forward sweep transforms this to a triangular system of equations by adding a multiple of the j^{th} equation to the $j+1^{th}$ equation in order to eliminate the α_{j+1} term. The algorithm is

$$\hat{\beta}_1 = \beta_1 \qquad \hat{\gamma}_1 = \gamma_1 \qquad \hat{f}_1 = f_1$$

$$\hat{\beta}_{j+1} = -\frac{\alpha_{j+1}}{\hat{\beta}_j}\hat{\gamma}_j + \beta_{j+1} \qquad \hat{\gamma}_{j+1} = \gamma_{j+1}$$

$$\hat{f}_{j+1} = -\frac{\alpha_{j+1}}{\hat{\beta}_j}\hat{f}_j + f_{j+1} \qquad 1 \le j \le J-2$$

The system then takes the form

$$\begin{vmatrix} \hat{\beta}_1 & \hat{\gamma}_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \hat{\beta}_2 & \hat{\gamma}_2 & 0 & \cdots & & \\ 0 & 0 & \hat{\beta}_3 & \hat{\gamma}_3 & 0 & \cdots & \\ \cdot & & & & & & \\ \cdot & & & & & \hat{\beta}_{J-2} & \hat{\gamma}_{J-2} \\ \cdot & & & & & & \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \hat{\beta}_{J-1} \end{vmatrix} \begin{vmatrix} U_1 \\ \\ \\ \\ \\ \\ U_{J-1} \end{vmatrix} = \begin{vmatrix} \hat{f}_1 \\ \\ \\ \\ \\ \\ \hat{f}_{J-1} \end{vmatrix}$$

This system is then solved by a simple backward substitution

$$U_{J-1} = \hat{f}_{J-1}/\hat{\beta}_{J-1}$$

$$U_j = \frac{1}{\hat{\beta}_{J-1}}\left(\hat{f}_j - \hat{\gamma}_{j+1} U_{j+1}\right) \qquad J-2 \ge j \ge 1$$

With a computer where division is much slower than multiplication we would probably compute $1/\hat{\beta}_j$, then $\alpha_{j+1}*(1/\hat{\beta}_j)$, $\hat{\beta}_{j+1}$, $\hat{f}_j$. Thus we would perform 1 division, 5 multiplications, and 3 additions for each component U_j (except U_1 and U_{J-1}). Thus we do a total of about 9(J-1) floating point operations. Thus we pay a slight additional price for use of the implicit scheme--we have to solve a system of linear equations. This price is frequently small compared with the total computation required to solve the problem. The advantage of the implicit scheme is the lack of a stability restriction on the size of Δt. For the explicit scheme of section 2.1 we must have $\mu = \sigma\Delta t/\Delta x^2 < \frac{1}{2}$, or $\Delta t \le \Delta x^2/(2\sigma)$. The implicit scheme is stable for all values of Δt. We will discuss this later.

First we will describe another way to derive the solution of our tridiagonal system of equations. This is taken from the book by Richtmyer and Morton and results in essentially the same algorithm as Gaussian elimination. However, it involves a somewhat different point of view which is sometimes useful. We have to solve the system of equations

$$\alpha_j U_{j-1} + \beta_j U_j + \gamma_j U_{j+1} = f_j \qquad 1 \le j \le J-1 \qquad (3.4\text{-}2)$$

$$U_0 = U_J = 0$$

Suppose we consider those sequences U_j which satisfy the equation and the left boundary condition $U_0 = 0$, but not the right boundary condition $U_J = 0$. This is a one parameter family of solutions since we may specify U_1 arbitrarily but then the remaining values of U_j, $2 \le j \le J$, are determined by the equation for f_j. Suppose we assume that U_J can be specified as the parameter rather than U_1; that is, we assume U_J can be specified arbitrarily, then the remaining U_j determined to satisfy the equation for f_j as well as the left boundary condition $U_0 = 0$. It seems reasonable to look for such a solution in the form of a linear relation $U_j = E_j U_{j+1} + F_j$, $0 \le j < J$. If we substitute into the equation for f_j we obtain

$$\alpha_j\left(E_{j-1}U_j + F_{j-1}\right) + \beta_j U_j + \gamma_j U_{j+1} = f_j \qquad 1 \le j \le J-1$$

Therefore $U_j + \dfrac{\gamma_j}{\alpha_j E_{j-1} + \beta_j} U_{j+1} = \dfrac{f_j - \alpha_j F_{j-1}}{\alpha_j E_{j-1} + \beta_j}$ and we have the following condition

$$E_j = \frac{-\gamma_j}{\alpha_j E_{j-1} + \beta_j} \qquad F_j = \frac{f_j - \alpha_j F_{j-1}}{\alpha_j E_{j-1} + \beta_j}$$

The equation for U_0 is $U_0 = E_0 U_1 + F_0$. Since U_1 is arbitrary we must have $E_0 = F_0 = 0$. Thus we have the forward sweep

$$E_0 = 0 \qquad F_0 = 0$$

$$E_j = -\gamma_j/(\alpha_j E_{j-1} + \beta_j) \qquad F_j = \frac{f_j - \alpha_j F_{j-1}}{\alpha_j E_{j-1} + \beta_j} \qquad 1 \le j \le J-1$$

followed by the backward sweep

$$U_J = 0 \qquad U_j = E_j U_{j+1} + F_j \qquad J-1 \ge j \ge 1 .$$

A comparison with the Gaussian elimination algorithm shows

$$E_j = \frac{-\gamma_j}{\hat{\beta}_j} \quad \text{or} \quad \hat{\beta}_j = \alpha_j E_{j-1} + \beta_j = \frac{-\alpha_j \gamma_{j-1}}{\hat{\beta}_{j-1}} + \beta_j$$

Problem 3.4-2. Assume that we have diagonal dominance with positive diagonals, $\beta_j > |\alpha_j| + |\gamma_j|$. Prove (use induction) that $|E_j| \le 1$ for all j and $\hat{\beta}_j > 0$ for all j.

Either process for solving the system will fail only if $\hat{\beta}_j$ vanishes. The above shows that this cannot occur. Also, the numbers involved in the process E_j and F_j do not become unreasonably large as long as the solution U_j is reasonably bounded. If $|U_j| \le M$, then from $U_j = E_j U_{j+1} + F_j$, $|E_j| \le 1$, we have $|F_j| \le 2M$. Since none of the numbers involved become large, we will probably not lose accuracy due to cancellation in computing the solution to the recurrence relations. This reasoning is usually valid, although sometimes not as the problem below will show. However, we can use the error analysis of Wilkinson to show that the above solution will not produce a disastrous accumulation of roundoff error.

Problem 3.4-3. Consider the recurrence relation $x_{n+1} = \alpha x_n + \beta$. Suppose $\beta = A - \alpha A$, $x_0 = A$; then the solution is $x_n = A$. What would you expect for the behavior of roundoff error? Try it on a computer. You might try $\alpha = 1/\sqrt{2}$, $A = \pi$, or $\alpha = 2$, $A = 2$, or $\alpha = \pi$, $A = \sqrt{2}$, or $\alpha = \pi$, $A = 1$. Can you explain the results?

Problem 3.4-4. We could solve the system of equations (3.4-2) as follows. First set $U_1 = 1$ and solve the equations, that is

$$U_0^{(1)} = 0, \quad U_1^{(1)} = 1, \quad U_j^{(1)} = \frac{1}{\gamma_{j-1}}\left(f_{j-1} - \beta_{j-1} U_{j-1}^{(1)} - \alpha_{j-1} U_{j-2}^{(1)}\right) \qquad 2 \le j \le J$$

Now solve the corresponding homogeneous equations

$$U_0^{(2)} = 0, \quad U_1^{(2)} = 1, \quad U_j^{(2)} = \frac{1}{\gamma_{j-1}}\left(-\beta_{j-1} U_{j-1}^{(2)} - \alpha_{j-1} U_{j-2}^{(2)}\right) \qquad 2 \le j \le J$$

Then it is easy to see that we can obtain the solution of equations (3.4-2) by a linear combination of $U^{(1)}$ and $U^{(2)}$, namely

$$U_j = U_j^{(1)} - \frac{U_J^{(1)}}{U_J^{(2)}} U_j^{(2)}$$

Show that this method will produce a disastrous accumulation of roundoff error for the heat equation problem, $\alpha_j = \gamma_j = -\frac{\mu}{2}$, $\beta_j = 1 + \mu$. You might try it on a computer, then explain the result. Or you may do it analytically. Hint: The general solution of the recurrence relation

$$\alpha U_{j+1} + \beta U_j + \gamma U_{j-1} = 0$$

is given by $U_j = Az_+^j + Bz_-^j$ where $z_\pm$ are the roots of $\alpha z^2 + \beta z + \gamma = 0$ (assume $z_+ \neq z_-$). The values of A and B are determined by the starting values U_0 and U_1. Show that $U_j^{(2)}$ will grow very rapidly. What is the significance of this growth to roundoff error?

Problem 3.4-5. The implicit scheme requires the solution of the equation $Bv^n = f^n$ for each time step. Since $B = I - \frac{\mu}{2} C$ does not depend on n, we could compute B^{-1} once, store it, and then simply compute $v^n = B^{-1}f^n$. Why is this a bad idea? Consider both storage and computing time. However, we could speed up the process outlined above by storing the appropriate three vectors. What should be stored?

We have engaged in a long discussion of methods used to solve the system of equations produced by the implicit difference scheme for the heat equation. Now we should consider the reason we use this scheme. This scheme is unconditionally stable. We can base our choice of Δt solely on accuracy considerations; there is no stability restriction.

This unconditional stability will frequently result in a much larger time-step and a corresponding reduction in computing time. The following problem will illustrate this.

Problem 3.4-6. Consider the heat equation $u_t = \sigma u_{xx} + f$. The term $f = f(x,t)$ represents a source or sink of heat. Let $u(x,t) = 4x(1-x)\sin \omega t$. Then u is a solution provided $f = 4x(1-x)\omega \cos \omega t + 8\sigma \sin \omega t$. We take the boundary conditions to be $u(0,t) = u(1,t) = 0$. Code this problem for a computer using both the explicit scheme of section 2.1 and the implicit scheme given above. Solve this problem for $0 \le t \le T$ using appropriate values of σ, ω, T, Δt, Δx. Compare accuracy and computer time for the two schemes.

To perform the stability analysis we assume periodic boundary conditions and consider the modes $U_j^n = [M(k)]^n e^{i\pi k x_j}$. If we substitute into equation (3.4-1) and divide out the factor $M^n e^{i\pi k x_j}$ we obtain

$$M = 1 + \frac{\mu}{2}\left(2\cos\theta - 2 + (2\cos\theta - 2)M\right), \quad \theta = \pi k \Delta x \ .$$

Solving for $M = M(k)$ we have

$$M = \frac{1 - \mu(1-\cos\theta)}{1 + \mu(1-\cos\theta)} = \frac{1 - A}{1 + A} \qquad \text{where } A = \mu(1-\cos\theta)$$

Since $A \ge 0$ it is easy to see that $|M(k)| \le 1$. Therefore the amplification factor is bounded by 1 independent of k, Δx, and μ. The implicit scheme is unconditionally stable. The proof for the boundary condition $u(0,t) = u(1,t) = 0$ can be done the same way.

Next we will give a proof of stability for the implicit scheme based on the energy method. The idea of the energy method is to define a norm for the vectors U^n and then prove something like the following inequality $\|U^{n+1}\| \leq (1+k\Delta t)\|U^n\|$. The constant k must be independent of n, Δx, Δt, and the solution U^n. Then $\|U^n\| \leq (1+k\Delta t)^n\|U^0\| \leq e^{kt_n}\|U^0\|$ and this implies stability. The norm, in some sense, measures the length of the vector. For example, the Euclidean or L_2 norm is most frequently used, $\|U\|_2 = \sqrt{\sum_{j=0}^{J} |u_j|^2}$. We might use the L_1 norm $\|U\|_1 = \sum_{j=0}^{J} |U_j|$ or the maximum norm $\|U\|_\infty = \max_{0 \leq j \leq J} |U_j|$. If A is any symmetric positive definite matrix, then the following relation defines a norm $\|U\|_A = \sqrt{U^T Au} = \sqrt{\sum_i \sum_j u_j a_{ij} u_j}$. As we noted in chapter 1, the following properties characterize a norm: 1) $\|U\| \geq 0$ and $\|U\| = 0$ iff $U = 0$, 2) $\|\alpha U\| = |\alpha|\,\|U\|$ for any scaler α, and 3) $\|U + W\| \leq \|U\| + \|W\|$.

We will use the ordinary L_2 norm to prove that the scheme given by equations (3.4-1) is stable. First we will need the following identity.

Problem 3.4-7. Given a vector ψ_j, $0 \leq j \leq J$ where $\psi_0 = \psi_J = 0$, prove that

$$\sum_{j=1}^{J-1} \psi_j\left(\psi_{j+1} - 2\psi_j + \psi_{j-1}\right) = -\sum_{j=0}^{J-1} \left(\psi_{j+1} - \psi_j\right)^2$$

Now take equations (3.4-1) and multiply by $\left(U_j^{n+1} + U_j^n\right)$. We obtain

$$\left(U_j^{n+1} + U_j^n\right)\left(U_j^{n+1} - U_j^n\right) = \frac{\mu}{2}\left(U_j^{n+1} + U_j^n\right)\left[U_{j+1}^{n+1} + U_{j+1}^n - 2\left(U_j^{n+1} + U_j^n\right) + U_{j-1}^{n+1} + U_{j-1}^n\right]$$

If we let $\psi_j = U_j^{n+1} + U_j^n$, we can sum these equations and use the above problem to obtain

$$\sum_{j=1}^{J-1} \left(U_j^{n+1}\right)^2 - \sum_{j=1}^{J-1} \left(U_j^n\right)^2 = \frac{\mu}{2} \sum_{j=1}^{J-1} \psi_j \left(\psi_{j+1} - 2\psi_j + \psi_{j-1}\right) =$$

$$- \frac{\mu}{2} \sum_{j=1}^{J-1} \left(\psi_{j+1} - \psi_j\right)^2 \leq 0$$

Therefore

$$\sum_{j=1}^{J-1} \left(U_j^{n+1}\right)^2 \leq \sum_{j=1}^{J-1} \left(U_j^n\right)^2 \qquad \text{or} \qquad \|U^{n+1}\|_2 \leq \|U^n\|_2$$

The energy decreases (or at least does not increase) at each time step, and therefore we certainly have a stable scheme.

Note the similarity between this proof and the derivation for the following energy inequality for the heat equation which we discussed in chapter 1. Multiply both sides of this equation by u and integrate with respect to both x and t to obtain

$$\int_0^1 \int_0^t uu_t dtdx = \tfrac{1}{2} \int_0^1 \int_0^t (u^2)_t dtdx = \tfrac{1}{2} \int_0^1 u^2(x,t)dx - \int_0^1 u^2(x,0)dx$$

$$= \sigma \int_0^1 \int_0^t uu_{xx} dtdx = \sigma \int_0^t [u(1,t)u_x(1,t) - u(0,t)u_x(0,t)]dx - \sigma \iint (u_x)^2 dxdt$$

Since $u(1,t) = u(0,t) = 0$ we have

$$\tfrac{1}{2}\int_0^1 u^2(x,t)dx - \tfrac{1}{2}\int_0^1 u^2(x,0)dx = -\sigma\int_0^t\int_0^1 (u_x)^2 dxdt \le 0 \qquad (3.4\text{-}3)$$

This implies that the function $E(t) = \frac{1}{2}\int_0^1 u^2(x,t)dx$ is a non-increasing function of t.

We proved that the explicit scheme for the heat equation described in section 2.1 was stable by use of a so-called maximum principle. We can regard this as a proof by the energy method since we showed that

$$\|U^{n+1}\|_\infty \le \|U^n\|_\infty$$

where $\|U^n\|_\infty = \max\limits_{0\le j\le J} |U_j^n|$.

Problem 3.4-8. Consider the implicit scheme for the simple hyperbolic equation $u_t + cu_x = 0$ with periodic boundary conditions.

$$U_j^{n+1} = U_j^n - \frac{\lambda}{4}\left[U_{j+1}^{n+1} - U_{j-1}^{n+1} + U_{j+1}^n - U_{j-1}^n\right] \qquad \lambda = c\Delta t/\Delta x$$

Show that this scheme is unconditionally stable by use of the Fourier analysis method. Prove the same thing by means of the energy method. Is the matrix equation for U^{n+1} tridiagonal? How would you solve the matrix equation? Suppose the values of U^{n+1} at the boundary are given; that is, $U_{-J}^{n+1} = g(t_{n+1})$, $U_J^{n+1} = g(t_{n+1})$. In this case the matrix is tridiagonal. Show that this tridiagonal matrix is nonsingular.

3.5 Implicit Schemes in More Than One Dimension: Alternating-Direction-Implicit (ADI) Methods

We have seen that there is a great advantage in the use of implicit schemes, especially for the heat equation because of the restrictive stability condition $\Delta t \le \Delta x^2/(2\sigma)$. What happens if we try an implicit scheme for a problem in two dimensions? For example, consider the heat equation on a square

$$\frac{\partial u}{\partial t} = \sigma\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \qquad 0 \le x \le 1, \qquad 0 \le y \le 1$$

$$u(x,y,0) = f(x,y)$$

We require u to vanish on the boundary; that is, $u(x,y,t) = 0$ if $x = 0$, or $x = 1$, or $y = 0$, or $y = 1$. We let $u^n_{jk} = u(x_j, y_k, t_n)$, $x_j = j\Delta x$, $y_k = k\Delta y$, $\Delta x = 1/J$, $\Delta y = 1/K$, $0 \le j \le J$, $0 \le k \le K$. The following would then be a two-dimensional implicit scheme.

$$\begin{aligned} U^{n+1}_{jk} = U^n_{jk} &+ \mu_x\left(U^{n+1}_{j+1,k} + U^n_{j+1,k} - 2\left(U^{n+1}_{jk} + U^n_{jk}\right) + U^{n+1}_{j-1,k} + U^n_{j-1,k}\right) \\ &+ \mu_y\left(U^{n+1}_{j,k+1} + U^n_{j,k+1} - 2\left(U^{n+1}_{jk} + U^n_{jk}\right) + U^{n+1}_{j,k-1} + U^n_{j,k-1}\right) \end{aligned}$$

where $\mu_x = \sigma\Delta t/(2\Delta x^2)$, $\mu_y = \sigma\Delta t/(2\Delta y^2)$. In matrix form this scheme is $(I+C)U^{n+1} = (I-C)U^n$. The matrix C is given by (here we have $\gamma = 1 + 2\mu_x + 2\mu_y$)

$$C = \left[\begin{array}{c|c|c}
\begin{matrix} \gamma & -\mu_x & 0 & \cdots & \\ -\mu_x & \gamma & -\mu_x & 0 & \cdots \\ \vdots & & & & \\ & & -\mu_x & \gamma & -\mu_x \\ 0 & \cdots & 0 & -\mu_x & \gamma \end{matrix} &
\begin{matrix} -\mu_y & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\mu_y \end{matrix} &
\begin{matrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{matrix} \\ \hline
\begin{matrix} -\mu_y & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\mu_y \end{matrix} &
\begin{matrix} \gamma & -\mu_x & 0 & \cdots & \\ -\mu_x & \gamma & -\mu_x & 0 & \cdots \\ \vdots & & & & \\ 0 & \cdots & 0 & -\mu_x & \gamma \end{matrix} &
\begin{matrix} -\mu_y & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\mu_y \end{matrix} \\ \hline
& \vdots &
\end{array}\right.$$

We have described the matrix operator for a two-dimensional explicit problem in section 2.5. The order of the matrix C (and the dimension of the vectors U^n) is (J-1) × (K-1). The matrix C can be partitioned into submatrices each of order J-1; that is,

$$\begin{array}{c} \begin{matrix} 1 & 2 & 3 & \cdots & & K-1 \end{matrix} \\ C = \left|\begin{matrix} D_x & D_y & 0 & \cdots & & \\ D_y & D_x & D_y & \cdots & & \\ 0 & D_y & D_x & D_y & \cdots & \\ \vdots & & & & & \\ 0 & \cdots & & & D_y & D_x \end{matrix}\right| \end{array}$$

where D_x is a tridiagonal matrix and D_y a diagonal matrix.

$$D_x = \begin{vmatrix} \gamma & -\mu_x & 0 & \cdots & & & \\ -\mu_x & \gamma & -\mu_x & 0 & \cdots & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & \cdots & & & 0 & -\mu_x & \gamma \end{vmatrix} \qquad D_y = \begin{vmatrix} -\mu_y & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & -\mu_y \end{vmatrix} \qquad \gamma = 1+2\mu_x+2\mu_y$$

In order to determine U^{n+1} we must solve the system $(I+C)U^{n+1} = (I-C)U^n$. The matrix $B = I+C$ is not tridiagonal--it is block tridiagonal. Thus B is a banded matrix with band width 2J-1, that is $b_{ij} = 0$ if $|i-j| > J-1$. To solve such a banded matrix by Gaussian elimination would require far too much computing. This solution has to be done at each time step, and there may be hundreds or thousands of time steps.

Problem 3.5-1. Show that the matrix I+C above is nonsingular. Estimate the number of floating point operations required to solve $(I+C)U^{n+1} = (I-C)U^n$ if the matrix is treated as a banded matrix.

In 1955 Peaceman and Rachford and also Douglas devised a very effective scheme for the heat equation. This is the alternating-direction-implicit method. It is unconditionally stable, has second-order accuracy, and requires nothing more than the solution of a tridiagonal matrix system. We will apply the ADI method to the heat equation problem described above. We use the notation $\delta_x^2U^n = U^n_{j+1,k} - 2U^n_{j,k} + U^n_{j-1,k}$, $\delta_y^2 = U^n_{j,k+1} - 2U^n_{j,k} + U^n_{j,k-1}$. The straight forward implicit scheme described above is then

$$U^{n+1} = U^n + \mu_x\delta_x^2\left(U^{n+1} + U^n\right) + \mu_y\delta_y^2\left(U^{n+1} + U^n\right).$$

The ADI scheme is a two-step scheme, defined as follows:

$$\hat{U}^1 = U^n + \mu_x\delta_x^2\left(\hat{U}^1 + U^n\right) + \mu_y\delta_y^2\left(2U^n\right) \tag{3.5-1}$$

$$U^{n+1} = U^n + \mu_x\delta_x^2\left(\hat{U}^1 + U^n\right) + \mu_y\delta_y^2\left(U^{n+1} + U^n\right)$$

The reader should convince himself that both operators involved here are tridiagonal matrix operators. This is true if we order the $\hat{U}^1$ vector with the x index first, and U^{n+1} with the y index first; that is

$$\hat{U}^1 = \left(\hat{U}^1_{11}, \hat{U}^1_{21}, \hat{U}^1_{31}, \ldots, \hat{U}^1_{J-1,1}, \hat{U}^1_{1,2} \ldots, \hat{U}^1_{J-1,K-1}\right)^T$$

$$U^{n+1} = \left(U^{n+1}_{11}, U^{n+1}_{12}, \ldots, U^{n+1}_{1,K-1}, U^{n+1}_{2,1}, U^{n+1}_{2,2} \ldots, U^{n+1}_{J-1,K-1}\right)^T$$

Since the matrices are tridiagonal, the solution of these equations requires only a modest effort.

If our problem were three-dimensional, $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$, then we would have a third operator δ_z^2 similar to δ_x^2 and δ_y^2. The ADI scheme would then be a three-step scheme

$$\hat{U}^1 = U^n + \mu_x\delta_x^2\left(\hat{U}^1 + U^n\right) + \mu_y\delta_y^2\left(2U^n\right) + \mu_z\delta_z^2\left(2U^n\right) \tag{3.5-2}$$

$$\hat{U}^2 = U^n + \mu_x\delta_x^2\left(\hat{U}^1 + U^n\right) + \mu_y\delta_y^2\left(\hat{U}^2 + U^n\right) + \mu_z\delta_z^2\left(2U^n\right)$$

$$U^{n+1} = U^n + \mu_x\delta_x^2\left(\hat{U}^1 + U^n\right) + \mu_y\delta_y^2\left(\hat{U}^2 + U^n\right) + \mu_z\delta_z^2\left(U^{n+1} + U^n\right)$$

It might seem more reasonable to use the term $\mu_x \delta_x^2 \left(\hat{U}^2 + U^n\right)$ rather than $\mu_x \delta_x^2 \left(\hat{U}^1 + U^n\right)$ since we would then be using the latest and presumably best approximation to U^{n+1}. However, this would not produce an unconditionally stable scheme.

Problem 3.5-2. Show that the ADI scheme given by equations (3.4-1) has truncation error $O(\Delta t^2) + O(\Delta x^2) + O(\Delta y^2)$.

We will analyze the stability of this scheme by use of Fourier analysis. Since our problem is two-dimensional, we will need a two-dimensional Fourier analysis. If f(x,y) is a suitably smooth function defined on the square $-1 \le x \le 1$, $-1 \le y \le 1$, then we have the Fourier series representation

$$f(x,y) = \sum_{s=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} a_{rs} e^{i\pi(rx+sy)}$$

$$a_{rs} = \tfrac{1}{4} \int_{-1}^{1} \int_{-1}^{1} f(x,y)\, e^{-i\pi(rx+sy)} dxdy$$

This representation has a discrete analogue just as in the one-dimensional case. Given a vector U_{jk}, $-J \le j < J$, $-K \le k < K$, then we have

$$U_{jk} = \sum_{s=-K}^{K-1} \sum_{r=-J}^{J-1} a_{rs} e^{i\pi(rx_j + sy_k)}$$

$$a_{rs} = \frac{1}{4JK} \sum_{k=-K}^{K-1} \sum_{j=-J}^{J-1} U_{jk} e^{-i\pi(rx_j + sy_k)} \qquad x_j = j/J, \quad y_k = k/K$$

If we have a linear finite difference scheme with constant coefficients and periodic boundary conditions in two dimensions, we can determine its stability by the same method used for the one-dimensional problems. We simply substitute the Fourier modes into the difference schemes and compute the amplification factor. If we substitute the Fourier mode $U^n_{jk} = e^{i\pi(rx_j+sy_k)}$ into the ADI scheme defined by equations (3.5-1) we obtain the following:

$$\hat{U}^1_{jk} = \hat{M}_1\, e^{i\pi(rx_j+sy_k)} \qquad U^{n+1}_{jk} = M\, e^{i\pi(rx_j+sy_k)}$$

where

$$\hat{M}_1 = 1 + \mu_x\, 2(\cos\theta-1)(\hat{M}_1+1) + 4\mu_y(\cos\psi-1)$$

$$M = 1 + \mu_x\, 2(\cos\theta-1)(\hat{M}_1+1) + \mu_y\, 2(\cos\psi-1)(M+1)$$

$$\theta = \pi r\Delta x\ , \qquad \psi = \pi s\Delta y$$

If we let $\mu_x\, 2(\cos\theta-1) = g_x$, $\mu_y\, 2(\cos\psi-1) = g_y$, then

$$g_x(\hat{M}_1+1) = \hat{M}_1 - 1 - 2g_y$$

$$(1-g_x)\hat{M}_1 = 1 + g_x + 2g_y$$

$$(1-g_y)M = 1 + \hat{M}_1 - 1 - 2g_y + g_y$$

$$(1-g_x)(1-g_y)M = (1-g_x)\hat{M}_1 - (1-g_x)g_y$$

$$(1-g_x)(1-g_y)M = 1 + g_x + 2g_y - g_y + g_x g_y$$

$$M = \frac{(1+g_x)(1+g_y)}{(1-g_x)(1-g_y)}$$

Since $g_x \le 0$, $g_y \le 0$, we have $|M| \le 1$ for all values of r, s, n, Δx, Δy, Δt. Therefore we have an unconditionally stable scheme.

Problem 3.5-3. Consider the following scheme:

$$\hat{U}^1 = U^n + \mu_x \delta_x^2 \left(\hat{U}^1 + U^n\right) + 2\mu_y \delta_y^2 U^n + 2\mu_z \delta_z^2 U^n$$

$$\hat{U}^2 = U^n + \mu_x \delta_x^2 \left(\hat{U}^1 + U^n\right) + \mu_y \delta_y^2 \left(\hat{U}^2 + U^n\right) + \mu_z \delta_z^2 \left(\hat{U}^1 + U^n\right)$$

$$U^{n+1} = U^n + \mu_x \delta_x^2 \left(\hat{U}^2 + U^n\right) + \mu_y \delta_y^2 \left(\hat{U}^2 + U^n\right) + \mu_z \delta_z^2 \left(U^{n+1} + U^n\right)$$

Show that it is not unconditionally stable. Thus we should not use the "best available estimate" for U^{n+1} at each step.

Problem 3.5-4. Show that the three-dimensional ADI scheme described by equations (3.5-2) is unconditionally stable.

Problem 3.5-5. Suppose you have an ADI problem for a three-dimensional heat equation where the fields U^n and U^{n+1} are too large to fit in the fast memory of the computer. Suppose you can store these fields on a drum. Assume you have a 50×50×50 mesh, 32,000 fast memory locations available for data storage, a drum rotation time of 34 milliseconds, and a transfer rate of 100,000 words per second. Assume your computer averages 2 μsec per floating point operation, including logical overhead such as indexing, tests, etc. Can you devise an ADI algorithm including the storage allocation and buffered transfer from the drum so that your computation will not be I/O bound?

Problem 3.5-6. Devise an ADI scheme for the inhomogeneous heat equation on a rectangle

$$\frac{\partial u}{\partial t} = \sigma\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + f(x,t)$$

which has second order truncation error, $\tau = 0(\Delta t^2) + 0(\Delta x^2) + 0(\Delta y^2)$.

Problem 3.5-7. Suppose you must solve the inhomogeneous heat equation as in problem 3.5-6. Suppose there is room for only two two-dimensional fields in our computer, $U^n_{j,k}$, $U^{n+1}_{j,k}$ or $U^n_{j,k}$, $f^{n+\frac{1}{2}}_{j,k}$, for example. Is it possible to devise an ADI algorithm storing only two fields at once and computing the $f^{n+\frac{1}{2}}_{j,k}$ array only once per time step?

3.6 The Method of Fractional Time Steps.

This is a method somewhat related to the ADI method which has been developed by Soviet mathematicians. The idea is to represent the difference operator for a multidimensional problem as the product of one-dimensional operators. If the norm of each of the one-dimensional operators is bounded by unity, then the norm of the product is bounded by unity, and we have a stable scheme. Also, the amplification factor for the multidimensional scheme is simply the product of the amplification factors of the one-dimensional scheme. If the one-dimensional operators are stable, then we might expect the multidimensional scheme to be stable.

We will illustrate the method by applying it to the two-dimensional heat equation. We first advance the solution from the time t_n to $t_n + \Delta t/2 = t_{n+\frac{1}{2}}$ by use only of the terms involving x-derivatives. Then we advance from $t_{n+\frac{1}{2}}$ to t_n using only the y-derivative terms.

$$\hat{U}^{n+\frac{1}{2}} = U^n + \frac{\mu_x}{2}\delta_x^2\left(\hat{U}^{n+\frac{1}{2}} + U^n\right) \qquad \mu_x = \sigma\Delta t/(2\Delta x^2) \qquad (3.6\text{-}1)$$

$$U^{n+1} = \hat{U}^{n+\frac{1}{2}} + \frac{\mu_y}{2}\delta_y^2\left(U^{n+1} + \hat{U}^{n+\frac{1}{2}}\right) \qquad \mu_y = \sigma\Delta t/(2\Delta y^2)$$

If we define the operators B_x and B_y by $B_x = \frac{\mu_x}{2}\delta_x^2$, $B_y = \frac{\mu_y}{2}\delta_y^2$, then

$$\hat{U}^{n+\frac{1}{2}} = (I-B_x)^{-1}(I+B_x)U^n, \qquad U^{n+1} = (I-B_y)^{-1}(I+B_y)\hat{U}^{n+\frac{1}{2}} .$$

If the one-dimensional operators satisfy the conditions

$$\|(I-B_x)^{-1}(I+B_x)\| \le 1 , \qquad \|(I-B_y)^{-1}(I+B_y)\| \le 1$$

then $\|U^{n+1}\| \leq \|(I-B_y)^{-1}(I+B_y)(I-B_x)^{-1}(I+B_x)\| \; \|U^n\| \leq \|U^n\|$. If we can prove the above bounds for the one-dimensional operators, then we have stability for the two-dimensional operators. In section 3.4 we used the energy method to show that the above inequalities do hold for the one-dimensional operators.

Problem 3.6-1. Determine the truncation error for the scheme given by equations (3.6-1).

Problem 3.6-2. Consider the nonlinear hyperbolic equation defined on a square with periodic boundary conditions.

$$\frac{\partial u}{\partial t} = u \frac{\partial u}{\partial x} + u \frac{\partial u}{\partial y}$$

Apply the method of fractional time steps to this problem. You might use the Lax-Wendroff (Taylor series) technique for the one-dimensional operators. What is the truncation error for your scheme? What is the result of applying the Taylor series technique directly to the two-dimensional problem?

3.7 The Use of Dissipation to Stabilize Finite Difference Schemes.

We will introduce this section with a study of the following differential equation (see section 1.3.7).

$$\frac{\partial u}{\partial t} + c\,\frac{\partial u}{\partial x} = \sigma\,\frac{\partial^2 u}{\partial x^2}\,, \qquad -1 \le x \le 1 \qquad (3.7\text{-}1)$$

$$u(x,0) = f(x)$$

We assume periodic boundary conditions $u(x\pm2,t) = u(x,t)$, $f(x\pm2) = f(x)$. The term $\sigma\,\partial^2u/\partial x^2$ is a dissipative term--it tends to reduce the energy in the solution. We can see this if we multiply equation (3.7-1) by u and then integrate over x and t. We obtain

$$\tfrac{1}{2}\,\frac{\partial(u^2)}{\partial t} + \tfrac{1}{2}c\,\frac{\partial(u^2)}{\partial x} = \sigma\,\frac{\partial\left(u\,\frac{\partial u}{\partial x}\right)}{\partial x} - \sigma\left(\frac{\partial u}{\partial x}\right)^2$$

$$\tfrac{1}{2}\int_{-1}^{1}\left[u^2(x,t) - u^2(x,0)\right]dx + \tfrac{1}{2}c\int_{0}^{t}\left[u^2(1,t) - u^2(-1,t)\right]dt$$

$$= \sigma\int_{0}^{t}\left[u(1,t)u_x(1,t) - u(-1,t)u_x(-1,t)\right]dt - \sigma\int_{0}^{t}\int_{-1}^{1}\left(u_x\right)^2 dxdt$$

We may consider $E(t) = \tfrac{1}{2}\int_{-1}^{1} u^2(x,t)dx$ as a measure of the energy in the flow at a given time. Looked at in another way E(t) is just the L_2 norm of u at a given t. Note that if $E(t) = 0$, then $u(x,t) \equiv 0$ for $-1 \le x \le 1$. If u is a velocity, then it is quite natural to regard the integral of its square as an energy. If we use our periodic boundary condition, then the above equation becomes

$$E(t) = E(0) - \sigma \int_0^1 \int_{-1}^1 (u_x)^2 \, dxdt$$

The integral term on the right is certainly non-positive, in fact the term is negative unless u is constant. Therefore $E(t) \le E(0)$, and we have an energy inequality for our equation (3.7-1). The diffusion term $\sigma \, \partial^2 u/\partial x^2$ takes energy out of the solution. If $\sigma = 0$, then the energy is constant since $E(t) = E(0)$ for all $t \ge 0$.

We can obtain more information from the solution of equations (3.7-1) We will look for solutions in the form $u(x,t) = A_k(t)e^{i\pi kx}$. Substitution into the equation yields

$$\left(A_k' + i\pi kcA_k + \pi^2 k^2 \sigma A_k\right)e^{i\pi kx} = 0$$

If we require $A_k(0) = a_k$, then the solution must be

$$u(x,t) = a_k \, e^{-\sigma\pi^2k^2t - i\pi kct} \, e^{i\pi kx} \text{ where } u(x,0) = a_k \, e^{i\pi kx}$$

From this we can obtain the general solution. If $f(x) = \sum_{k=-\infty}^{\infty} a_k e^{i\pi kx}$, then, since our equation is linear and since we know the solution for each term in the Fourier series, we can write the general solution as

$$u(x,t) = \sum_{k=-\infty}^{\infty} a_k \, e^{-\sigma\pi^2k^2t} \, e^{i\pi k(x-ct)}$$

We may wish to impose some requirement on f(x), such as $\sum_{-\infty}^{\infty} k|a_k| < \infty$ or perhaps $\sum_{-\infty}^{\infty} k^2|a_k| < \infty$. Why might we need such as requirement?

The form of this solution tells us something about the nature of the energy dissipation. It is highly sensitive to frequency. If the term $\sigma\pi^2 k^2 t$ is quite small, then the solution approximates that of the hyperbolic equation $u_t + cu_x = 0$, namely $\sum_{-\infty}^{\infty} a_k e^{i\pi(x-ct)}$. However, each term in the series is reduced by the amount $e^{-\sigma\pi^2 k^2 t}$. Obviously, the reduction is much greater for the higher frequencies. We know that instability in a difference scheme is usually due to rapid growth of the higher frequencies. This suggests the addition of a diffusion type term to the difference equation, say the finite difference analog of $\sigma\, \partial^2 u/\partial x^2$. If σ is sufficiently small, then this term might kill the high frequency growth without affecting the desired solution too much.

As an example, consider the unstable difference scheme for the hyperbolic equation $u_t + cu_x = 0$.

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad \lambda = c\Delta t/\Delta x \ .$$

The Lax-Wendroff scheme for this same equation is

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{\lambda^2}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \tag{3.7-2}$$

(see section 3.2). It is stable provided $|\lambda| < 1$. We could regard this scheme as an obvious difference approximation to the equation $u_t + cu_x = \epsilon u_{xx}$ where $\epsilon = \lambda^2 \Delta x^2/(2\Delta t) = c^2 \Delta t/2$. We simply approximate u_t by $(U_j^{n+1} - U_j^n)/\Delta t$, u_x by $(U_{j+1}^n - U_{j-1}^n)/2\Delta x$, and u_{xx} by $(U_{j+1}^n - 2U_j^n + U_{j-1}^n)/\Delta x^2$. Note that $\epsilon = 0(\Delta t)$, therefore in the limit as the mesh spacing approaches zero, our differential equation becomes $u_t + cu_x = 0$. Therefore, we might expect the solution of the difference scheme to converge to the solution of the hyperbolic equation. Also note that we have taken an unstable scheme for $u_t + cu_x$, added a diffusion type term and thereby stabilized the scheme. In the case of the Lax-Wendroff scheme, we even improved the accuracy from $O(\Delta t)$ to $O(\Delta t^2)$. This improved accuracy comes from the Taylor series expansion which produced the difference scheme. As a fortuitous by-product we obtain the dissipative nature of the Lax-Wendroff scheme. There is a theorem due to Kriess which states that a wide class of difference schemes for hyperbolic equations can be stabilized by the addition of a diffusion type term [Kreiss, 1964].

Problem 3.7-1. Consider the difference scheme

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \epsilon\Delta t\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right)\Big/\Delta x^2$$

For what range of ϵ is this scheme stable. We have already shown it to be stable for $\epsilon = \lambda^2\Delta x^2/(2\Delta t) = O(\Delta x)$.

A second example can also be obtained starting with the leapfrog scheme

$$U_j^{n+1} = U_j^{n-1} - \lambda\left(U_{j+1}^n - U_{j-1}^n\right). \tag{3.7-3}$$

We consider the DuFort-Frankl approximation to the term $\partial^2 u/\partial x^2$, namely $\left(U_{j+1}^n - U_j^{n+1} - U_j^{n-1} + U_{j-1}^n\right)\Big/\Delta x^2$. If we form $U_j^{n+1} = U_j^{n-1} - \lambda\left(U_{j+1}^n - U_{j-1}^n\right) + \left(U_{j+1}^n - U_j^{n+1} - U_j^{n-1} + U_{j-1}^n\right)$, we see that this is equivalent to

$$U_j^{n+1} = \tfrac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad (3.7\text{-}4)$$

Note that this is equivalent to adding the term ϵu_{xx} to the hyperbolic equation, with $\epsilon = \Delta x^2/\Delta t$, and then using a DuFort-Frankl approximation for u_{xx}.

Both the leapfrog scheme of equation (3.7-3) and the scheme above (equation (3.7-4)) are stable. However, the amplification factor for the leapfrog scheme lies on the unit circle for all frequencies k. Therefore no Fourier mode is attenuated although the higher frequencies will suffer large phase shift (if there were no phase shift, the leapfrog scheme would be perfectly accurate for all frequencies--nature is usually not this generous). Since the high frequencies are not accurately represented, it may be better to dissipate them; that is, force the magnitude of the amplification factor to be less than one. Otherwise, in a nonlinear hyperbolic equation such as $u_t + uu_x = 0$, these high frequencies may interact to produce an explosive error growth (nonlinear instability, see chapter 8). The addition of the DuFort-Frankl form of the diffusion term to the leapfrog scheme does just this. In figure 3.7-1 below we have plotted the magnitude of the amplification factors for the leapfrog scheme (equation 3.7-3), the Lax-Wendroff (equation 3.7-2), and the scheme of equation (3.7-4) (sometimes called the Friedrichs scheme). For the Lax-Wendroff scheme the amplification factor is $M(k) = 1 - i\lambda\sin\theta + \lambda^2(\cos\theta - 1)$, $\theta = \pi k\Delta x$ (see section 3.2).

For the leapfrog scheme there are really two amplification factors (we will speak more of this in chapter 4) since it is a three-level scheme.

However, both factors have magnitude one. The amplification factor for the third scheme is $\cos\theta - i\lambda\sin\theta$. A dissipative scheme will have $|M| < 1$ if $0 < |\theta|$. As we will see in chapter 8 this question of dissipation frequently is quite important in the choice of a difference scheme. Most physical systems are actually dissipative even if our idealized model of the physical system is not dissipative. Therefore, it may be reasonable to add a proper amount of dissipation to our model--the problem is to determine what is proper. Of course, a closed system must conserve energy. But frequently we model only part of the total system; for example, we neglect the heating of a fluid due to viscosity, but include the viscosity damping in the momentum equations (the Navier-Stokes equations, for example). Thus our model is dissipative. If we neglect viscosity altogether, our model would be conservative.

Problem 3.7-2. Is the implicit scheme discussed in section 3.4 dissipative when applied to the hyperbolic equation $u_t + cu_x = 0$.

Problem 3.7-3. Consider the problem $u_t = \sigma u_{xx} + g(x,t)$, $u(0,t) = u(1,t) = 0$, $u(x,0) = f(x)$. Use the implicit difference scheme on this problem. Show that if we choose Δx small enough, then we can run this scheme forever at this fixed Δx ($n \to \infty$) and still the error will remain less than ϵ for all n. That is, the convergence of this scheme is uniform in time. Also, show that roundoff error will cause no trouble no matter how many time steps we take. Note that part of this problem is to formulate it precisely. We have merely supplied the meaning of the problem.

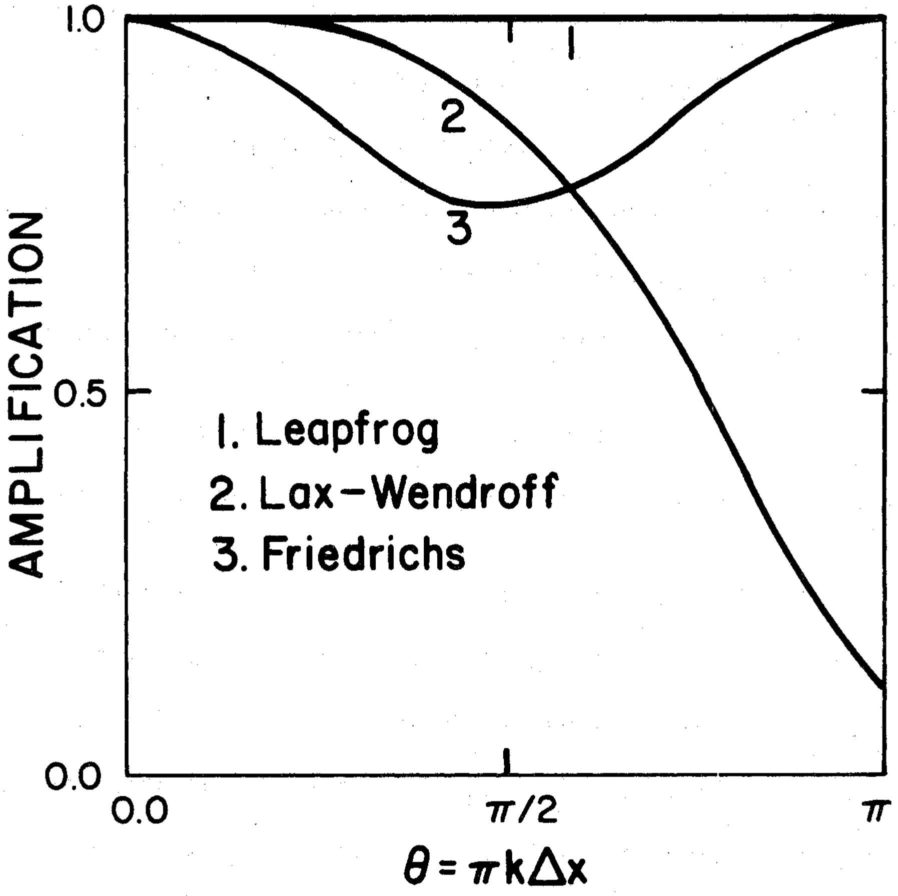


Figure 3.7-1

Amplification v.s. frequency for three difference schemes at $\lambda = 0.75$

In order to prove a stable scheme is convergent, we limit the time interval; that is, we require $n\Delta t \le T$. Convergence is not uniform in time. Also, the roundoff error may grow like e^{KT}. We cannot expect to run the difference scheme indefinitely without eventually losing all accuracy. However, if our scheme is sufficiently dissipative (for example $\|L_h\| \le 1 - K\Delta t$ where $K > 0$ is independent of the mesh spacing), then we do not get an error buildup. Devise a computer program to check the error in the solution of the above heat equation. You might use the code written for problem 3.4-6.

Problem 3.7-4. Consider the hyperbolic equation $u_t + cu_x = f$ with periodic boundary conditions. Consider the implicit scheme for this problem. Do you think you could run this scheme indefinitely with no serious buildup of error? In other words, is convergence likely to be uniform in time? Why? Write a computer program to verify your conclusion.

Problem 3.7-5. Consider the following unstable scheme for the hyperbolic equation $u_t + cu_x = 0$ (assume periodic boundary conditions).

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad \lambda = c\Delta t/\Delta x$$

If we add the dissipative approximation to ϵu_{xx}, $\lambda^2\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right)\Big/2$, we obtain the Lax-Wendroff scheme. Suppose instead we add the term $\lambda^2\left(U_{j+2}^n - 2U_j^n + U_{j-2}^n\right)\Big/8$. This is an approximation to $\frac{\Delta t^2}{2} c^2 u_{xx} = \frac{\Delta t^2}{2} u_{tt}$, so we may also regard the following scheme as being derived from a Taylor series.

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{\lambda^2}{8}\left(U_{j+2}^n - 2U_j^n + U_{j-2}^n\right).$$

Determine the truncation error and stability for this difference scheme. Is this added term dissipative, that is, is

$$\sum_{j=-J}^{J-1} U_j^n\left(U_{j+2}^n - 2U_j^n + U_{j-2}^n\right) \le 0 .$$

3.8 The Effect of Lower Order Terms on Stability

Suppose we take the heat equation $u_t = \sigma u_{xx}$ and add terms to the right side which contain only lower order derivatives, for example, $u_t = \sigma u_{xx} + au_x$. Or we might modify the hyperbolic equation $u_t + cu_x = 0$ to yield $u_t + cu_x + au = 0$. Suppose we have a stable difference scheme for the equation $u_t + cu_x = 0$ and modify it to include the au term. What is the effect on stability? In general, there is no effect. For example, if we have the scheme $U^{n+1} = L_h U^n$ for $u_t + cu_x = 0$, and we modify it so that $U^{n+1} = L_h U^n + \Delta t a U^n$, then our new operator is $(L_h + a\Delta t I)$. But $\|L_h + a\Delta t I\| \leq \|L_h\| + |a|\Delta t$, and if $\|L_h\| \leq 1 + O(\Delta t)$, then the same statement is true for the augmented operator. Thus we would expect no effect on stability. However, we have to be a little careful with this argument as the following example will show.

Suppose we consider the modified heat equation $u_t = \sigma u_{xx} + au_x$, $u(x,0) = f(x)$. We assume periodic boundary conditions. We have already obtained the solution to this equation in section 3.7. Now consider the explicit difference scheme

$$U_j^{n+1} = U_j^n + \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) + \frac{a\Delta t}{2\Delta x}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad (3.8\text{-}1)$$

Problem 3.8-1. Show that the above scheme is stable if $\mu = \sigma\Delta t/\Delta x^2 < \frac{1}{2}$. Show that the scheme is strongly stable if $\Delta t < 2\sigma/a^2$.

By strongly stable we mean that the amplification factor $A_h(k)$ for the mode $e^{ik\pi x}$ satisfies $|A_h(k)| \leq 1$ independent of the mesh spacing h and the frequency k. By stable we mean $|A_h(k)| \leq 1 + C\Delta t$

where C is independent of h and k. If $A_h(k)$ satisfies this condition, and for all h there is at least one k such that $|A_h(k)| \geq 1 + C_1\Delta t$ ($C_1 > 0$ and independent of h), then we say that the scheme is "weakly unstable." Note that some modes in the solution of the difference equation will grow like e^{Ct_n} in this case. However, the solutions of the differential equation $u_t = \sigma u_{xx} + au_x$ do not grow; in fact, they will decay. This weak instability can make a difference scheme useless for some applications, such as long-running problems in numerical weather prediction.

The first condition in the above problem is our normal stability condition for the heat equation. If σ is very small, we can expect trouble since for $\sigma = 0$ the above scheme is an unstable approximation to the hyperbolic equation $u_t = au_x$. And indeed we do have trouble as σ approaches zero since the second condition requires $\Delta t < 2\sigma/a^2$. For fixed σ and a this condition will certainly be satisfied if we take the mesh spacing Δx to be sufficiently small since $\mu < \frac{1}{2}$ implies $\Delta t < \Delta x^2/(2\sigma)$. But this may require a very small Δx. Our general argument shows that the lower order term $a\partial u/\partial x$ cannot influence stability if the mesh spacing is small enough. We may not wish to use such a small Δx. The numerical analyst must be somewhat suspicious of arguments which are true "for sufficiently small" Δx or Δt.

If σ is much smaller than a, then the above equation is more like a hyperbolic equation than is a parabolic equation. Therefore we might try the following difference scheme since it is stable for $\mu = 0$, $\lambda \neq 0$.

$$U_j^{n+1} = U_j^n + \frac{\mu}{2}\left[U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1} + U_{j+1}^n - 2U_j^n + U_{j-1}^n\right]$$
$$+ \frac{\lambda}{4}\left[U_{j+1}^{n+1} - U_{j-1}^{n+1} + U_{j+1}^n - U_{j-1}^n\right] \qquad (3.8\text{-}2)$$

where $\mu = \sigma\Delta t/\Delta x^2$, $\lambda = a\Delta t/\Delta x$.

Sometimes it is not desirable to make the first order terms implicit. This is particularly true if we are dealing with a system of equations where the first order terms involve several different variables but the second order term contains only the variable on the left side of the equation. An example is the following Navier-Stokes equations.

$$u_t + uu_x + vu_y + p_x = \frac{1}{R}\left(u_{xx} + u_{yy}\right) = \frac{1}{R}\nabla^2 u$$

$$v_t + uv_x + vv_y + p_y = \frac{1}{R}\nabla^2 v \quad , \qquad u_x + v_y = 0$$

The second order term in the u equation ($\nabla^2 u$) involves only u, and in the v equation only v, therefore we can use an implicit formula to difference this term and have only a tridiagonal matrix equation to solve. If we made the first order terms implicit, we would have a non-tridiagonal matrix equation to solve. We might also have to solve a nonlinear equation.

If we difference the lower order hyperbolic term in the leapfrog style and make the parabolic term implicit, then we can modify the difference scheme of equation (3.8-2) to obtain the following:

$$U_j^{n+1} = U_j^{n-1} + \mu\left[U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1} + U_{j+1}^{n-1} - 2U_j^{n-1} + U_{j-1}^{n-1}\right]$$
$$+ \lambda\left(U_{j+1}^n - U_{j-1}^n\right) \qquad (3.8\text{-}3)$$

Problem 3.8-2. Analyze the stability of the schemes given in equations (3.8-2) and (3.8-3).

Next we will consider a type of weak instability which can arise in the solution of the equation $u_t + cu_x + au = 0$ by use of the leapfrog scheme.

Problem 3.8-3. Assume the solution of the hyperbolic equation $u_t + cu_x + au = 0$ has periodic boundary conditions $u(-1,t) = u(1,t)$ and initial value $u(x,0) = f(x)$. Obtain the Fourier series representation for the solution $u(x,t)$.

$$u(x,t) = \sum_{k=-\infty}^{\infty} a_k \, e^{ik(x-ct)-at}$$

Hint: Let $f(x) = a_k e^{ikx}$, $u(x,t) = A(t)e^{ikx}$ and solve for $A(t)$. Assume $\sum_{-\infty}^{\infty} k|a_k| < \infty$, then show that the Fourier series is a solution.

Note the factor e^{-at} in the solution. Such an exponential decay in the solution can cause trouble when the leapfrog scheme is used. This scheme possesses a "weak instability" similar to that shown by Milne's method for ordinary differential equations [Henrici, 1962, p. 242]. If the leapfrog scheme is used to solve the primitive equations which govern the motion of the atmosphere, the terms representing the Coriolis force can cause such a weak instability [Kasahara, Washington, O'Brien]. These are undifferentiated terms like au in the above equations. The leapfrog scheme for the above equation is

$$U_j^{n+1} = U_j^{n-1} - \lambda\left(U_{j+1}^n - U_{j-1}^n\right) - 2\Delta t a U_j^n \qquad \lambda = c\Delta t/\Delta x$$

Problem 3.8-4. Assume that $U_j^n = K^{(n)} e^{ikx_j}$. Show that

$$K^{(n)} = AZ_+^n + BZ_-^n$$

where $Z_\pm$ are roots of

$$Z^2 + 2i\gamma Z - 1 = 0, \quad \gamma = \beta - i\alpha, \quad \beta = \lambda \sin k\Delta x, \quad \alpha = a\Delta t$$

$$Z_\pm = -i\gamma \pm \sqrt{1 - \gamma^2}$$

Note that if $a = 0$, then $|Z_\pm| = 1$. Show that if $|\gamma| \ll 1$, then

$$Z_\pm = \pm (1 - \tfrac{1}{2}\gamma^2) - i\gamma + 0(\gamma^4)$$

Show that for small γ we have $|Z_-| > 1$, $|Z_+| < 1$ if $a \neq 0$.

Therefore the term BZ_-^n will grow and we have a weak instability since the solution $u(x,t)$ should decay as t increases. Note that $Z_- = -1 + 0(\Delta t)$ and thus $|Z_-^n| \leq e^{Kt}$ for some K. Also $B = 0(\Delta t^2)$ and therefore we have a weak instability since the growth will not be objectionable if Δt is small enough. However, if we have to run out to large values of t, it may not be possible to take Δt small enough to insure that the term Be^{Kt} is small. Therefore we might consider the following scheme.

Problem 3.8-5. Analyze the stability of the following scheme for $u_t + cu_x + au = 0$ to show that it does not suffer from the weak instability described above.

$$U_j^{n+1} = U_j^{n-1} - \lambda\left(U_{j+1}^n - U_{j-1}^n\right) - a\Delta t\left(U_j^{n+1} + U_j^{n-1}\right), \quad \lambda = c\Delta t/\Delta x$$

Note that this scheme is effectively explicit.

3.9 An Experiment Concerning the Accuracy of Finite Difference Approximations.

In these experiments we solve the equation $u_t + u_x = 0$ for $0 \le x \le L$, $u(x,0) = f(x)$. In the first three cases we used the Lax-Wendroff scheme

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{\lambda^2}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right), \qquad \lambda = \Delta t/\Delta x \qquad (3.9\text{-}1)$$

In the first case

$$f(x) = \begin{cases} 0 & 0 \le x \le \pi \\ \frac{1}{2}+\frac{1}{2}\cos x & \pi \le x \le 2\pi \\ 1 & 2\pi \le x \le L \end{cases}$$

The boundary conditions are $u(0,t) = 0$, $u(L,t) = 1$. As long as $t \le (L-2\pi)$, the solution is $u(x,t) = f(x-t)$, and thus we can compute the error for t in this range. We computed the solution of equation (3.9-1) and also the error which is given in the table below.

	Error $\lambda = 0.99$		
Time	$\Delta x = 0.377$	$\Delta x = 0.188$	$\Delta x = 0.094$
0.95	9.0(-4)	4.4(-4)	2.2(-4)
1.90	1.7(-3)	8.5(-4)	3.5(-4)
4.04	3.4(-3)	1.4(-3)	5.8(-4)

In the second case we used the same Lax-Wendroff scheme with periodic boundary conditions $u(0,t) = u(2\pi,t)$ and initial function $f(x) = \sin x$, $0 \le x \le 2\pi$. The error is given in the second table.

	Error $\lambda = 0.99$		
Time	$\Delta x = 0.251$	$\Delta x = 0.126$	$\Delta x = 0.063$
0.95	3.1(-4)	8.5(-5)	2.0(-5)
3.96	1.3(-3)	3.3(-4)	8.2(-5)
8.01	2.6(-3)	6.6(-4)	1.6(-4)

The third case is the same as the second except $\lambda = 0.8$.

Error $\lambda = 0.8$

Time	$\Delta x = 0.251$	$\Delta x = 0.126$	$\Delta x = 0.063$
1.02	6.1(-3)	1.5(-3)	3.7(-4)
3.96	2.3(-2)	6.0(-3)	1.5(-3)
8.06	4.8(-2)	1.2(-2)	3.0(-3)

In the fourth case we used the scheme

$$U_j^{n+1} = \frac{U_{j+1}^n + U_{j-1}^n}{2} - \frac{\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right) \qquad \lambda = \Delta t/\Delta x$$

with periodic boundary conditions $u(0,t) = u(2\pi,t)$ and $u(x,0) = f(x) = \sin x$. The error is given below.

Error $\lambda = 0.99$

Time	$\Delta x = 0.251$	$\Delta x = 0.126$	$\Delta x = 0.063$
0.95	3.7(-3)	2.0(-3)	9.8(-4)
3.96	1.5(-2)	8.0(-3)	4.0(-3)
8.08	3.1(-2)	1.6(-2)	8.0(-3)

Problem 3.9-1. Derive the following expression for the truncation error of the Lax-Wendroff scheme.

$$\tau = \frac{\Delta t^3}{6}\left(u_{t^3} + \lambda^{-2}u_{x^3}\right) + \frac{\Delta t^4}{24}\left(u_{t^4} - \lambda^{-2}u_{x^4}\right) + O(\Delta t^5)$$

We denote $\partial^3 u/\partial t^3$ by u_{t^3}. We have assumed that all the derivatives required for the above derivation are continuous.

Problem 3.9-2. Explain the following facts concerning the above results: 1) In the first table we seem to have $E(\Delta x) = 0(\Delta x)$ and in the second $E(\Delta x) = 0(\Delta x^2)$. Here $E(\Delta x)$ denotes the error. 2) The error is much smaller in the second table than in the third. 3) In the fourth table $E(\Delta x) = 0(\Delta x)$.

4. DIFFERENCE SCHEMES FOR SYSTEMS OF EQUATIONS

Thus far we have only considered problems with a single equation for a single unknown function u. Usually one has several unknown functions and must thus deal with a system of equations. For example, the equations for incompressible two-dimensional viscous fluid flow are the following (u and v are the velocity components in the x and y directions, p the pressure and R the Reynolds number).

$$u_t + uu_x + vu_y + p_x = \frac{1}{R}(u_{xx} + u_{yy})$$

$$v_t + uv_x + vv_y + p_y = \frac{1}{R}(v_{xx} + v_{yy})$$

$$u_x + v_y = 0$$

Note that these are nonlinear equations. Only two of these equations involve a time derivative. The pressure must be obtained by some means other than a marching procedure--it is a diagnostic variable rather than a prognostic variable. We will say more about this in section 4.5. In order to set up a finite difference scheme, we need more than the above equations--we must specify the boundary and initial conditions. The proper treatment of boundary conditions causes the numerical analyst considerable difficulty. Here there is but little theory to guide him. If our theoretical stability analysis implies that a difference scheme is stable, then it usually is, except an instability may develop near the boundary. Sometimes we also have a failure of the theory because of nonlinear terms. Our stability analysis is usually valid only for linear equations with periodic boundary conditions. However, the extension from a single equation to a system usually does not cause a problem in practice, although it may make a theoretical analysis more difficult. Also, it may greatly increase the computer time required for a solution. Most schemes that are stable for a single equation will also be stable for a system of the same type.

4.1 VonNeumann stability analysis for systems of equations. The von Neumann condition for a single equation is based on the computation of the amplification factor for the scheme. This is a complex number $M_h(k)$ and we require $M_h(k) = 1 + O(\Delta t)$ for stability. In the case of a system of equations this amplification factor is a matrix $M_h(k)$. If the scheme involves only two time levels $(U^{n+1} = L_h U^n)$ and there are N unknown functions $u_1(x,t)$, ..., $U_N(x,t)$, then the order of this amplification matrix $M_h(k)$ is equal to the number of unknown functions, namely N. To analyze stability we must determine a bound for the power of the matrix operator L_h, that is $\|L_h^n\| \leq M$ for $n\Delta t \leq T$. By using the Fourier representation, we reduce the stability problem to that of finding a bound for the power of the amplification matrix $M_h(k)$. We pay a price for this reduction, since our stability analysis is now valid only for periodic boundary conditions. In the case of a single unknown (N=1) this factor $M_h(k)$ is a scaler, which makes the analysis much easier. For the case $N > 1$ we must deal with the norm of a matrix; that is, find a bound $\|M_h^n(k)\| < M$. The original matrix operator L_h has order approximately N*J where J is the number of mesh points. Thus the Fourier representation has reduced the order of the matrix considerably, but we must still deal with a matrix. If the norm $\|M_h^n(k)\|$ is bounded independent of the mesh spacing h and the wave number k, then the eigenvalues $\lambda_i(h,k)$ of $M_h(k)$ must satisfy the condition

$$|\lambda_i(h,k)| \le 1 + C\Delta t \qquad (4.1\text{-}1)$$

where C is independent of h and k. The reader should verify this statement. The above condition on the eigenvalue is the von Neumann necessary condition for stability. It is frequently easier to find such a bound for the eigenvalues than it is to find one for the norm. Considerable effort has been expended to find conditions on the matrix or difference scheme which will insure that the von Neumann condition is sufficient for stability, as well as necessary (see Richtmyer and Morton, 1967). It is usually rather difficult to bound the norm of a non-symmetric matrix. Note that the von Neumann condition is sufficient if there is only one unknown function (N=1).

We will now consider the von Neumann method to determine the stability of schemes for systems of equations. This is based on Fourier analysis for a vector. This is a trivial extension of the scaler case--we simply look at each component separately. Suppose we have a vector function $u(x)$, $-1 \le x \le 1$; that is, $u(x) = (u_1(x), u_2(x), \ldots, u_N(x))$. We may represent each component in a Fourier series.

$$u_\nu(x) = \sum_{-\infty}^{\infty} a_{\nu k}\, e^{i\pi kx} .$$

Then the function can be represented in the form

$$u(x) = \sum_{-\infty}^{\infty} a_k\, e^{i\pi kx}$$

where the a_k are vectors of order N. Similarly for a mesh function we have a finite Fourier analysis

$$U_j = \sum_{k=-J}^{J-1} a_k e^{i\pi k x_j}$$

$$a_k = \frac{1}{2J} \sum_{j=-J}^{J-1} U_j e^{i\pi k x_j}$$

where U_j and a_k are vectors.

For an example we will use the wave equation $u_{tt} - c^2 u_{xx} = 0$. We will write this as a system of equations

$$v_t - cw_x = 0 \qquad v = v(x,t)$$

$$-1 \le x \le 1$$

$$w_t - cv_x = 0 \qquad w = w(x,t)$$

The function v is then a solution of the wave equation. We assume periodic boundary conditions. The initial conditions are $v(x,0) = f_1(x)$, $w(x,0) = f_2(x)$. We can write this system in matrix form as $u_t + Au_x = 0$, $u(x,0) = f$, where

$$u = \begin{vmatrix} v \\ w \end{vmatrix} \qquad A = \begin{vmatrix} 0 & -c \\ -c & 0 \end{vmatrix} \qquad f = \begin{vmatrix} f_1 \\ f_2 \end{vmatrix} .$$

The mesh is $x_j = j/J$, $-J \le j \le J$. The notation is the same as for a single equation

$$u_j^n = \begin{vmatrix} v_j^n \\ w_j^n \end{vmatrix} .$$

The Lax-Wendroff scheme can be derived by the same sort of Taylor series as before except we are now dealing with vectors and matrices. Namely, $u_j^{n+1} = u_j^n + \Delta t {u_t}_j^n + \frac{1}{2}\Delta t^2 {u_{tt}}_j^n + O(\Delta t^3)$, and $u_{tt} = A^2 u_{xx}$. Therefore the scheme is

$$U_j^{n+1} = U_j^n - \frac{B}{2}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{B^2}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right)$$

$$B = \frac{\Delta t}{\Delta x}A = \begin{vmatrix} 0 & -\lambda \\ -\lambda & 0 \end{vmatrix}, \qquad \lambda = c\Delta t/\Delta x \ .$$

Next we use the Fourier representation

$$U_j^n = \sum_{k=-J}^{J-1} [M_h(k)]^n a_k e^{i\pi k x_j} \tag{4.1-2}$$

where $M_h(k)$ is the amplification matrix. If we substitute this into the difference scheme and equate the coefficients of the complex exponentials, we obtain an equation for the factor $M_h(k)$. This is exactly the same as for a single equation, except we use vectors and matrices instead of scalers. We can obtain the same expression for $M_h(k)$ if we let $U_j^n = e^{i\pi k x_j}$ and substitute this expression for U_j^n into the difference scheme (that is, take the Fourier transform or work with one frequency component at a time). We obtain

$$U_j^{n+1} = M_h(k)U_j^n = (I - i\sin\theta B + (\cos\theta - 1)B^2)U_j^n, \qquad \theta = \pi k \Delta x$$

and thus $M_h(k) = I - i\sin\theta B + (\cos\theta - 1)B^2$. This is the same as the single equation case (equation (3.2-2)) if we replace the scaler λ by the matrix B.

A little computation shows us that the matrix $M_h(k)$ is

$$M_h(k) = \begin{vmatrix} 1 + \lambda^2(\cos\theta - 1) & i\lambda\sin\theta \\ i\lambda\sin\theta & 1 + \lambda^2(\cos\theta - 1) \end{vmatrix} \qquad (4.1\text{-}3)$$

To determine the eigenvalues of this matrix $M_h(k)$ we can use the following result.

Problem 4.1-1. Let the matrix M of order N be represented by a polynomial in the matrix B; that is, $M = C_0 I + C_1 B + \dots + C_n B^n$. If the eigenvalues of B are ξ_i $1 \le i \le N$, then the eigenvalues λ_i of M are $\lambda_i = C_0 + C_1\xi_i + C_2\xi_i^2 + \dots + C_n\xi_i^n$. Hint: For any matrix B there is a unitary matrix U such that $U^*U = I$ and U^*BU is upper triangular. The eigenvalues of a triangular matrix are the diagonal elements.

Using the result of this problem we see that the eigenvalues of M are $\xi_i = 1 \pm i\lambda\sin\theta + \lambda^2(\cos\theta - 1)$. Note that the eigenvalues of B are $\pm\lambda$. We have already shown in section 3.2 that $|\xi_i| \le 1$ independent of $\theta = \pi k\Delta x$. Therefore the von Neumann criterion is satisfied for this difference scheme. The von Neumann criterion is only a necessary condition so we still have no proof that the scheme is stable.

In order to prove stability we must find a bound for the following norm $\|[M_h(k)]^n\|$. We will prove that $\|M_h(k)\| \le 1$ independent of k and h ($h = \Delta x$) provided $|\lambda| \le 1$ ($\lambda = c\Delta t/\Delta x$). This provides a bound for the norm of the n^th^ power of the matrix, namely $\|[M_h(k)]^n\| \le 1$. We will use the Euclidean (sometimes called L_2) norm. The elements in our matrix are complex numbers since we used the complex form of the Fourier series.

The norm of a vector $\{u_i\}$, $1 \le i \le N$, containing complex elements is

$$\|u\|_2 = \sqrt{\sum_{i=1}^{N} \bar{u}_i u_i} = \sqrt{\sum_{i=1}^{N} |u_i|^2}$$

(here $\bar{u}_i$ denotes the complex conjugate of u_i). The norm of the matrix M is defined by

$$\|M\|_2 = \max_{\|u\|=1} \|Mu\|_2$$

Problem 4.1-2. We define the spectral radius of a matrix A as $\sigma(A) = \max_j |\lambda_j(A)|$ where $\lambda_h(A)$ are the eigenvalues of A. We denote the transpose conjugate of a matrix A by A^*, thus the elements of A^* are $(a^*)_{ij} = \bar{a}_{ji}$. Prove the following relation for the L_2 norm of a matrix, $\|M\|_2 = \sqrt{\sigma(M^*M)}$. See problem 1.2-34.

Now we will estimate the norm of the matrix M given by equation (4.1-3). We could use the result of problem 4.1-2, but we will give an independent proof since our matrix is so easy to deal with. If we denote the elements of M by m_{ij}, $1 \le i, j \le N$ and let $w = Mu$, then $\|Mu\| = \sum_{i=1}^{N} \bar{w}_i w_i =$

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{N} \overline{m_{ij}u_j}\right)\left(\sum_{k=1}^{N} m_{ik}u_k\right) = \sum_{j=1}^{N} \bar{u}_j \sum_{k=1}^{N}\left(\left(\sum_{i=1}^{N} \bar{m}_{ij}m_{ik}\right)u_k\right).$$

A moment's reflection will show that we can write the right hand triple sum as $u^*(M^*M)u$ (note that $w^*Aw = \sum_j \sum_i \bar{w}_i\, a_{ij}\, w_j$ for any vector w and matrix A). If we compute the matrix M^*M for the matrix M of equations (4.1-2) we obtain the scaler matrix

$$\begin{vmatrix} \alpha^2+\beta^2 & 0 \\ 0 & \alpha^2+\beta^2 \end{vmatrix} = \left|\alpha^2+\beta^2\right| \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \qquad \begin{aligned} \alpha &= 1 + \lambda^2(\cos\theta-1) \\ \beta &= \lambda\sin\theta \end{aligned}$$

If $\|u\|_2 = 1$, then $u^*u = 1$ and $u^*M^*Mu = (\alpha^2+\beta^2)u^*Iu = (\alpha^2+\beta^2)u^*u = \alpha^2+\beta^2$. In section 3.2 we have shown that $\alpha^2 + \beta^2 = 1 - \lambda^2(1-\cos\theta)^2(1-\lambda^2)$. Obviously $(\alpha^2+\beta^2) \leq 1$ independent of θ (thus independent of h) provided $|\lambda| < 1$. Hence we have stability.

Normally it is not so easy to determine a bound for the norm of the powers of the amplification matrix $\|M_h^n(k)\|$. In practice, instead of computing the norm, we usually use the von Neumann stability criterion. This requires knowledge of the eigenvalues of $M_h(k)$ rather than the norm. Determination of the eigenvalues can be difficult but usually not so difficult as the norm. The von Neumann condition is a necessary condition for stability, but it is not sufficient. A stable scheme must satisfy the the von Neumann condition, but a scheme which satisfies this condition may not be stable. The von Neumann condition simply requires that the eigenvalues λ_i of $M_h(k)$ satisfy the condition $|\lambda_i| = 1 + O(\Delta t)$; that is, there exists a constant c independent of h $(h=\Delta x)$ and the frequency k such that $|\lambda_i| \leq 1 + c\Delta t$. The fact that this condition is only necessary for stability and not sufficient, is not too serious. After all, this analysis of stability is dependent on the Fourier representation of the solution. This only works for linear equations with constant coefficients and periodic boundary conditions. Usually these requirements are not met in practice, so our theory is not rigorously applicable. However, the von Neumann criterion is an invaluable guide to the selection of finite difference schemes.

If M is a normal matrix, then the von Neumann condition is sufficient to bound the norm $\|M^n\|$ of the powers of M. Other sufficient conditions are given in chapter 4 of the book by Richtmyer and Morton.

Problem 4.1-3. Consider the hyperbolic system $u_t + Au_x = 0$ where A is a real matrix with real, distinct eigenvalues and u is a vector function. Assume we have an initial value problem with periodic boundary conditions. Use the Friedrichs finite difference scheme.

$$U_j^{n+1} = \tfrac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) - \frac{\lambda}{2} A\left(U_{j+1}^n - U_{j-1}^n\right), \qquad \lambda = \Delta t/\Delta x .$$

Use the von Neumann criterion to obtain the stability condition for this scheme ($|\Delta t \xi_i/\Delta x| < 1$ where ξ_i are the eigenvalues of A). Show that this is also a sufficient condition for stability in this case.

Problem 4.1-4. Show that the implicit scheme below for the hyperbolic system of problem 4.1-3 is unconditionally stable (use the von Neumann criterion).

$$U_j^{n+1} = U_j^n - \frac{\lambda}{4} A\left(U_{j+1}^{n+1} - U_{j-1}^{n+1} + U_{j+1}^n - U_{j-1}^n\right), \qquad \lambda = \Delta t/\Delta x$$

If $-J \le j < J$ and if there are N components in the vectors U_j^n, then at each time step we must solve a linear equation of order 2JN. Show that the matrix for this system of equations is block-tridiagonal. Is this matrix non-singular? What method would you use to solve this matrix equation on a computer? Can you provide an operational count for this method?

Problem 4.1-5. Consider the parabolic system of equations $u_t = Du_{xx}$. Here u is a vector of order N and D is a symmetric positive definite matrix of order N. Assume an initial value problem with periodic boundary conditions on the interval $-1 \le x \le 1$. Define the energy in this system by $\|u\|^2 = \int_{-1}^{1} \left(\sum_{i=1}^{N} u_i^2(x,t) \right) dx$ (u is the vector whose components are $u_i(x,t)$). Show that this energy is non-increasing. Extend the Crank-Nicholson implicit scheme of section 3.4 to this system. Use the von Neumann criterion to show that the scheme is unconditionally stable. The implicit scheme requires the solution of a matrix equation at each time step. Show that this matrix is nonsingular.

Problem 4.1-6. Find a family of matrices M_h depending on a real parameter h such that $\|M_h\|_\infty \le 2$ and $\sigma(M_h) < 1$ for all h. Here $\|M_h\|_\infty$ is based on the maximum norm and $\sigma(M_h)$ is the spectral radius of M_h. The family should have the additional property that $\|M_h^n\|$ is not bounded as $n \to \infty$ even if $hn < 1$. This is most easily done with a matrix of order 2.

Problem 4.1-7. What is the error in the following argument? Let the family of matrices M_h of order N have spectral radius bounded by $\sigma(M_h) \le 1 + c\Delta t$ (assume $\Delta t = \Delta t(h)$ is a function of h). Assume there are N independent eigenvectors of M_h for each h. Denote these eigenvectors by $v_{h,i}$ $1 \le i \le N$. Then for any vector u we have $u = \sum_{i=1}^{N} \alpha_i v_{h,i}$. Therefore if $\|u\| = 1$, we have $\|M_h^n u\| = \| \sum_{i=1}^{N} \lambda_{h,i}^n \sigma_i v_{h,i} \|$. Since $|\lambda_{h,i}| \le 1 + c\Delta t$, $|\lambda_{h,i}^n| \le e^{Ct_n}$, and therefore $\|M_h^n u\| \le e^{Ct_n}$ provided $\|u\| \le 1$.

4.2 Multilevel Difference Schemes. The leapfrog scheme for the hyperbolic equation $u_t + cu_x = 0$ is an example of a multilevel difference scheme.

$$U_j^{n+1} = U_j^{n-1} - \lambda\left(U_{j+1}^n - U_{j-1}^n\right) \qquad \lambda = c\Delta t/\Delta x \qquad (4.2\text{-}1)$$

We must know U^{n-1} and U^n to predict U^{n+1}. Many schemes use only the single time level U^n to predict U^{n+1}. Our theory of stability--the von Neumann criterion, for example--applies only to such single-level schemes. However, we can reduce a multilevel scheme to a single-level scheme. Consider the leapfrog scheme given above. Define the column vector W^n by $W^n = \left(U_{-J}^{n+1}, U_{-J+1}^{n+1}, \ldots, U_{J-1}^{n+1}, U_{-J}^n, U_{-J+1}^n, \ldots, U_{J-1}^n\right)^T$. That is $W^n = \left(U^{n+1}, U^n\right)^T$. The difference scheme for U^n can be written as $U^{n+1} = IU^{n-1} + BU^n$ where I is the identity matrix of order 2J and B is also a matrix of order 2J, namely,

$$B = \begin{vmatrix} 0 & -\lambda & 0 & \cdots & & & \lambda \\ \lambda & 0 & -\lambda & 0 & \cdots & & 0 \\ 0 & \lambda & 0 & -\lambda & 0 & \cdots & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ -\lambda & \cdots & & & 0 & \lambda & 0 \end{vmatrix}$$

Since $U^{n+2} = IU^n + BU^{n+1}$, we can write the scheme in terms of W^n as

$$W^{n+1} = L_h W^n \quad \text{where } L_h = \begin{vmatrix} B & I \\ I & 0 \end{vmatrix} \qquad (4.2\text{-}2)$$

Note that L_h is a matrix of order 4J. To determine the stability of the difference scheme for W^n we must investigate $\|L_h^n\|$. Our proof that stability implies convergence for a consistent scheme no longer applies to this multilevel scheme. We have changed the structure of our problem since the values of W approximate the solution u on two time levels. However this change creates no real problem. We refer the reader to chapter 7 of the book by Richtmyer and Morton.

We will now use the Fourier series method to perform a stability analysis for the difference scheme written in terms of W^n. The vector $\{W^n_j\}$ $1 \le j \le 4J$ can be represented in the following form (why?)

$$W^n_j = \begin{cases} \sum\limits_{k=-J}^{J-1} a_k^{(n)} e^{i\pi k x_j} & 1 \le j \le 2J, \quad x_j = (j-J-1)/J \\[2ex] \sum\limits_{k=-J}^{J-1} b_k^{(n)} e^{i\pi k x_j} & 2J+1 \le j \le 4J, \quad x_j = (j-3J-1)/J \end{cases}$$

Note the coefficients $a_k^{(n)}$ and $b_k^{(n)}$ are the Fourier coefficients for U^{n+1} and U^n respectively.

Problem 4.2-1. Substitute the above expression for W^n into equation (4.2-2) or equivalently, into (4.2-1). Obtain the following equation for the coefficients $a_k^{(n)}$ and $b_k^{(n)}$.

$$\begin{vmatrix} a_k^{(n+1)} \\ b_k^{(n+1)} \end{vmatrix} = M_h(k) \begin{vmatrix} a_k^{(n)} \\ b_k^{(n)} \end{vmatrix}$$

Find an expression for the 2×2 amplification matrix $M_h(k)$. Note that

$$\left|\begin{matrix} a_k^{(n)} \\ b_k^{(n)} \end{matrix}\right| = [M_h(k)]^n \left|\begin{matrix} a_k^{(0)} \\ b_k^{(0)} \end{matrix}\right|$$

where the coefficients $a_k^{(0)}$ and $b_k^{(0)}$ are determined from the given initial vector W^0 (we must have U^0 and U^1 in order to define W^0). Find the eigenvalues of the matrix $M = M_h(k)$. Show that the maximum norm $\|M\|_\infty$ is bounded independent of Δx and n. Using this result prove that the scheme is stable. Show that the L_2 norm of M is greater than one. Does the L_2 norm satisfy the condition $\|M_h(k)\|_2 \leq 1 + 0(\Delta t)$ independent of h and k. Given any matrix must this condition be satisfied in order that $\|M_h^n(k)\|$ be bounded? Note that the L_2 norm is given by $\|M\|_2 = \sqrt{\sigma(M^*M)}$ where $\sigma(A)$ denotes the spectral radius of A.

Problem 4.2-2. Use the von Neumann criterion to show that the following multilevel scheme for the heat equation $u_t = \sigma u_{xx}$ is stable.

$$-3U_j^{n+1} + 4U_j^n - U_j^{n-1} = 2\mu\left(U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}\right), \qquad \mu = \sigma\Delta t/\Delta x^2$$

4.3 The Courant-Friedrichs-Lewy stability condition for hyperbolic equations. We will first look at the wave equation written as a system of two first-order equations.

$$\frac{\partial v}{\partial t} - c\frac{\partial w}{\partial x} = 0 \qquad\qquad -1 \leq x \leq 1$$

$$\frac{\partial w}{\partial t} - c\frac{\partial v}{\partial x} = 0 \qquad\qquad 0 \leq t$$

$$v(x,0) = v_0(x)$$

$$w(x,0) = w_0(x)$$

We assume periodic boundary conditions $v(x\pm2,t) = v(x,t)$, $w(x\pm2,t) = w(x,t)$. If we change variables by setting $u_1 = v+w$ and $u_2 = v-w$, then the above equations become

$$\frac{\partial u_1}{\partial t} - c\,\frac{\partial u_1}{\partial x} = 0$$

$$\frac{\partial u_2}{\partial t} + c\,\frac{\partial u_2}{\partial x} = 0$$

$$u_1(x,0) = v_0(x) + w_0(x)$$

$$u_2(x,0) = v_0(x) - w_0(x)$$

The solution of this system is

$$u_1(x,t) = v_0(x+ct) + w_0(x+ct)$$

$$u_2(x,t) = v_0(x-ct) - w_0(x-ct)$$

The value of u_1 is constant along the lines $x+ct = K$ and the value of u_2 is constant along the lines $x-ct = K$. If we consider any point $P = (x,t)$ in the x-t plane and draw the two lines downward from this point with slope $\pm c^{-1}$, we obtain a"domain of influence" for the point P. The values of u_1 and u_2 (hence v and w) at P are determined by the values at the intersection of these lines with the initial line $t = 0$ (see Figure 4.3-1 below).

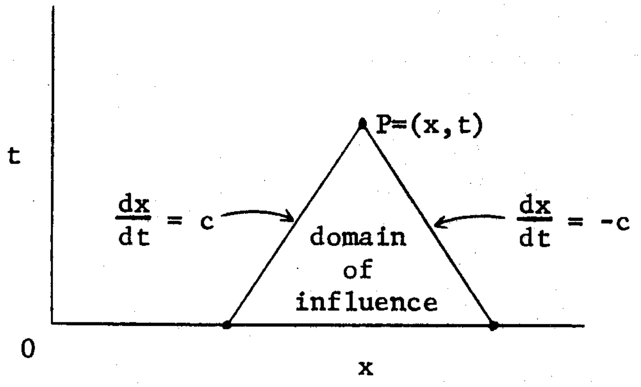


Figure 4.3-1

Values outside the domain of influence for the point P can have no effect on the values at P.

Next we will look at more general hyperbolic systems.

$$\frac{\partial u}{\partial t} + A \frac{\partial u}{\partial x} + Bu = g \tag{4.3-1}$$

where u is a vector of dimension N and A and B are constant matrices of order N (see section 1.3.7). We assume that there is a nonsingular matrix P such that $P^{-1}AP = D$ is a diagonal matrix. Then the diagonal elements are the eigenvalues of A and the columns of P are the eigenvectors. If we change variables to $w = P^{-1}u$, then we obtain the system

$$\begin{aligned} &\frac{\partial(P^{-1}u)}{\partial t} + P^{-1}APP^{-1} \frac{\partial u}{\partial x} + P^{-1}BPP^{-1}u = P^{-1}g \\ &\frac{\partial w}{\partial t} + D \frac{\partial w}{\partial x} + Cw = \hat{g} \end{aligned} \tag{4.3-2}$$

where $C = P^{-1}BP$, $\hat{g} = P^{-1}g$. Suppose we denote the elements of the diagonal matrix D by ξ_i, $1 \le i \le N$. Consider the lines $L_i(K)$ given by $x - \xi_i t = K$. Along such a line $w(x,t)$ is a function of t alone and the chain rule for derivatives gives us $\frac{dw}{dt} = w_t + \frac{dx}{dt} w_x = w_t + \xi_i w_x$. If we write out the vector equation (4.3-2) into components we have

$$\frac{\partial w_i}{\partial t} + \xi_i \frac{\partial w_i}{\partial x} = - \sum_{j=1}^{N} c_{ij} w_j + \hat{g}_i$$

or

$$\frac{dw_i}{dt} = - \sum_{j=1}^{N} c_{ij} w_j + \hat{g}_i$$

Therefore we can solve for w along any line by integration of a system of ordinary linear differential equations. Now consider any point (x,t) in the x-t plane. From this point we may draw the N lines with slope $dx/dt = \xi_i$ where ξ_i are the eigenvalues of A (see Figure 4.3-2).

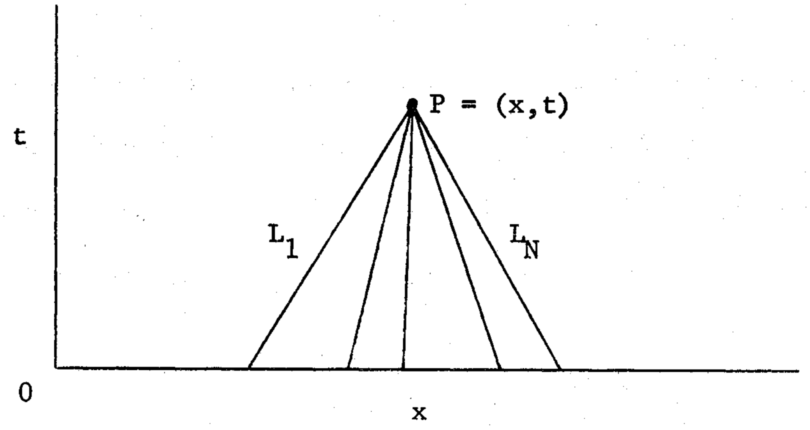


Figure 4.3-2

The values of w (and therefore of u = Pw) are found by integrating along the lines L_1 through L_N. Therefore only initial values which lie inside the triangle determined by these lines can influence the value at the point P. The lines L_i are called characteristics of the differential equation. Only hyperbolic equations can be properly solved by this "method of characteristics."

Next we will discuss the use of these concepts in the evaluation of finite difference schemes. Suppose we lay out a mesh to solve equations (4.3-1). We assume the mesh ratio $\Delta t/\Delta x = \lambda$ is constant, independent of Δx. Suppose we have an explicit three-point difference scheme. Then the values of U_j^{n+1} can be obtained from a knowledge of U_{j-1}^n, U_j^n, and U_{j+1}^n. The domain of influence of the point (x_j, t_{n+1}) in the difference scheme is the interval $[x_{j-1}, x_{j+1}]$ on the n^{th} time level and $[x_{j-2}, x_{j+2}]$ on the $(n-1)^{st}$ time level. It is clear that the domain of influence for the difference scheme is bounded by lines of slope $\Delta x/\Delta t = \pm\lambda^{-1}$ extending downward from a mesh point (see Figure 4.3-3).

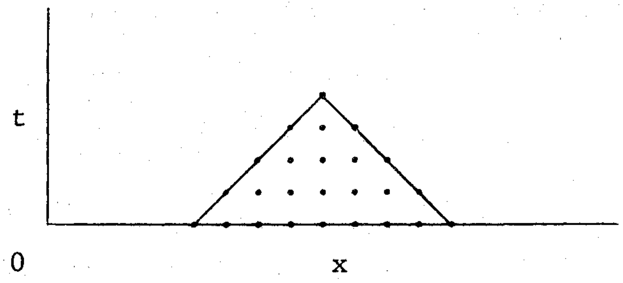


Figure 4.3-3

If the solution is initially zero on the lower side of the triangle, then it will be zero within the triangle. This is true even if we halve the mesh spacing since the triangle is determined by lines of slope $\pm\lambda^{-1}$ and λ^{-1} is not dependent on the mesh spacing. Now suppose the demain of influence for the differential equation is not contained within that for the finite difference scheme (see Figure 4.3-4).

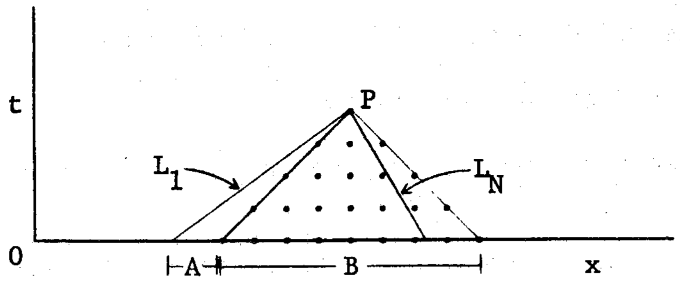


Figure 4.3-4

If the initial function on the interval A is not zero, then we can assume the value of the solution at the point P will not vanish--if it did vanish, we could change the initial function on A. However, if the initial function does vanish on the interval B, then the solution of the finite difference scheme must vanish at P regardless of how small we take the mesh spacing Δx. But this means convergence is impossible--zero values cannot converge to a nonzero value. This leads us to the following condition.

Definition 4.3-1. We say a finite difference scheme for a hyperbolic system satisfies the Courant-Friedrichs-Lewy (C-F-L) condition if the domain of influence of the finite difference scheme contains the domain of influence of the differential equation. The C-F-L condition is a necessary condition for stability, but it is not sufficient (can you find an example of a scheme which satisfies the C-F-L condition but is unstable?).

If the eigenvalues of A are ξ_i, then the C-F-L condition clearly requires that $\lambda^{-1} > |\xi_i|$ for $1 \le i \le N$, or $\Delta t|\xi_i|/\Delta x < 1$. Thus the value of Δt is limited. The values $\xi_i = \frac{dx}{dt}$ govern the speed with which waves or disturbances are propagated by the differential equation. A moment's reflection will show that the maximum speed of propagation for disturbances in the mesh is $\Delta x/\Delta t = \lambda^{-1}$. The C-F-L condition then states that the mesh disturbances must have a velocity no less than that of the physical disturbances.

Problem 4.3-1. Consider the hyperbolic system

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0 \qquad A = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix}$$

with initial conditions $u(x,0) = f(x)$ and boundary condition $u(0,t) = g(x)$. Use the "upstream difference" scheme

$$U_j^{n+1} = U_j^n - \lambda A\left(U_j^n - U_{j-1}^n\right). \qquad \lambda = \Delta t/\Delta x$$

Choose λ so that the C-F-L condition is satisfied for this scheme. Is it possible to satisfy the C-F-L condition with this scheme if A is given by

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

Problem 4.3-2. What is the domain of influence for the following implicit scheme for $u_t + cu_x = 0$?

$$U_j^{n+1} = U_j^n - \frac{\lambda}{2}\left(U_{j+1}^{n+1} - U_{j-1}^{n+1}\right) \qquad \lambda = c\Delta t/\Delta x\ .$$

What does the C-F-L condition say about the stability of this scheme? Does this agree with the von Neumann criterion?

4.4 Implicit difference schemes for systems. Implicit schemes offer the same unconditional stability for systems as for a single equation.

However, the solution of the implicit equations for a system can require a prohibitive amount of arithmetic (see problem 4.1-4). At each time step it is necessary to solve a block-tridiagonal matrix equation rather than a scaler tridiagonal matrix equation. If there are N variables in the system, then these blocks are N×N matrices. The amount of arithmetic required for each block goes up roughly in the order N^3.

Suppose we have a hyperbolic system of equations

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0$$

The Crank-Nicholson type of implicit scheme for this system is given in problem 4.1-4. It is

$$U_j^{n+1} = U_j^n - \frac{\lambda}{4}A\left(U_{j+1}^{n+1} - U_{j-1}^{n+1} + U_{j+1}^n - U_{j-1}^n\right)$$

As noted above, we must solve a block tridiagonal matrix equation at each time step to obtain U^n. We might try a modification of this implicit scheme which requires only the solution of N scaler tridiagonal matrix equations each time step (N is the order of the matrix A; that is, the number of unknowns in the original equation). The idea is very similar to the Gauss-Seidel iteration for the inversion of a matrix. We illustrate this under the assumption that A is of order three (N=3). We use the notation $U_{\hat{x},j} = U_{j+1} - U_{j-1}$. Let the components of U_j^n be $U_{1,j}^n$, $U_{2,j}^n$ and $U_{3,j}^n$. The scheme is the following:

$$U_1^{n+1} = U_1^n - \frac{\lambda}{4}a_{11}\left(U_{1\hat{x}}^{n+1} + U_{1\hat{x}}^n\right) - \frac{\lambda}{4}a_{12}\left(2U_{2\hat{x}}^n\right) - \frac{\lambda}{4}a_{13}\left(2U_{3\hat{x}}^n\right) \qquad (4.4\text{-}1)$$

$$U_2^{n+1} = U_2^n - \frac{\lambda}{4} a_{21}\left(U_{1\hat{x}}^{n+1} + U_{1\hat{x}}^n\right) - \frac{\lambda}{4} a_{22}\left(U_{2\hat{x}}^{n+1} + U_{2\hat{x}}^n\right) - \frac{\lambda}{4} a_{23}\ 2U_{3\hat{x}}^n$$

$$U_3^{n+1} = U_3^n - \frac{\lambda}{4} a_{31}\left(U_{1\hat{x}}^{n+1} + U_{1\hat{x}}^n\right) - \frac{\lambda}{4} a_{32}\left(U_{2\hat{x}}^{n+1} + U_{2\hat{x}}^n\right) - \frac{\lambda}{4} a_{33}\left(U_{3\hat{x}}^{n+1} + U_{3\hat{x}}^n\right)$$

We have dropped the mesh subscript j from $U_{1,j}^n$, thus U_1^n does not mean the three-dimensional vector with j = 1, but rather the 2J dimensional vector formed from the first component of U_j^n (a terrible notation). Note that this scheme involves the inversion of three scaler tridiagonal matrices at each time step.

Problem 4.4-1. Show that the truncation error of the scheme of equations (4.4-1) is $\tau = O(\Delta t) + O(\Delta x^2)$. How could you improve the accuracy to $\tau = O(\Delta t^2) + O(\Delta x^2)$?

The above scheme can be shown to be unconditionally stable if the matrix is symmetric and positive definite. If A is symmetric and not positive definite, then the scheme is unconditionally unstable [Gary, 1964].

Mitchell and his collaborators have a series of papers on implicit difference schemes for hyperbolic systems of equations. Some of these require nothing worse than the inversion of a scaler tridiagonal matrix [Mitchell, 1966].

4.5 An initial value problem coupled with a boundary value problem.

In some cases we have initial value problems in which some variables are not differentiated with respect to time. The Navier-Stokes equations for viscous fluid flow mentioned at the beginning of this chapter are one such example. We will consider a contrived example of such a system which is simple enough so that we can analyze it. The system is

$$\frac{\partial u}{\partial t} + \alpha w = \sigma \frac{\partial^2 u}{\partial x^2} \qquad u = u(x,t) \qquad (4.5\text{-}1)$$

$$\frac{\partial^2 w}{\partial x^2} + u = 0 \qquad w = w(x,t)$$

$$u(x,0) = f(x) \qquad 0 \le x \le 1 \qquad 0 \le t$$

The boundary conditions are $u(0,t) = u(1,t) = w(0,t) = w(1,t) = 0$. Note that once $u(x,t)$ is known for a given t, then w can be found for the same t by solving a two-point boundary value problem. For this we need u and the boundary conditions on w. Therefore we do not need an initial condition on w. If w were known, then we could find u by integrating the inhomogeneous heat equation

$$\frac{\partial u}{\partial t} = -\alpha w + \sigma \frac{\partial^2 u}{\partial x^2} .$$

We must solve for u and w simultaneously, but only for u do we use a marching method. Before we discuss difference schemes for this problem, we will study the differential equation.

Problem 4.5-1. Assume $f(x)$ can be represented as

$$f(x) = \sum_{k=1}^{\infty} a_k \sin \pi kx \qquad \text{where } \sum_{k=1}^{\infty} |a_k| < \infty.$$

Show that $u = \sum_{k=1}^{\infty} a_k e^{-\left(\frac{\alpha}{\pi^2 k^2} + \sigma\pi^2 k^2\right)t} \sin \pi kx$ is the solution. Find the expression for w. Prove that these series converge and satisfy the differential equation for $t > 0$.

Problem 4.5-2. Assume the functions u and w and the appropriate derivatives are continuous. Show that the following energy equality holds.

$$\tfrac{1}{2}\int_0^1 \left\{[u(x,t)]^2 - [u(x,0)]^2\right\}dx = -\alpha\int_0^t\int_0^1 (w_x)^2 dxdt - \sigma\int_0^t\int_0^1 (u_x)^2 dxdt$$

Use this result to show that the solution of equation (4.5-1) is unique, provided it is sufficiently smooth (assume $\alpha,\sigma > 0$).

Now we are ready to consider finite difference schemes for this problem. First we will look at the obvious analog of the explicit scheme for the heat equation.

$$U_j^{n+1} + U_j^n - \Delta t\alpha W_j^n = \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \qquad 1 \le j \le J-1 \qquad (4.5\text{-}2)$$

$$\mu = \sigma\Delta t/\Delta x^2$$

The values of W_j^n are obtained by solving the following system

$$W_{j+1}^n - 2W_j^n + W_{j-1}^n = \Delta x^2 U_j^n \qquad 1 \le j \le J-1 \qquad (4.5\text{-}3)$$

$$W_0^n = W_J^n = 0$$

The matrix for this system is tridiagonal. We start the integration by using the initial condition $U_j^0 = f(x_j)$, $x_j = j/J$. Then we can solve equations (4.5-3) for the vector W^0. Then we can use equation (4.5-2) to obtain the vector U^1. Now we can repeat the process finding first W^1, then U^2.

Problem 4.5-3. Using the finite sin series $U_j^n = \sum_{k=1}^{J-1} A_k \sin\pi k x_j$ determine a stability condition for the above difference scheme.

Consider the predictor-corrector scheme defined as follows. Assume U^n is known.

$$\hat{W}_{j+1} - 2\hat{W}_j + \hat{W}_{j-1} = -\Delta x^2 U_j^n \qquad 1 \le j \le J-1$$

$$\hat{W}_0 = \hat{W}_J = 0$$

$$\hat{U}_j = U_j^n - \Delta t\alpha\hat{W}_j + \frac{\mu}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n + \hat{U}_{j+1} - 2\hat{U}_j + \hat{U}_{j-1}\right)$$

$$\hat{U}_0 = \hat{U}_J = 0$$

$$W_{j+1}^{n+\frac{1}{2}} - 2W_j^{n+\frac{1}{2}} + W_{j-1}^{n+\frac{1}{2}} = -\frac{\Delta x^2}{2}\left(U_j^n + \hat{U}_j\right)$$

$$W_0^{n+\frac{1}{2}} = W_J^{n+\frac{1}{2}} = 0$$

$$U_j^{n+1} = U_j^n - \Delta t\alpha W_j^{n+\frac{1}{2}} + \frac{\mu}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n + U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}\right)$$

$$U_0^{n+1} = U_J^{n+1} = 0$$

Problem 4.5-4. Determine the truncation error and stability condition for this predictor-corrector scheme. Assume $\alpha > 0$ and $\sigma > 0$.

Next we will define an iterative difference scheme. We wish to iterate toward the solution of the following system of equations.

$$W_{j+1}^{n+\frac{1}{2}} - 2W_j^{n+\frac{1}{2}} + W_{j-1}^{n+\frac{1}{2}} = \frac{\Delta x^2}{2}\left(U_j^n + U_j^{n+1}\right) \qquad 1 \le j \le J-1$$

$$W_0^{n+\frac{1}{2}} = W_J^{n+\frac{1}{2}} = 0 \qquad (4.5\text{-}4)$$

$$U_j^{n+1} = U_j^n - \Delta t\alpha W_j^{n+\frac{1}{2}} + \frac{\mu}{2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n + U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}\right)$$

$$1 \le j \le J-1$$

$$U_0^{n+1} = U_J^{n+1} = 0$$

Problem 4.5-5. Suppose we are able to solve the above implicit equations (4.5-4) at each time step. Prove that this difference scheme is unconditionally stable (assume $\alpha,\sigma > 0$).

Problem 4.5-6. Extend the predictor-corrector scheme of problem 4.5-4 to obtain an iterative scheme for solving equations (4.5-4). Under what conditions will this iterative scheme converge?

Problem 4.5-7. Eliminate $W^{n+\frac{1}{2}}$ from equations (4.5-4) to obtain a matrix equation for U^{n+1} in the form $AU^{n+1} = BU^n$. Here A and B are matrices of order J-1. Show that A is pentadiagonal; that is, $|a_{ij}| = 0$ if $|i-j| > 2$. Prove that the matrix A is nonsingular.

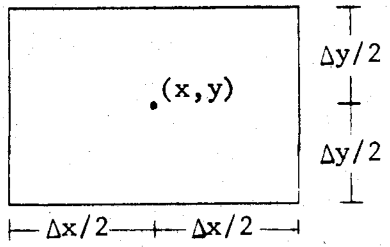
5. CONSTRUCTION OF DIFFERENCE SCHEMES FOR ELLIPTIC EQUATIONS

In this chapter we will be concerned with the construction of finite difference approximations for the solution of elliptic partial differential equations. This involves the selection of a set of mesh points to approximate the region and its boundary. Using this mesh we must derive a finite difference approximation to the differential equation. This leads to the question of convergence and error estimation. That is, how good is our approximation. These are the questions we will treat in this chapter. These schemes require the solution of a large system of linear equations. There is a formidable amount of literature concerning the solution of these large systems. We will consider these techniques in chapter 6. Much of the material in this chapter is patterned after chapter 6 in the book by Varga [1962]. The books by Wachspress [1966], Greenspan [1965], and Forsythe and Wasow [1960] have also proved useful, as did an article by Spanier [1967].

5.1 Derivation of the heat equation. We wish to determine an equation for the temperature $u(x,y,t)$ in a two-dimensional domain. We assume that the flux of heat energy through a line segment ds is $k(\partial u/\partial n)ds$ where k is the conductivity of the medium and $\partial u/\partial n$ is the derivative of u in a direction normal to the line segment. Let ρ be the density of our material, c the specific heat per unit mass, $\Delta u = u(x,y,t+\Delta t) - u(x,y,t)$ the change in temperature during the time interval Δt, $\Delta A = \Delta x\Delta y$ measure the area of a rectangular region containing the point (x,y). Then the increase of energy in the region is given by

$$\Delta E = \rho c \Delta A \Delta u$$

and this is equal to the energy which flows into the region due to the temperature gradient at the boundary. The region is pictured below.



The net flow of energy per unit time is

$$\left[k\left(x,y+\frac{\Delta y}{2}\right) u_y\left(x,y+\frac{\Delta y}{2}\right) - k\left(x,y-\frac{\Delta y}{2}\right) u_y\left(x,y-\frac{\Delta y}{2}\right)\right]\Delta x$$

$$+\left[k\left(x+\frac{\Delta x}{2},y\right) u_x\left(x+\frac{\Delta x}{2},y\right) - k\left(x-\frac{\Delta x}{2},y\right) u_x\left(x-\frac{\Delta x}{2},y\right)\right]\Delta y$$

If we equate these two energy terms and take the limit as Δx, Δy, Δt approach zero, we obtain the heat equation

$$\rho c\,\frac{\partial u}{\partial t} = \frac{\partial\left(k\,\frac{\partial u}{\partial x}\right)}{\partial x} + \frac{\partial\left(k\,\frac{\partial u}{\partial x}\right)}{\partial x} = \left(ku_x\right)_x + \left(ku_y\right)_y$$

A somewhat "cleaner" way to derive this formula involves the use of the divergence theorem [Kaplan, 1952, chapter 5]. We will state the theorem in three dimensions although it applies to the plane as well. If $\underline{F}$ is a vector, then

$$\int_V \nabla\cdot\underline{F}\;dv = \int_{\partial V} \underline{F}\cdot\underline{n}\;ds$$

Here $\nabla\cdot\underline{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$ is the divergence of $\underline{F} = F_x\underline{i} + F_y\underline{j} + F_z\underline{k}$ and $\underline{n}$ is the outward-drawn normal to the volume V. The surface integral on the right is taken over the boundary of V, denoted by ∂V. The heat flux through a surface element into V is given by the derivative of u in the direction of the outward normal to the surface multiplied by the heat conductivity, $k\partial u/\partial n$. We have $k\partial u/\partial n = k\underline{\nabla}u\cdot\underline{n}$ where $\underline{\nabla}$ u is the gradient of u ($\nabla u = u_x\underline{i} + u_y\underline{j} + u_z\underline{k}$) and therefore the rate at which heat energy is being conducted into the region V through the boundary ∂V is given by

$$\int_{\partial V} k\underline{\nabla}u\cdot\underline{n}\,ds$$

The rate at which heat energy is changing within the volume V is given by

$$\frac{\partial}{\partial t}\int_V \rho c u\,dv$$

Therefore we obtain (assume ρ and c do not depend on time t)

$$\int_V \rho c\frac{\partial u}{\partial t}dv = \int_{\partial V} \underline{\nabla}u\cdot\underline{n}\,ds = \int_V \underline{\nabla}\cdot(k\underline{\nabla}u)dv \qquad (5.1\text{-}1)$$

Therefore at any point in the region V (assume that all the derivatives are continuous), we must have the heat equation

$$\rho c\frac{\partial u}{\partial t} = \underline{\nabla}\cdot(k\underline{\nabla}u) \qquad \text{or}$$

$$\rho c\frac{\partial u}{\partial t} = \left(ku_x\right)_x + \left(ku_y\right)_y + \left(ku_z\right)_z \qquad (5.1\text{-}2)$$

Cases for which the conductivity is not continuous are of great practical interest [Wachspress, 1965]. Our region may be composed of two different materials which meet along an interface. Along this interface, the conductivity may not be continuous. Along this interface we will require the flux of heat energy to be continuous, that is $k\partial u/\partial n$ is continuous. If k is not continuous, then clearly $\partial u/\partial n$ cannot be continuous. Therefore the derivatives used in the heat equation do not exist along the interface. We must solve the equation in regions where k is continuous and then piece the solutions together by means of the requirement that $k\partial u/\partial n$ be continuous along the interface. We will see that the integral formulation of equation (5.1-1) is better suited for this than the differential equation (5.1-2).

In order to solve the heat equation we must have some boundary conditions. We could specify the temperature on the boundary to obtain $u(x,y,t) = g(x,y,t)$ on ∂V where g is a known function defined for points (x,y) on the boundary ∂V. This is called a boundary condition of the first kind or a Dirichlet boundary condition. We might also specify the heat flux to obtain $\partial u/\partial n = g$ on ∂V. This is called a boundary condition of the second kind, or a Neumann boundary condition. Another common boundary condition is $\alpha u + \beta\partial u/\partial n = g$ where α and β are functions defined on ∂V such that $\alpha \neq 0$ and $\beta \neq 0$. This is a boundary condition of the third kind. We can give a simple physical interpretation of this boundary condition by consideration of the temperature in a rod. This temperature is governed by the one-dimensional heat equation $u_t = \sigma u_{xx}$ where $\sigma = k/\rho c$ is assumed constant. Suppose the rod occupies the interval $0 \leq x \leq 1$. Suppose the end of the rod is in contact with a heat reservoir at $x = 0$. Also, suppose there is a thin film on the end of the rod, perhaps an oxide coating. We assume the film is so thin that it has a constant temperature gradient. We suppose the film extends from $x = -\delta$ to $x = 0$. Then the heat flux through the film is $k_f\left(u(-\delta,t) - u(0,t)\right)\big/(-\delta)$, and this must equal the flux through the end of the rod $ku_x(0,t)$. We assume the temperature at the end of the film is that of the heat reservoir, thus $u(-\delta,t) = g(t)$. If we combine these equations we obtain

$$k_f(g(t) - u(0,t))/(-\delta) = ku_x(0,t) \qquad \text{or}$$

$$u(0,t) - \frac{k\delta}{k_f}\, u_x(0,t) = g(t) \qquad \text{or}$$

$$u(0,t) + \beta\,\frac{\partial u}{\partial n} = g(t) \qquad \beta = \frac{k\delta}{k_f}$$

and $\partial u/\partial n$ denotes the outward normal derivative. All of these boundary conditions yield a properly posed problem.

We will consider only steady state problems in this chapter, although the methods used to derive the steady state equations will apply to the heat equation. By a steady state solution we mean one which is independent of time t. In one dimension our problem thus becomes a boundary value problem for an ordinary differential equation. That is, the problem

$$\sigma\,\frac{\partial u}{\partial t} = \left(ku_x\right)_x \qquad \begin{aligned} u(0,t) &= u_0 \\ u(1,t) &= u_1 \end{aligned}$$

under the assumption $u = u(x)$ becomes

$$\frac{d}{dx}\left(k(x)\,\frac{du}{dx}\right) = 0 \qquad\qquad u(0) = u_0$$

$$u(1) = u_1$$

We will first treat the derivation of difference schemes for one-dimensional problems. The basic principles are the same in higher dimensions but the details are considerably more complex.

5.2 <u>The derivation of difference schemes for ordinary second-order boundary value problems</u>. We will consider the differential equation

$$\frac{d}{dx}\left(c_2(x)\,\frac{du}{dx}\right) + \frac{d}{dx}\left(c_1(x)u\right) + c_0(x)u = f(x) \tag{5.2-1}$$

on the interval $A \le x \le B$ with boundary conditions

$$\begin{aligned} \alpha_A u(A) + \beta_A u'(A) &= g_A \qquad & \alpha_A^2 + \beta_A^2 \ne 0 \\ \alpha_B u(B) + \beta_B u'(B) &= g_B \qquad & \alpha_B^2 + \beta_B^2 \ne 0 \end{aligned} \tag{5.2-2}$$

We will denote the operator on the left side of equation (5.2-1) by L(u) and define the inner product (u,v) by

$$(u,v) = \int_A^B u(x)v(x)dx$$

We say a problem is self-adjoint if (u,Lv) = (Lu,v) for any two functions u and v which have continuous second derivatives and satisfy the boundary conditions (5.2-2) with $g_A = g_B = 0$. We will see that self-adjoint problems lead to a symmetric system of finite difference equations provided an appropriate difference scheme is used.

Problem 5.2-1. Show that the problem defined by equations (5.2-1) with $c_1(x) \equiv 0$ and boundary conditions (5.2-2) is self-adjoint.

5.2.1 Difference schemes based on a Taylor series expansion. Suppose we assume the boundary conditions are $u(0) = g_0$ and $u(1) = g_1$ and write the differential equation in the form

$$c_2(x)u'' + (c_2'(x) + c_1(x))u' + (c_1'(x) + c_0(x))u = f(x)$$

For simplicity we will assume the coefficients are constant, $c_i(x) = c_i$, $0 \le i \le 2$. We will assume that we have an equally spaced mesh $x_j = j/J$, $0 \le j \le J$. We denote $u(x_j)$ by $u_j = u(x_j)$. We can represent the boundary conditions by $u_0 = g_0$, $u_J = g_1$. Then we must obtain a system of equations for u_j, $1 \le j \le J-1$. We will make a further assumption that $c_1 = 0$. We will look for a difference approximation in the form

$$\alpha_{j,1}\, u_{j-1} + \alpha_{j,2} u_j + \alpha_{j,3}\, u_{j+1} \cong c_2 u''(x_j) + c_0 u(x_j) \qquad (5.2\text{-}3)$$

If we let $h = 1/J$ denote the mesh spacing, then we can expand u_{j-1} and u_{j+1} in a Taylor series in h with the derivatives evaluated at x_j. If we substitute these expansions into equation 5.2-3 and require equality for the h^0, h^1, and h^2 terms we obtain three equations which can then be solved for $\alpha_{j,1}$, $\alpha_{j,2}$, and $\alpha_{j,3}$. That is (let $\alpha_k = \alpha_{j,k}$, k=1,2,3)

$$\alpha_1\, u_j + \alpha_2 u_j + \alpha_3 u_j = c_0\, u_j$$

$$-\alpha_1 h u_j' \qquad + \alpha_3 h u_j' = 0$$

$$\alpha_1 \frac{h^2}{2} u_j'' \qquad + \alpha_3 \frac{h^2}{2} u_j'' = c_2\, u_j''$$

The solution is $\alpha_1 = \alpha_3 = c_2/h^2$ and $\alpha_2 = c_0 - 2c_2/h^2$. From the differential equation we have $c_2u'' + c_0u = f(x)$, therefore the finite difference scheme is

$$\frac{c_2}{h^2} U_{j-1} + \left(c_0 - \frac{2c_2}{h^2}\right) U_j + \frac{c_2}{h^2} U_{j+1} = f(x_j) \tag{5.2-4}$$

$$U_0 = g_0, \qquad U_J = g_1$$

If we write out this equation in matrix form, we have AU = F where

$$A = \left|\begin{array}{ccccccc} \alpha_2 & \alpha_3 & 0 \; . \; . \; . & & & & \\ \alpha_1 & \alpha_2 & \alpha_3 & 0 \; . \; . \; . & & & \\ 0 & \alpha_1 & \alpha_2 & \alpha_3 & 0 \; . \; . \; . & & \\ & & & & & & \\ & & . \; . \; . & 0 & \alpha_1 & \alpha_2 & \alpha_3 \\ & & & . \; . \; . \; 0 & & \alpha_1 & \alpha_2 \end{array}\right| \quad F = \left|\begin{array}{c} f_1 - \alpha_1 g_0 \\ f_2 \\ f_3 \\ \cdot \\ \cdot \\ \cdot \\ f_{J-1} - \alpha_3 g_1 \end{array}\right| \quad U = \left|\begin{array}{c} U_1 \\ U_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ U_{J-1} \end{array}\right| \tag{5.2-5}$$

Note that we have a tridiagonal matrix equation to solve. We have already discussed methods to solve such systems in section 3.4. We could have obtained the same formula by simple substitution of the centered difference approximation for u'' (that is $u''(x_j) \cong (u_{j+1} - 2u_j + u_{j-1})/h^2$) into the differential equation. However, there is sometimes an advantage in regarding this as a Taylor series substitution, as demonstrated in the book by Greenspan [1965].

5.2.2 Irreducible, diagonally dominant matrices. Now we will digress for a moment to consider some concepts from matrix theory which are especially relevant to the solution of elliptic equations. For more details on these concepts, we refer the reader to the book by Varga.

Definition 5.2-1. A permutation matrix is a square matrix P such that each row and each column contains a single nonzero element which is equal to one.

If we premultiply A by such a matrix, then PA is a matrix in which the rows of A have been interchanged or permuted. Postmultiplication, AP, permutes the columns.

Problem 5.2-2. Show that if P is a permutation matrix, then $P^T P = I$.

Definition 5.2-2. A square matrix A is reducible if there is a permutation matrix P such that

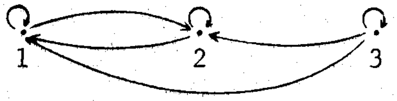
$$PAP^T = \begin{vmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{vmatrix}$$

where $A_{1,1}$ and $A_{2,2}$ are square submatrices. A square matrix is irreducible if it is not reducible.

Problem 5.2-3. Show that the following matrices are reducible.

$$\begin{vmatrix} 3 & 1 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{vmatrix} \qquad \begin{vmatrix} 1 & 0 & 5 & 1 \\ 2 & 4 & 1 & 2 \\ 1 & 0 & 3 & 1 \\ 3 & 0 & 2 & 6 \end{vmatrix}$$

Definition 5.2-3. We will define a graph associated with the n×n matrix (a_{ij}). We choose n distinct points in the plane $P_1, \ldots, P_n$ which we call nodes. A directed path of length 1 is curve joining P_i to P_j. A set of directed paths is called a directed graph. We may also define a directed graph on n nodes in nongeometric terms as a set of order pairs (i_k,j_k) $1 \le k \le m$ where $1 \le i_k \le n$, $1 \le j_k \le n$. Associated with a matrix of order n is a directed graph defined by all pairs (i,j) such that $a_{ij} \neq 0$. The directed graph associated with the first matrix in problem 5.2-3 is



Definition 5.2-4. A directed path of length r joining nodes i and j is a set of pairs in the graph

$$(i_0,i_1),\ (i_1,i_2),\ (i_2,i_3),\ \ldots,\ (i_{r-1},i_r)$$

where $i_0 = i$ and $i_r = j$. A directed graph is strongly connected if there is a directed path between every pair of nodes i and j.

Problem 5.2-4. Show that a matrix is irreducible if and only if it is strongly connected. Hint: Suppose there is no path from node i_1 to node i_2. Let M be the set of nodes which can be connected to i_1. Permute the matrix so that these M nodes correspond to the last M rows of the matrix. This permutation shows that the matrix is reducible.

Problem 5.2-5. Show that if $\alpha_1 \neq 0$, $\alpha_3 \neq 0$, then the matrix A in equations 5.2-5 is irreducible.

Definition 5.2-5. A matrix is diagonally dominant if

$$|a_{i,i}| \ge \sum_{\substack{j \\ j\neq i}} |a_{i,j}| \qquad \text{for all } i$$

with inequality for at least one i.

Problem 5.2-6. Show that a diagonally dominant irreducible matrix is nonsingular. Hint: Suppose $Ax = 0$ and let $|x_k| = \max_j |x_j|$. Then $|a_{kk}| \le \sum_{j \ne k} |a_{kj}| \frac{|x_j|}{|x_k|}$. In case $|x_j| < |x_k|$ and $a_{kj} \ne 0$, this leads to a contradiction. Now consider the remaining cases.

If the coefficients in equation 5.3-4 satisfy $c_2 > 0$ and $c_0 \le 0$, then the matrix A in equation 5.3 is irreducibly diagonally dominant and thus nonsingular. Irreducible diagonal dominance frequently holds for matrices derived from elliptic problems.

5.2.3 <u>Derivation of the difference scheme using integration by parts</u>. Suppose we wish to approximate the self-adjoint problem

$$\frac{d}{dx}\left(c_2(x)\,\frac{du}{dx}\right) + c_0(x)u = f \qquad \begin{aligned} u(0) &= g_0 \\ u(1) &= g_1 \end{aligned} \tag{5.2-6}$$

on the mesh $0 = x_0 < x_1 < \ldots < x_J = 1$. If we integrate this equation from $x_{j-\frac{1}{2}} = (x_{j-1} + x_j)/2$ to x_j we obtain

$$c_2(x)u'(x)\Big|_{x_{j-\frac{1}{2}}}^{x_j} + \int_{x_{j-\frac{1}{2}}}^{x_j} c_0(x)u(x)dx = \int_{x_{j-\frac{1}{2}}}^{x_j} f(x)dx \tag{5.2-7}$$

If $c_2(x)$ is piecewise continuous with possible discontinuities at the interior mesh points x_j, then the above differential equation does not hold at the mesh points. However, the integration which produced equation 5.2-7 is still valid since we did not integrate across a mesh point. To derive the difference scheme we use the condition that $c_2(x)u'(x)$ must be continuous (we make this assumption because it is true for the heat equation).

If we form a second equation similar to 5.2-7 by integration from x_j to $x_{j+\frac{1}{2}}$ and add the two equations using the continuity of $c_2(x)u'(x)$ we obtain

$$c_2(x_{j+\frac{1}{2}})u'(x_{j+\frac{1}{2}}) - c_2(x_{j-\frac{1}{2}})u'(x_{j-\frac{1}{2}}) + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} c_0(x)u(x)dx = F(x_{j+\frac{1}{2}}) - F(x_{j-\frac{1}{2}})$$

where $F(x) = \int_{x_0}^{x} f(\tau)d\tau$. If we approximate the integral on the left side of the equation by $\left[C(x_{j+\frac{1}{2}}) - C(x_{j-\frac{1}{2}})\right]u(x_j)$ where $C(x) = \int_{x_0}^{x} c_0(\tau)d\tau$ and use an obvious difference approximation for $u'(x_{j+\frac{1}{2}})$ we obtain

$$c_2(x_{j+\frac{1}{2}}) \frac{(U_{j+1}-U_j)}{(x_{j+1}-x_j)} - c_2(x_{j-\frac{1}{2}}) \frac{(U_j-U_{j-1})}{x_j-x_{j-1}} + (C_{j+\frac{1}{2}}-C_{j-\frac{1}{2}})U_j = F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}}$$

(5.2-8)

This equation holds for $1 \le j \le J-1$. If we use $u_0 = g_0$ and $u_J = g_1$, we can solve this system of equations.

Problem 5.2-7. If $c_2(x) > 0$ for $0 \le x \le 1$ and $c_0(x) \le 0$ for $0 \le x \le 1$, then show that the system of equations (5.2-8) for $(U_1,\ldots,U_{J-1})$ has a unique solution, that is the matrix is nonsingular. Show that this matrix is symmetric.

Problem 5.2-8. Suppose instead of Dirichlet boundary conditions in equation 5.2-6 we have Neumann boundary conditions $u'(0) = g_0$, $u'(1) = g_1$. Modify the above integral method to obtain a finite difference scheme for this problem.

5.2.4 <u>A difference scheme based on a variational formulation</u>. We will modify our notation slightly and ask for a solution of the equation

$$-\left(c_2(x)u'(x)\right)' + c_0(x)u(x) = f(x) \qquad 0 \le x \le 1 \qquad (5.2\text{-}9)$$

where $u(0) = g_0$, $u(1) = g_1$. We suppose $f(x)$, $c_2(x)$ and $c_0(x)$ are continuous, with $c_2(x) > 0$ and $c_0(x) \ge 0$. The variational problem is to find a minimum of the functional

$$F(w) = \int_0^1 \left[c_2(x)(w'(x))^2 + c_0(x)(w(x))^2 - 2w(x)f(x)\right]dx \qquad (5.2\text{-}10)$$

We ask for the minimum over the set of all functions $w \in C^2[0,1]$ which satisfy the boundary conditions $w(0) = g_0$, $w(1) = g_1$. ($w \in C^2[0,1]$ if $w(x)$ has a continuous second derivative on the interval $0 \le x \le 1$.) We call such functions $w(x)$ admissible. We will now show that if such a minimum exists and is given by $F(u)$, than u must satisfy the differential equation (5.2-9). Let $u(x)$ be the minimizing function. If $w(x) = u(x) + \epsilon v(x)$ is an admissible function, then $F(u) \le F(w)$ by assumption. Note that $w(x)$ will be admissible for any real ϵ, if u is admissible, $v \in C^2[0,1]$, and $v(0) = v(1) = 0$.

Problem 5.2-9. Show that the admissible minimizing function $u(x)$ satisfies the differential equation. Hint: Show that

$$F(u+\epsilon v) = F(u) + 2\epsilon G(u,v) + \epsilon^2 F(v) \qquad \text{where}$$

$$G(u,v) = \int_0^1 \left[-(c_2(x)u'(x))' + c_0(x)u(x) - f(x)\right]v(x)dx \qquad (5.2\text{-}11)$$

Show that if $F(u) \leq F(u+\epsilon v)$, then $G(u,v) = 0$ for all v which satisfy the conditions above. Now use the continuity of u(x) to show that the bracketed term in equation (5.2-11) must vanish--that is, u must satisfy the differential equation.

Problem 5.2-10. Show that if u(x) is admissible and satisfies the differential equation (5.2-9), then u minimizes the quadratic form in equation (5.2-10).

Now we are ready to construct the finite difference scheme. Suppose we use the mesh $0 = x_0 < x_1 < \dots < x_J = 1$. We use the integral approximations shown below.

$$\int_{x_j}^{x_{j+1}} c_2(x)(w'(x))^2 dx \cong c_2(x_{j+\frac{1}{2}}) \frac{(w_{j+1} - w_j)^2}{(x_{j+1} - x_j)} \qquad x_{j+\frac{1}{2}} = \frac{x_{j+1} + x_j}{2}$$

$$\int_{x_j}^{x_{j+1}} c_0(x)(w(x))^2 dx \cong c_0(x_{j+\frac{1}{2}}) \left(\frac{w_{j+1} + w_j}{2}\right)^2 (x_{j+1} - x_j)$$

$$\int_{x_j}^{x_{j+1}} w(x) f(x) dx \cong \left(\frac{w(x_{j+1}) + w(x_j)}{2}\right) f(x_{j+\frac{1}{2}})(x_{j+1} - x_j)$$

If we now substitute these approximations into the expression for F(w), we obtain an expression which is quadratic in the components of the vector $W = (W_1, W_2, \dots, W_{J-1})^T$ (note that W_j approximates $w(x_j)$). We can write the approximation for F(w) in matrix form as

$$F(w) \cong W^T A W - 2W^T b + d$$

where $b = (b_1, \dots, b_{J-1})$ is a vector and d a scaler.

Problem 5.2-11. Assume an equally spaced mesh $x_j = j/J$. Write out an expression for A, b and d in terms of c_2, c_0, and f. Note that A is a symmetric tridiagonal matrix.

Now in order to find the vector W which will minimize the above approximation for F(w) we differentiate the expression with respect to the components W_j and set these derivatives to zero. We then obtain

$$AW = b$$

This is our finite difference scheme.

5.2.5 The effect of a discontinuous coefficient $c(x)$ in $(cu')' = f$. First we will consider an improper method for the solution of this problem. Suppose we write the equation as

$$c(x)u''(x) + c'(x)u'(x) = f(x) \tag{5.2-12}$$

Suppose we have boundary conditions $u(0) = 0$, $u(1) = 1$. If we simply ignore the discontinuity in $c(x)$, we might construct a difference scheme as follows:

$$x_j = jh\ , \qquad 0 \le j \le J\ , \qquad h = 1/J$$

$$U_0 = 0\ , \qquad U_J = 1$$

$$c(x_j)\,\frac{(U_{j+1} - 2U_j + U_{j-1})}{h^2} + c'(x_j)\,\frac{U_{j+1} - U_{j-1}}{2h} = f(x_j) \tag{5.2-13}$$

$$1 \le j \le J-1$$

Now if c(x) is as given below and J is odd, then the discontinuity in c(x) will never fall on a mesh point. Therefore we might naively expect the above scheme to give a reasonable approximation. It does not. If

$$c(x) = \begin{Bmatrix} 1 & 0 \le x < \frac{1}{2} \\ 2 & \frac{1}{2} \le x \le 1 \end{Bmatrix}, \qquad f(x) \equiv 0 , \tag{5.2-14}$$

then a solution of equation (5.2-12) based on the continuity of cu' is

$$u(x) = \begin{cases} \frac{4x}{3} & 0 \le x \le \frac{1}{2} \\ \frac{1}{3} + \frac{2x}{3} & \frac{1}{2} \le x \le 1 \end{cases}$$

Problem 5.2-12. Show that the solution of the system of equations (5.2-13) is $u_j = x_j$, independent of the mesh spacing h.

Obviously the difference scheme does not converge. A difference scheme based on the integral method of section 5.2.2 is

$$U_0 = 0 , \qquad U_J = 1$$

$$c(x_{j+\frac{1}{2}})\left(\frac{U_{j+1} - U_j}{h}\right) - c(x_{j-\frac{1}{2}})\left(\frac{U_j - U_{j-1}}{h}\right) = hf(x_j) \text{ for } 1 \le j \le J-1 \tag{5.2-15}$$

Here we want the discontinuity in c to occur at a mesh point. If the discontinuity is at $x = \frac{1}{2}$ we should take J to be even since $h = 1/J$.

Problem 5.2-13. With c(x) given by equations (5.2-14) show that the solution of differential equation (5.2-12) is also a solution of the difference equations (5.2-15). Thus the difference scheme gives us an exact result in this case.

Next we will give the results of a numerical experiment with the scheme of equation (5.2-15). We defined u(x) and c(x) by

$$u(x) = \begin{cases} e^{(x-\frac{1}{4})} & 0 \le x \le \frac{1}{2} \\ e^{\frac{x}{2}} & \frac{1}{2} \le x \le 1 \end{cases}$$

$$c(x) = \begin{cases} 1 & 0 \le x \le \frac{1}{2} \\ 2 & \frac{1}{2} \le x \le 1 \end{cases}$$

Then the right side f(x) is

$$f(x) = \begin{cases} e^{x-\frac{1}{4}} & 0 \le x \le \frac{1}{2} \\ \frac{1}{2}e^{\frac{x}{2}} & \frac{1}{2} \le x \le 1 \end{cases}$$

The boundary conditions are $U_0 = e^{-\frac{1}{4}}$, $U_J = e^{\frac{1}{2}}$. Using the above definitions of f(x) and c(x) we can compute U_j from equation (5.2-15). Then we determine the maximum relative error $E = \max_j |U_j - u(x_j)| / \max_j |u(x_j)|$. This error is listed in the table below.

J	Error	J	Error
10	3.3(-3)	11	8.2(-2)
20	1.6(-3)	21	8.0(-2)
40	8.1(-4)	41	3.1(-3)
80	4.1(-4)	81	7.9(-2)

In the case where J is even the discontinuity in c(x) occurs at a mesh point, as it must if the derivation of the difference scheme is to be valid. In this case the error seems to be proportional to Δx. Apparently we have a first-order scheme. If we take J to be odd we apparently do not have a convergent scheme. We must always be careful with limited experimentation such as this. Although we can sometimes gain considerable insight through such experiments, they do not prove anything. This is particularly true if we only run one case such as the above.

Problem 5.2-14. Determine the truncation error for the above example. Suppose the scheme of equation (5.2-8) was used instead of equation (5.2-15). Would there be any difference in the truncation error? Program scheme (5.2-8) and run the above example to compare the scheme of equation (5.2-8) with that of (5.2-15).

5.2.5 An example to illustrate the treatment of Neumann boundary conditions. In this section we will show that the difference approximation used for the Neumann boundary conditions must have the same accuracy as that used in the interior. The overall order of accuracy is the minimum of that for the boundary and the interior. Perhaps this is no surprise but we think it is worth an example. We will consider the equation

$$(cu')' = f \tag{5.2-16}$$

with Neumann boundary conditions $u'(0) = g_0$, $u'(1) = g_1$.

The first scheme is derived from an integral method. We use the mesh $x_j = (j-\frac{1}{2})h$, $1 \le j \le J$, $h = 1/J$. The difference equation for $j = 1$ is then

$$c(x_{3/2})u'(x_{3/2}) - c(x_{1/2})u'(x_{1/2}) = hf(x_1)$$

The difference approximation is

$$c(x_{3/2})\left(\frac{U_2 - U_1}{h}\right) - c(x_{1/2})g_0 = hf(x_1)$$

Note that $x_{3/2} = h$, $x_{1/2} = 0$, $x_1 = h/2$. In the interior, the difference approximation is

$$c(x_{j+\frac{1}{2}})\left(\frac{U_{j+1} - U_j}{h}\right) - c(x_{j-\frac{1}{2}})\left(\frac{U_j - U_{j-1}}{h}\right) = hf(x_j) \qquad (5.2\text{-}17)$$

At the upper boundary

$$c(x_{J+\frac{1}{2}})g_1 - c(x_{J-\frac{1}{2}})\left(\frac{U_J - U_{J-1}}{h}\right) = hf(x_J)$$

The reader should note that if we divide these equations by h, then substitute for U_j, the solution of the differential equation $u(x_j)$, the error will be $O(h^2)$. We have a scheme of second-order accuracy.

In the second case we use a slightly different mesh, $x_j = jh$ where $h = 1/J$. For the unknowns U_j, $1 \le j \le J-1$, we have the equation (5.2-17) except $x_j = j/J$ in this case and $x_j = (j-\frac{1}{2})/J$ in the first case. For the unknowns U_0 and U_J we use the boundary conditions

$$\frac{U_1 - U_0}{h} = g_0 \qquad\qquad \frac{U_J - U_{J-1}}{h} = g_1$$

Note that these boundary equations have only first-order accuracy.

We could obtain a third scheme in case $c(x)$ and $c'(x)$ are continuous. This scheme will have second-order accuracy. We use the mesh $x_j = (j-1)h$,

$0 \leq j \leq J$, $h = 1/(J-2)$. Note that $x_0 = -h$ and $x_J = 1+h$. Thus these points are outside the interval $0 \leq x \leq 1$ on which the differential equation is defined. Nevertheless we will define mesh variables U_0 and U_J on these "fictitious" mesh points. We use the differential equation in the form

$$c(x)u''(x) + c'(x)u'(x) = f(x)$$

The difference scheme for interior points is

$$c(x_j)\left(\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2}\right) + c'(x_j)\,\frac{U_{j+1} - U_{j-1}}{2h} = f(x_j) \qquad 1 \leq j \leq J-1$$

To complete the set of equations we use the boundary conditions

$$\frac{U_2 - U_0}{2h} = g_0 \qquad\qquad \frac{U_J - U_{J-2}}{2h} = g_1$$

Note that these equations are of second-order accuracy. Also note that these difference schemes for the Neumann boundary condition do not have a unique solution. If the vector $\{U_j\}$ is a solution, then so is $\{U_j+K\}$ where K is an arbitrary constant. This is proper since the differential equation (5.2-16) with Neumann boundary conditions has the same property. This means that $\{U_j\}$ is the solution of a singular system of linear equations. Therefore we have to modify the usual Gauss elimination in order to obtain $\{U_j\}$. We assume that all the pivots (the diagonal elements in the upper triangular reduction of our system) will be nonzero except the last (or bottom-most). We know the exact solution in these test cases. Thus we simply set the last component U_J equal to the exact value $u(x_J)$. Then we use the backward substitution to obtain the remaining components.

In the third case we have Dirichlet boundary conditions for equation (5.2-16). We use the following difference scheme which is based on the integral method.

$$c(x_{j+\frac{1}{2}})\left(\frac{U_{j+1} - U_j}{h}\right) - c(x_{j-\frac{1}{2}})\left(\frac{U_j - U_{j-1}}{h}\right) = hf(x_j) \qquad 1 \le j \le J-1 \tag{5.2-18}$$

where $U_0 = g_0$, $U_1 = g_1$, $h = 1/J$.

These experiments were run on a Control Data 6600 computer. In the three cases below we have $c(x) \equiv 2$, $u(x) = e^{x/2}$. Then $f(x) = \frac{1}{2}e^{x/2}$. In the Dirichlet case the boundary conditions are $g_0 = 1$, $g_1 = e^{\frac{1}{2}}$; in the Neumann case $g_0 = \frac{1}{2}$, $g_1 = \frac{1}{2}e^{\frac{1}{2}}$. The error is the maximum relative error, namely $\epsilon = \max_j |u(x_j) - U_j| \,/\, \max_j |u(x_j)|$. The results are listed in the table below.

J	Error I. Second-order Neumann	 II. First-order Neumann	 III. Dirichlet
10	4.6(-5)	7.7(-3)	5.4(-6)
20	1.2(-6)	3.8(-3)	1.3(-6)
40	3.1(-6)	1.9(-3)	3.2(-7)
80	7.8(-7)	9.5(-4)	8.0(-8)

Note that the error for cases I and III is reduced by a factor of 4 when the mesh spacing is halved. In cases II the error seems to be proportional to Δx rather than Δx^2. Error estimates for Laplace's equation on a rectangle using Neumann boundary conditions are discussed in a paper by Giese [1958]. Numerical solutions of Laplace's equation for Neumann boundary conditions show this same sensitivity to accuracy at the boundary.

5.2.7 <u>A convergence proof by the method of Gerschgorin</u>. We wish to prove that the solution of the finite difference scheme for equation (5.2-1) converges. The method of Gerschgorin which we will use is described in considerable detail in the book by Forsythe and Wasow [1960, pp. 283-328]. We will illustrate the method by applying it to the simple equation

$$(c_2 u')' = 0 \qquad u(0) = g_0 \qquad c_2(x) > 0 \qquad (5.2\text{-}19)$$
$$u(1) = g_1$$

The same method can be applied to certain difference schemes for Laplace's equation. We will first consider the trivial problem

$$u'' = 0 \qquad u(0) = g_0$$
$$u(1) = g_1$$

Problem 5.2-15. Show that the solution of equation (5.2-19) satisfies the following maximum principle. The maximum (and minimum) of $u(x)$ is taken on at the boundary. That is, $\min[g_0, g_1] \le u(x) \le \max[g_0, g_1] \quad 0 \le x \le 1$. Does this maximum principle apply to the equation $-(c_2 u')' + c_0 u = 0$ where $c_2(x) > 0$, $c_0(x) \ge 0$?

Now we will demonstrate that the same sort of maximum principle applies to the finite difference approximation for equation (5.2-19). Use the difference scheme of equation (5.2-18), namely

$$c_2(x_{j+\frac{1}{2}})\,(U_{j+1} - U_j) - c_2(x_{j-\frac{1}{2}})\,(U_j - U_{j-1}) = 0 \qquad 1 \le j \le J-1$$

$$U_0 = g_0\,, \qquad U_J = g_1.$$

Problem 5.2-16. Show that if $g_0 \neq g_1$, $\min[g_0,g_1] < U_j < \max[g_0,g_1]$ $1 \leq j \leq J-1$.

Next we will introduce some notation, even though it is not needed in this trivial case. The same notation will be used for the treatment of Laplace's equation in two dimensions. We denote the interval $0 \leq x \leq 1$ by R; its boundary, $x = 0$ and $x = 1$, by ∂R; the mesh points by $R_h = \{x_j, 0 \leq j \leq J\}$; and the mesh boundary by $\partial R_h = \{x_0, x_J\}$. Let L_h be the finite difference operator defined by

$$L_h(U) = c_2(x_{j+\frac{1}{2}})\ (U_{j+1} - U_j) - c_2(x_{j-\frac{1}{2}})\ (U_j - U_{j-1}) \qquad 1 \leq j \leq J-1$$

We will need the following result.

Problem 5.2-17. Let $V = \{V_j\}$ be a mesh function such that $L_h(V) \geq 0$ for some subset R^* of $(R_h - \partial R_h)$. Note that L_h is not defined on the boundary ∂R_h. Then

$$\max_{x_j \in R^*} V_j \leq \max_{x_j \in R_h - R^*} V_j$$

We need to consider the truncation error of this difference scheme before we prove convergence.

Problem 5.2-18. Suppose u is a solution of equation (5.2-19) with enough continuous derivatives, and assume $c_2(x)$ is also sufficiently differentiable. Then show that $L_h(u) = 0(h^4)$, that is $|L_h(u)| \leq h^4 M$.

Gerschgorin's method requires the construction of a comparison mesh function W such that $L_h(W) \leq -1$ on R^* and $W \geq 0$ on $R_h - R^*$ where $R_h - R^*$ contains the boundary ∂R_h. If we let the solution of the difference scheme be U, and the solution of the differential equation be u, then the error function is $E = u-U$; and if τ is a bound for the truncation error, we have $|L_h(E)| \le \tau$. Now suppose $R_h^* = R_h - \partial R_h$. If we define the mesh functions $\varphi_\pm$ by $\varphi_\pm = \pm E - \tau W$, then $L_h(\varphi_\pm) \ge 0$ on $R_h - \partial R_h$, and therefore $\max_{R_h-\partial R_h} \varphi_\pm \le \max_{\partial R_h} \varphi_\pm$. But $u = U$ on ∂R_h, and thus $\varphi_\pm = -\tau W \le 0$ on ∂R_h. Therefore $\varphi_\pm \le 0$ on $R_h - \partial R_h$. This in turn implies $\pm E \le \tau W$ or $|u-U| \le \tau|W|$ on $R_h - \partial R_h$. This is our error estimate. Obviously, this error estimate depends on the choice of the comparison function.

We will choose a comparison function for the case $c_2(x) \equiv 1$, which leads to the trivial problem $u'' = 0$, $u(0) = g_0$, $u(1) = g_1$ with solution $g_0 + x(g_1-g_0)$. If we let $W_j = x_j\frac{(1-x_j)}{2}$, then $W_j \ge 0$ on ∂R_h. The operator L_h is given by $L_h(W) = W_{j+1} - 2W_j + W_{j-1}$, and for the W given above $L_hW = -h^2$. Therefore we can use

$$W_j = \frac{x_j(1 - x_j)}{2h^2}$$

as the comparison function. A bound for the truncation error is $\tau = h^4M$ according to problem 5.2-16. Therefore our error estimate is

$$|u_j - U_j| \le \tau W_j = \frac{Mh^4x_j(1-x_j)}{2h^2} \le \frac{Mh^2}{8}.$$

Problem 5.2-19. Return to the difference scheme of equation (5.2-18) where $c_2(x) > 0$. Assume the mesh is not necessarily equally spaced, but $0 = x_0 < x_1 < \ldots < x_J = 1$. This requires a slight modification in the difference scheme. Construct a comparison function for this case and obtain an error estimate. Hint: Try something of the form

$$\Sigma\, G_{j-\frac{1}{2}}(x_j - x_{j-1}) \,, \qquad G_{j+\frac{1}{2}} = \frac{x_{j+\frac{1}{2}}}{c_{j+\frac{1}{2}}} \,.$$

Varga [1962, pp. 165] has a more general proof of convergence for this problem based on the properties of a Stieltjes matrix. A Stieltjes matrix is a symmetric positive definite matrix with nonpositive off-diagonal elements; $a_{ij} \le 0$ if $i \neq j$. If a Stieltjes matrix A is irreducible, then $A^{-1} > 0$. Note that by $B > 0$ we mean $b_{ij} > 0$ for all i and j. This property can be used to obtain an error estimate for the equation $-(c_2 u')' + c_0 u = 0$ with $c_2 > 0$, $c_1 \ge 0$. Since this equation does not satisfy the maximum principle, we would not expect Gerschgorin's method to work at least as we have stated it.

5.3 A finite difference approximation for Laplace's equation on a rectangle. In this section we will be concerned with the following problem:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \rho \qquad u = u(x,y) \qquad \rho = \rho(x,y) \qquad (5.3\text{-}1)$$

$$0 \le x \le a \qquad 0 \le y \le b$$

$$u(x,0) = f_1(x) \qquad u(0,y) = f_3(y)$$

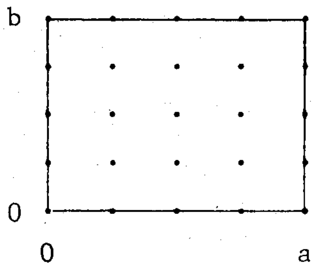
$$u(x,b) = f_2(x) \qquad u(a,y) = f_4(y)$$

We will assume that the data for this problem; that is, the functions ρ and f_i; are such that the solution is as smooth as required for our purposes. First we will construct the difference approximation. We define the mesh (x_j, y_k) as follows (J and K are positive integers):

$$x_j = j\Delta x, \qquad 0 \le j \le J+1, \qquad \Delta x = a/(J+1)$$

$$y_k = k\Delta y, \qquad 0 \le k \le K+1, \qquad \Delta y = b/(K+1)$$

The mesh then appears as follows for the case J = K = 3.



Using the usual centered approximation for the second derivatives, we obtain a set of equations for the values of u(x,y) at the mesh points. We let $u_{j,k}$ denote $u(x_j, y_k)$. Then we have an equation for each interior point. For an interior point we must have $1 \le j \le J$ and $1 \le k \le K$. Note that the values of u_{jk} are known at the boundary points $j = 0$, $j = J+1$, $k = 0$, and $k = K+1$. Therefore we have a linear system of J*K equations in J*K unknowns. The equations are

$$\frac{U_{j+1,k} - 2U_{j,k} + U_{j-1,k}}{\Delta x^2} + \frac{U_{j,k+1} - 2U_{j,k} + U_{j,k-1}}{\Delta y^2} = \rho_{jk} \qquad (5.3\text{-}2)$$

$$1 \le j \le J$$

$$1 \le k \le K$$

The values of $U_{0,k}$, $U_{J+1,k}$, $U_{j,0}$, and $U_{j,K+1}$ are given by the boundary conditions. We may write these equations in the form

$$U_{j,k} - \theta_x(U_{j-1,k} + U_{j+1,k}) - \theta_y(U_{j,k-1} + U_{j,k+1}) = -\delta^2\rho_{j,k} \qquad (5.3\text{-}3)$$

$$1 \le j \le J$$
$$1 \le k \le K$$

where $\theta_x = \dfrac{\Delta y^2}{2(\Delta x^2 + \Delta y^2)}$, $\theta_y = \dfrac{\Delta x^2}{2(\Delta x^2 + \Delta y^2)}$, $\delta^2 = \dfrac{\Delta x^2\,\Delta y^2}{2(\Delta x^2 + \Delta y^2)}$.

We will define a matrix operator based on a multiplication of these equations by -1, namely

$$L_h(U) = \theta_x(U_{j-1,k} + U_{j+1,k}) + \theta_y(U_{j,k-1} + U_{j,k+1}) - U_{j,k} \qquad (5.3\text{-}4)$$

Note that L_h is an $m \times n$ matrix where $m = J*K$, $n = (J+2)*(K+2)$. We use the subscript h to denote the fact that L_h depends on the mesh.

Problem 5.3-1. Let J = K = 3. Let the vector U be given by $U^T = (U_{1,1}, U_{2,1}, U_{3,1}, U_{1,2}, U_{2,2}, U_{3,2}, U_{1,3}, U_{2,3}, U_{3,3})$. The above difference scheme can then be written in matrix form as AU = f where A is a matrix of order 9 and f is a vector. Write out the matrix A and the vector f.

Problem 5.3-2. With J = K = 3 write out the (9×25) matrix operator L_h. Hint: Show that the matrix has the block form

$$L_h = (L_{i,j}) \qquad 1 \le i \le 3$$
$$1 \le j \le 5$$

where the $L_{i,j}$ are 3×5 matrices. Then describe the structure of these blocks.

5.3.1 The convergence of the finite difference scheme. In this section we will use the method of Gerschgorin to prove convergence. The technique is the same as that described in the preceding section for the ordinary differential equation $u'' = 0$.

Problem 5.3-3. Let $u(x,y)$ be a sufficiently differentiable solution of equation (5.3-1). Define the truncation error by $L_h(u) - \delta^2\rho = \tau$ where L_h and δ^2 are defined by equation (5.3-3). Show that

$$|\tau| \le \frac{\delta^2}{12}\left[\Delta x^2 \, M_x^{(4)} + \Delta y^2 \, M_y^{(4)}\right]$$

where $M_x^{(4)}$ and $M_y^{(4)}$ are bounds on the fourth derivatives of u with respect to x and y.

Let R_h denote the set of mesh points (x_j, y_k) $0 \le j \le J+1$, $0 \le k \le K+1$ and let ∂R_h denote the boundary points $j = 0$, $j = J+1$, $k = 0$, and $k = K+1$.

Problem 5.3-4. Let $W = \{W_{jk}\}$ be a mesh function defined on R_h. Suppose $L_h(W) \ge 0$ on a subset R_h^* of R_h. Assume that R_h^* does not contain any points of ∂R_h, that is $R_h^* \subset R_h - \partial R_h$. Then show that

$$\max_{x_{jk} \in R_h^*} W_{jk} \le \max_{x_{jk} \notin R_h^*} W_{jk}$$

Problem 5.3-5. Show that the function $W_{jk} = \frac{a^2}{4} - \left(x_j - \frac{a}{2}\right)^2$ is a suitable Gerschgorin comparison function. That is $W_{jk} \ge 0$ for boundary points $(x_{jk} \in \partial R_h)$. And also $L_h(W) \le -1$.

Problem 5.3-6. Let τ be the bound on the truncation error from problem 5.3-3. Let u be the solution of the differential equation, U the solution of the difference equation, and let $E = u-U$. Then define the functions $\varphi_\pm = \pm E - \tau W$. Continue as in section 5.2 to obtain an estimate on the error

$$\max_{R_h - \partial R_h} |u_{jk} - U_{jk}|$$

5.3.2 Properties of the matrix equation. In this section we will consider the matrix equation AU = f defined by equations (5.3-3). Here the vector U consists of the unknown values $\{U_{jk}\}$ for $1 \le j \le J$, $1 \le k \le K$. Note that the known boundary values are not included in the vector U. When we operate with the matrix L_h, then these known boundary values are included in the vector U. In this case $U = \{U_{jk} \mid 0 \le j \le J+1,\ 0 \le k \le K+1\}$ and therefore the matrix L_h is not a square matrix. To define the matrix A (or L_h) we must specify the order in which we write the components of the vector U. We will order U by running down the rows first. That is, $U^T = (U_{11}, U_{21}, \ldots, U_{J,1}, U_{1,2}, \ldots, U_{1,K}, \ldots, U_{J,K})$. The reader should verify the following block tridiagonal form for the matrix A.

$$\begin{array}{cc}
 & \begin{array}{cccccccc} 1 & 2 & 3 & 4 & \cdots & & & K \end{array} \\
A = & \left|\begin{array}{cccccccc}
D & B & 0 & 0 & \cdots & & & \\
B & D & B & 0 & \cdots & & & \\
0 & B & D & B & \cdots & & & \\
\cdot & & & & & & & \\
\cdot & & & & & & & \\
\cdot & & & & 0 & B & D & B \\
0 & \cdots & \cdots & & 0 & 0 & B & D
\end{array}\right|
\end{array}$$

There are K blocks each of which is a square matrix of order J. Each of these blocks represents a row in the mesh. Equation (5.3-3) links each mesh point to its four nearest neighbors. Thus each row is linked to the row immediately above and the row immediately below. Hence the block tridiagonal form of the matrix A. The matrices B are diagonal matrices and D tridiagonal matrices both of order J.

$$B = \begin{vmatrix} -\theta_y & 0 & 0 & \cdots & & 0 \\ 0 & -\theta_y & 0 & \cdots & & 0 \\ \cdot & & & & & \\ \cdot & & & 0 & -\theta_y & 0 \\ \cdot & & & & & \\ 0 & \cdots & & 0 & 0 & -\theta_y \end{vmatrix} \qquad D = \begin{vmatrix} 1 & -\theta_x & 0 & 0 & \cdots & & \\ -\theta_x & 1 & -\theta_x & 0 & \cdots & & \\ 0 & -\theta_x & 1 & -\theta_x & \cdots & & \\ \cdot & & & & & & \\ \cdot & & \cdots & 0 & -\theta_x & 1 & -\theta_x \\ \cdot & & & & & & \\ 0 & & & 0 & 0 & -\theta_x & 1 \end{vmatrix}$$

Problem 5.3-7. Consider a different ordering of the vector U, namely run through the columns first instead of the rows. Thus $U^T = (U_{1,1}, U_{1,2}, \ldots, U_{1,K}, U_{2,1}, \ldots, U_{2,K}, \ldots, U_{J,1}, \ldots, U_{J,K})$. Show that the matrix A has a block tridiagonal form for this ordering and also determine the submatrices.

Problem 5.3-8. Order the vector U by running along the diagonals of the mesh; that is, group the components $U_{j,k}$ for which j+k is constant. Then $U^T = (U_{1,1}, U_{2,1}, U_{1,2}, U_{3,1}, U_{2,2}, U_{3,1}, U_{4,1}, \ldots, U_{J,K})$. Show that A has block tridiagonal form and determine the form of the submatrices. Note that the blocks are not all of the same order, and the off-diagonal blocks are not square.

Later it will be convenient to write the matrix A as

$$A = I - \theta_x\left(C_x + C_x^T\right) - \theta_y\left(C_y + C_y^T\right)$$

where C_x and C_y are defined by

$$C_x = \begin{array}{llllll} 1 & 2 & \cdots & & & J{*}K \\ \hline 0 & 0 & \cdots & & & \\ 1 & 0 & \cdots & & & \\ 0 & 1 & 0 & \cdots & & \\ \vdots & & & & & \\ 0 & \cdots & & 1 & 0 & 0 \\ 0 & \cdots & & 0 & 1 & 0 \end{array} \qquad C_y = \begin{array}{llllll} 1 & 2 & 3 & \cdots & & K \\ \hline 0 & 0 & 0 & \cdots & & \\ I_J & 0 & 0 & \cdots & & \\ 0 & I_J & 0 & \cdots & & \\ 0 & 0 & I_J & \cdots & & \\ \vdots & & & & & \\ 0 & \cdots & & 0 & I_J & 0 \end{array}$$

where I_J is the identity matrix of order J and θ_x, θ_y are defined in equation (5.3-3). That is

$$C_{x_{r,s}} = \begin{cases} 1 & r = s+1 \\ 0 & \text{otherwise} \end{cases} \qquad C_{y_{r,s}} = \begin{cases} 1 & r = s+J \\ 0 & \text{otherwise} \end{cases}$$

where $1 \le r \le J{*}K$, $1 \le s \le J{*}K$.

Problem 5.3-9. Show that the matrix A is symmetric, irreducible, and diagonally dominant.

Note that $a_{ii} = 1$ for $1 \le i \le J{*}K$ and $a_{ij} \le 0$ if $i \ne j$. We will next determine the eigenvalues of A. This will show that A is positive definite.

Problem 5.3-10. Verify that the eigenvectors of A are given by $W^{(p,q)}$ $1 \le p \le J$, $1 \le q \le K$, where

$$W^{(p,q)}_{j,k} = \sin\left(\frac{jp\pi}{J+1}\right) \sin\left(\frac{kq\pi}{K+1}\right) \qquad 1 \le j \le J, \quad 1 \le k \le K$$

Also show that the eigenvalues of A are given by

$$\lambda^{(p,q)} = 4\theta_x \sin^2 \frac{\pi p}{2(J+1)} + 4\theta_y \sin^2 \frac{\pi q}{2(K+1)}$$

Note that the eigenvalues of A lie in the interval $0 < \lambda^{(p,q)} < 2$. Therefore A is positive definite.

5.4 Difference approximations for Laplace's equation in two dimensions.

In this section we will be concerned with the construction of finite difference approximations for Laplace's equation in a planar region. In the preceding section we were concerned with a rectangular region. Now we wish to consider more general regions. We will first use an integral method which is based on the Green's theorem relating line and surface integrals. Here we will generally follow the presentation given in a paper by Spanier [1967]. This is also described in chapter 6 of the book by Varga [1962].

5.4.1 A scheme based on the integral method.

We assume we have the elliptic equation

$$-\frac{\partial}{\partial x}\left(D\,\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(D\,\frac{\partial u}{\partial y}\right) + Eu = f \tag{5.4-1}$$

$$E = E(x,y) \geq 0$$

$$D = D(x,y) > 0$$

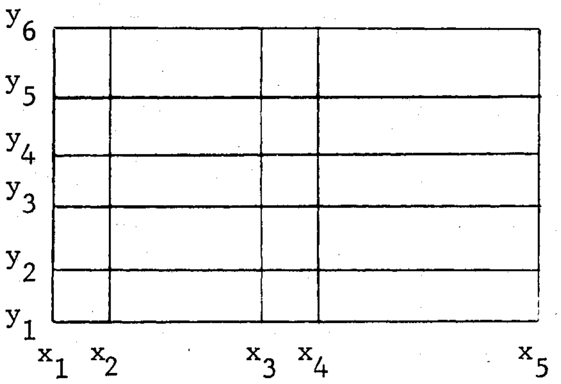
$$f = f(x,y)$$

defined on some region R of the plane with boundary conditions

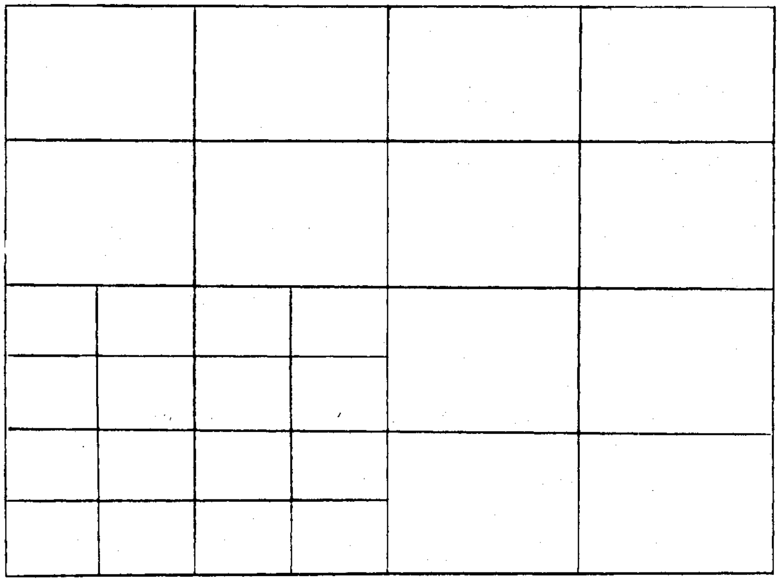
$$\alpha(\tau)\frac{\partial u}{\partial n} + \beta(\tau)u = g \,, \qquad \alpha \geq 0, \quad \beta \geq 0, \quad \alpha+\beta > 0 \tag{5.4-2}$$

given on the boundary ∂R. The parameter τ is used to describe the curve ∂R. For example, we might let R be the disk $x^2 + y^2 \leq 1$ and the boundary curve would be $x = \cos\tau$, $y = \sin\tau$, for $0 \leq \tau \leq 2\pi$. We will define a mesh region R_h

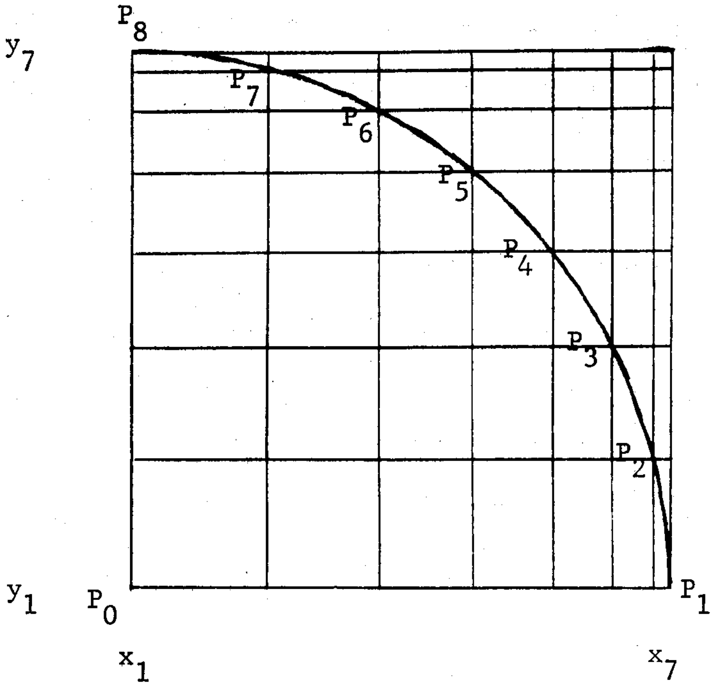
which approximates R. We first lay out a rectangular mesh whose mesh points are (x_j, y_k). That is, we specify points x_j for $1 \le j \le J$ and y_k for $1 \le k \le K$, such that $x_j < x_{j+1}$ and $y_k < y_{k+1}$. This gives us a mesh such as the one below.



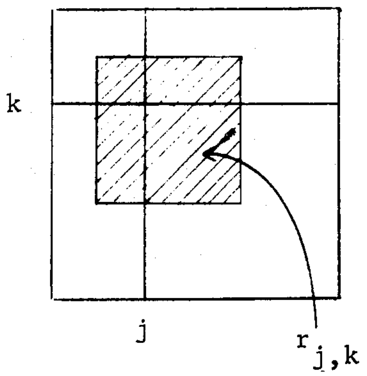
Note that we do not allow a "non-product" mesh such as the following:

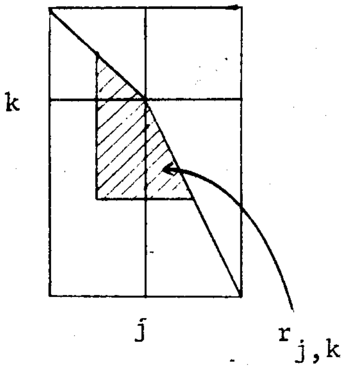


We next form a boundary curve in this mesh composed of straight line segments joining mesh points. These segments may be vertical, horizontal, or diagonal. To approximate a quadrant of the disk we might have used the following mesh.

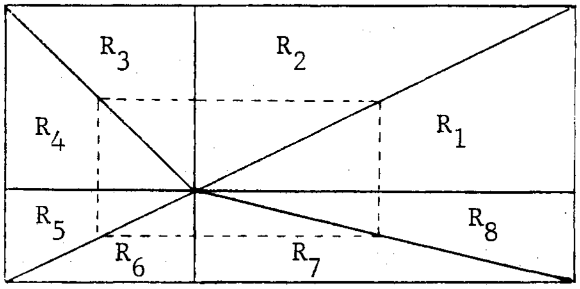


Note that we were forced to use unequal mesh spacing in order that the boundary mesh points $(P_1,\dots,P_8)$ lie on the boundary circle. The mesh boundary is then composed of the straight lines joining the points $P_0, P_1, \dots, P_8, P_0$. The set R_h is composed of all mesh points on or inside this boundary curve. Given any mesh point (x_j, y_k) we form a region $r_{j,k}$ by forming a rectangle of sides $x_j+\frac{1}{2}(x_{j+1}-x_j)$, $x_j-\frac{1}{2}(x_j-x_{j-1})$, $y_k+\frac{1}{2}(y_{k+1}-y_k)$, $y_k-\frac{1}{2}(y_k-y_{k-1})$. The region $r_{j,k}$ is formed by taking the intersection of this rectangle with the region R_h (note that we sometimes mean R_h to be the set of mesh points and other times the region bounded by the rectilinear mesh boundary ∂R_h). Examples of the definition of $r_{j,k}$ are illustrated below.





We allow the coefficient function D(x,y) to have discontinuities along the vertical, horizontal, or diagonal lines joining the mesh points. Thus D(x,y) is continuous in the eight regions $R_1,\ldots,R_8$ shown below.



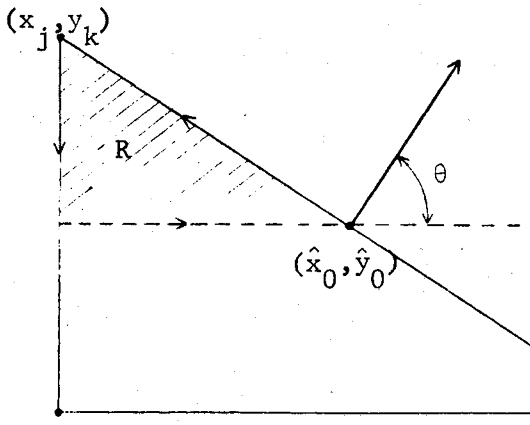
We integrate the differential equation over each of these regions and then use Green's theorem to replace surface integrals by line integrals. Green's theorem states that for any smooth functions T(x,y) and S(x,y)

$$\int\int_R (S_x - T_y)dxdy = \int_{\partial R} Tdx + Sdy$$

where the boundary integral is taken in the positive, or counter-clockwise sense. Application of this formula to the differential equation (5.4-1) yields

$$\int_{\partial R} -Du_x dy + Du_y dx + \int\int_R Eudxdy - \int\int_R fdxdy = 0 \qquad (5.4\text{-}3)$$

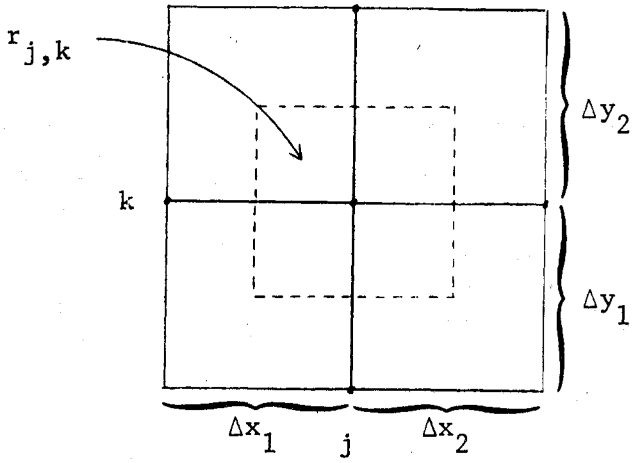
Note that in our case these line integrals are always over straight line segments. If we describe these segments by a parameter τ, then $x = \hat{x}_0 - \tau\sin\theta$, $y = \hat{y}_0 + \tau\cos\theta$ where θ is the angle that the normal to the segment makes with the positive x-axis and $(\hat{x}_0,\hat{y}_0)$ is the initial point for the segment.



We have $0 \le \tau \le L$ where L is the length of the segment. If $\partial u/\partial n$ denotes the derivative of u in a direction normal to the segment, then $\partial u/\partial n = u_x\cos\theta + u_y\sin\theta$. On the segment $dx = -\sin\theta d\tau$, $dy = \cos\theta d\tau$, therefore equation (5.4-3) becomes

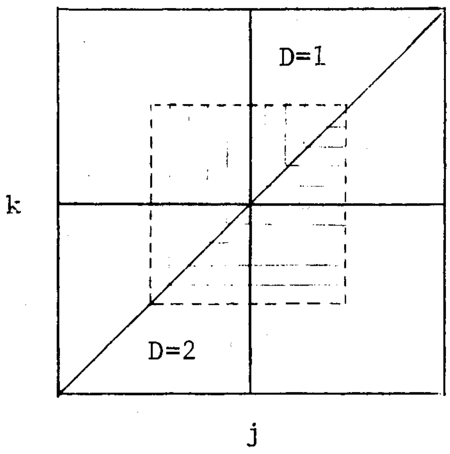
$$-\int_{\partial R} D(\tau)\,\frac{\partial u}{\partial n}\,d\tau + \iint_R Eudxdy = \iint_R fdxdy \qquad (5.4\text{-}4)$$

At each mesh point (x_j,y_k) of R_h where $u_{j,k}$ is unknown, we use the above integral to derive an equation for this unknown. For example, we will assume D(x,y) and E(x,y) are constant and derive an equation for the point below.



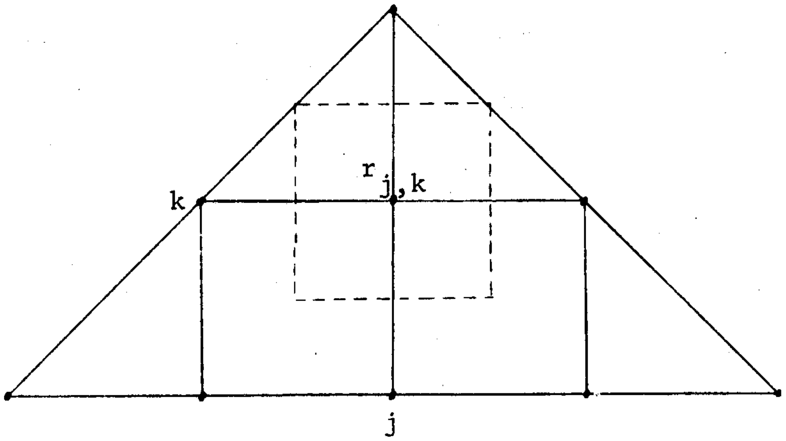
Problem 5.4-1. Derive a difference equation for the unknown $u_{j,k}$ at the mesh point shown above (D and E are constant). Hint: Approximate the normal derivative on the right side of $r_{j,k}$ by $\frac{u_{j+1,k} - u_{j,k}}{\Delta x_2}$, and similarly for the other sides.

Problem 5.4-2. Assume D(x,y) = 1 in the upper diagonal part of $r_{j,k}$ and D(x,y) = 2 in the lower diagonal part of $r_{j,k}$ (see the figure below). Using the fact that $D\partial u/\partial n$ must be continuous along an interface, derive a difference equation at the point (x_j,y_k).

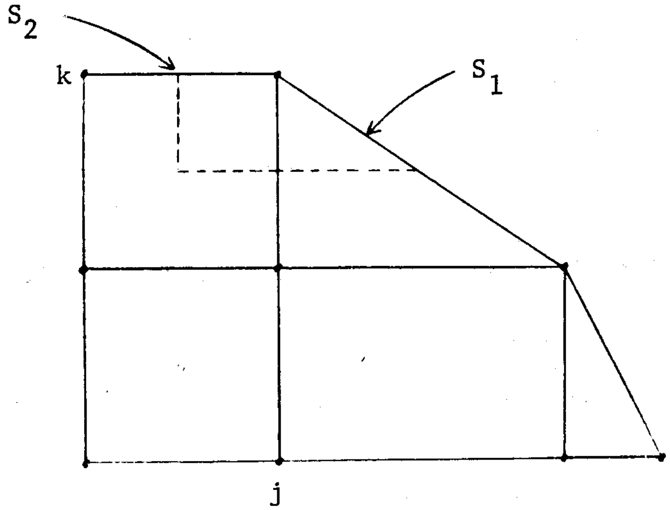


Hint: Use equation (5.4-4) integrating first over the vertical cross-hatched region, then over the horizontal cross-hatched region, then add the integrals. Note that the line integrals along the diagonal will cancel under the addition.

Next we will consider application of the boundary condition of equation (5.4-2). If $\alpha = 0$ at a point on the boundary, then the value of $U_{j,k}$ at that point is determined trivially from the boundary conditions. Note that if (x_j,y_k) lies in the interior of R_h, then the region $r_{j,k}$ also lies in the interior and we have the situation described in the previous problems. As an example, consider the figure below.



If $\alpha(\tau) \neq 0$, then we can derive an equation as follows. We will consider the case illustrated in the figure below.



Problem 5.4-3. Assume D and E are constant, and $\alpha(\tau) \neq 0$ along the boundary segments shown above. Derive an equation for the unknown boundary value $u_{j,k}$. Hint: On the side labeled S_1, approximate the line integral by

$$\int_{S_1} -D\frac{\partial u}{\partial n}\,d\tau \cong -D(x_{j+\frac{1}{2}},\ y_{k-\frac{1}{2}})u_n(x_{j+\frac{1}{2}},y_{k-\frac{1}{2}})\frac{L_1}{2}$$

Here L_1 is the length of the segment between (x_j,y_k) and (x_{j+1},y_{k-1}) and $(x_{j+\frac{1}{2}},y_{k-\frac{1}{2}})$ is the midpoint of this segment. Note that $u_n = \frac{\partial u}{\partial n}$ can be evaluated in terms of the boundary values of $U_{j,k}$ from the boundary condition of equation (5.4-2). What would you do if $\alpha(x_j,y_k) \neq 0$, but $\alpha(x_{j+\frac{1}{2}},y_{k-\frac{1}{2}}) = 0$?

The derivation of these difference equations can be done by the computer. Otherwise the scheme would be difficult to use. This is described by Spanier [1967]. One inputs a description of the mesh geometry along with the boundary conditions to the computer program, and the latter constructs the coefficients in the difference scheme in accordance with the method described above. This method will lead to a matrix equation for the unknown $u_{j,k}$, namely AU = F, where U is the vector of unknowns. Note that the matrix will depend on the ordering of the components of the vector U.

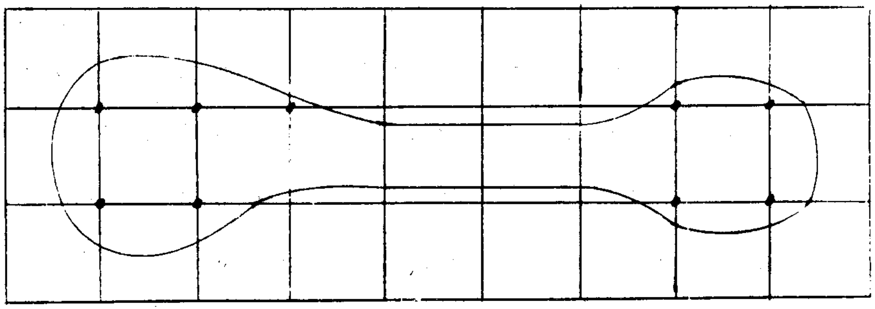
Problem 5.4-4. Show that the matrix A is symmetric, diagonally dominant, with positive diagonal entries. It is possible to show that A is positive definite. (Assume $\beta \neq 0$ on the two line segments joining at least one boundary point.) To prove this you must specify exactly what is done in case $\alpha = 0$ for some boundary points. Note that you must also specify the method used to approximate the integral

$$\iint_R Eu\,dx\,dy$$

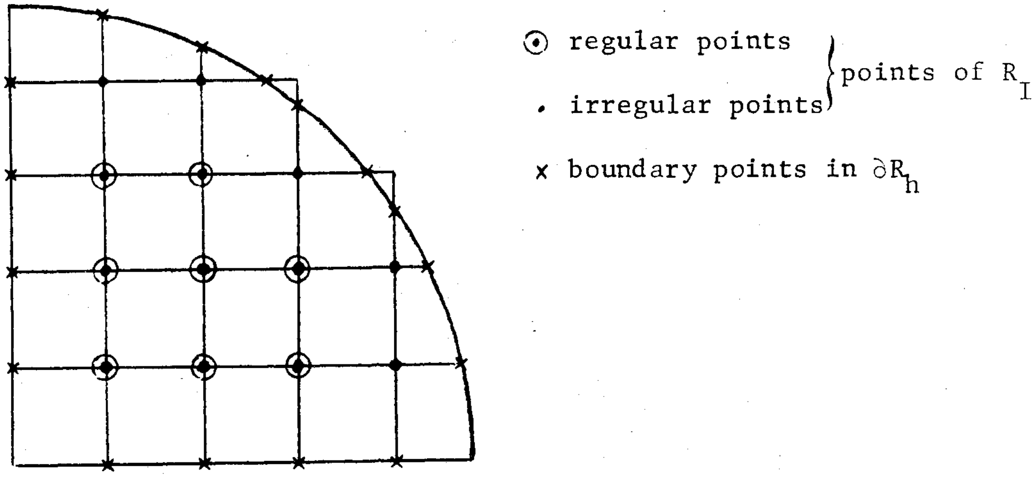
You might need to change the method used to approximate the line integral in problem 5.4-3.

5.4.2 Difference schemes based on interpolation. We will briefly describe two difference schemes for Laplace's equation with Dirichlet boundary conditions on a general (i.e., nonrectangular) region. For further information we refer the reader to the book by Forsythe and Wasow [1960] and also to Greenspan [1965].

We first describe the mesh. We cover the domain R with a rectangular mesh (x_j,y_k), $1 \leq j \leq J$, $1 \leq k \leq K$. We assume $x_j < x_{j+1}$, $y_k < y_{k+1}$. We also assume that if the point (x,y) is in R, then $x_1 \leq x \leq x_J$, $y_1 \leq y \leq y_K$. That is, we assume the domain R is contained within the mesh. We let R_I denote the set of mesh points which lie in the interior of R. We will assume R_I is connected in R. By this we mean that any two points of R_I can be joined by a series of mesh segments (that is $(x_j,y_k) \rightarrow (x_{j\pm1},y_k)$ or $(x_j,y_k) \rightarrow (x_j,y_{k\pm1})$) which lie in the interior of R. We wish to exclude cases such as the one illustrated below. The points of R_I are marked by "•".



We say that a point of R_I is a regular point if the line segments joining the point to its four neighbors all lie in the closure of R. The closure of R is the interior of R plus the boundary curve ∂R of R. If a point of R_I has a line segment joining it to a neighboring mesh point and this line segment is not contained in the interior of R, then the segment must intersect the boundary. We choose that boundary point on the segment which is cloest to the original point and add it to the mesh. The set of such points we denote by ∂R_h. Our mesh R_h is then composed of the union of R_I with ∂R_h. If we have Dirichlet boundary conditions, then we know the value of the solution at all points in ∂R_h. We need to obtain a difference equation for each point of R_I. The figure below illustrates the situation when R is the first quadrant inside a circle.



Now that we have defined the mesh we are ready to construct the finite difference approximation. Since we are mainly interested in methods to approximate the boundary conditions on a nonrectangular domain we will only discuss Laplace's equation $u_{xx} + u_{yy} = 0$. We will further restrict our discussion to Dirichlet boundary conditions.

Problem 5.4-5. We will first describe the approximation of u_{xx} at the point (x_j,y_k). Denote $u_{xx}(x_j,y_k)$ by $u_{xx}(P)$, and use $u(E)$ for $u(x_{j+1},y_k)$, $u(W)$ for $u(x_{j-1},y_k)$. Let $h_W = x_j - x_{j-1}$, $h_E = x_{j+1} - x_j$. Assume that $u(x,y)$ is sufficiently differentiable. Show that the following difference approximation is valid.

$$u_{xx}(P) = \frac{\dfrac{u(E) - u(P)}{h_E} - \dfrac{u(P) - u(W)}{h_W}}{\dfrac{h_W + h_E}{2}} + R_x''(P)$$

$$\text{where } R_x''(P) = \frac{h_W - h_E}{3} u_{x^3}(P) - \frac{h_W^2 - h_W h_E + h_E^2}{12} u_{x^4}(\xi, y_P)$$

$$x_P - h_W \le \xi \le x_P + h_E$$

Note that if we use an equally spaced mesh ($h_E = h_W = h$), then our error term is $0(h^2)$ instead of $0(h)$ for $h_E \neq h_W$. However, it may be to our advantage to use an unequally spaced mesh if the solution changes rapidly in one part of the region. We may then get by with fewer mesh points. Near the boundary we may need to add more mesh points.

At a regular interior point $P = (x_j, y_k)$ we will use the differential equation to obtain an equation for the unknown $u_{j,k}$. We denote the neighbors of P by $E = (x_{j+1}, y_k)$, $W = (x_{j-1}, y_k)$, $N = (x_j, y_{k+1})$, $S = (x_j, y_{k-1})$, and h_E, h_W, h_N, h_S have the obvious meaning. The difference scheme is

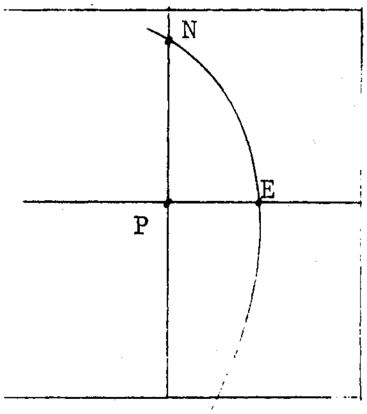
$$L_h(U)(P) = \frac{\dfrac{U(E) - U(P)}{h_E} - \dfrac{U(P) - U(W)}{h_W}}{\dfrac{h_E + h_W}{2}} + \frac{\dfrac{U(N) - U(P)}{h_N} - \dfrac{U(P) - U(S)}{h_S}}{\dfrac{h_N + h_S}{2}} = 0 \tag{5.4-5}$$

Some of these neighboring points may be boundary points, in which case the values of U at these points are known. These values would then be moved over to the right side of the equation to form a matrix equation (square matrix A) AU = F for the unknown interior values of U. Note that the above equation reduces to the usual equation (5.3-2) in case $h_E = h_W$, $h_N = h_S$.

Next we will describe three methods to obtain an equation for the irregular interior points of the mesh. The first is a very crude interpolation. If P is an irregular mesh point, one of its neighbors must lie on the boundary of the mesh. Denote the closest such neighboring point by Q and let $u(P) = u(Q)$ be the equation for $u(P)$. Note that $u(Q)$ is known since Q lies on the boundary ∂R. Instead of $u(Q)$ we might use an average or integral of $u(\tau)$ along the boundary ∂R, for example

$$u(P) = \int_E^N u(\tau)\,d\tau \Big/ \int_E^N d\tau$$

where the integrals refer to the path shown below.

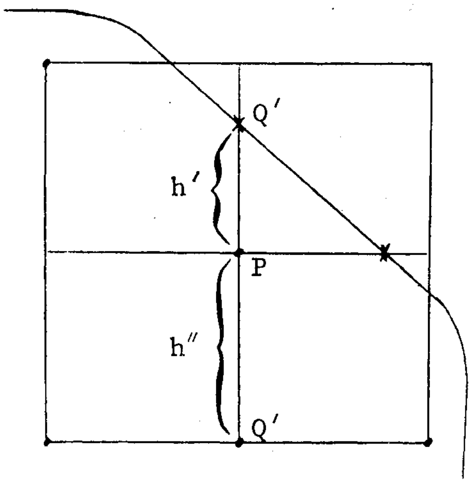


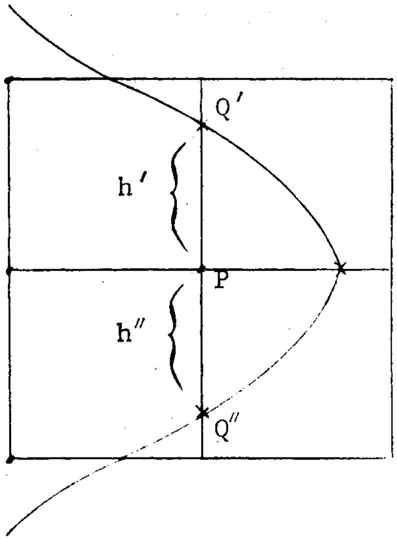
This is a rather crude approximation.

The next method requires the choice of a neighboring point Q' on the boundary mesh ∂R_h closest to P. We then let Q'' denote the neighboring point on the opposite side of P from Q'. We use simple interpolation to form the equation

$$L_h(U)(P) = U(P) - \frac{h''u(Q') + h'U(Q'')}{h' + h''} = 0$$

Two examples are illustrated below.





Problem 5.4-6. Let R be the triangle with vertices (0,0), (0,1.), (0.5,0). Let the mesh be determined by $x_j = (j-1)/6$, $1 \le j \le 4$, $y_k = (k-1)/6$, $1 \le k \le 7$. Then $\Delta x = \Delta y = 1/6$. Note that there are four points in the interior R_I of this mesh. Assume we wish to solve Laplace's equation with Dirichlet boundary conditions on this mesh. Write out the finite difference scheme obtained from the second of the above methods.

A third difference scheme is one due to Shortley and Weller [1938]. In this case we simply use equation (5.4-5) at all points of R_I.

Problem 5.4-7. Write out the finite difference equations for the mesh in problem 5.4-6 using the method of Shortley and Weller.

Problem 5.4-8. Show that all three of these methods lead to matrix equations for the unknown values of $U_{j,k}$ whose matrices A are irreducible and diagonally dominant and also satisfy $a_{ij} \le 0$ if $i \ne j$, $a_{ii} > 0$. We have assumed that the mesh is connected in the sense described in the beginning of this section. Give an example to show that the matrix may not be irreducible unless the mesh is connected. Are these matrices symmetric?

We know the matrices obtained from these methods are nonsingular since these matrices are irreducible and diagonally dominant. Therefore we can solve these systems of equations for the unknown vector U.

6. THE ITERATIVE SOLUTION OF LINEAR EQUATIONS

In the preceding chapter we considered finite difference approximations to elliptic partial differential equations such as Laplace's equation. These methods all require the solution of a linear system of equations. The order of this system may be very large--in some cases over 100,000. However, these linear systems frequently have special properties which permit us to construct particularly efficient iterative methods for their solution. The purpose of this chapter is to discuss two iterative methods for the solution of these systems--the successive overrelaxation (SOR) and alternating-direction-implicit (ADI) methods. A great deal of work has been done on the development of these methods. For more information the reader can consult books by Forsythe and Wasow [1960], Varga [1962], or Wachspress [1965]. There are many journal articles on this subject. We will start with some general comments on iterative methods and then consider the SOR and ADI methods.

6.1 General remarks on the convergence of iterative methods.

We will consider iterative methods for the solution of the matrix equation $Ax = b$, where x and b are n-dimensional vectors, and A is a matrix of order n. We are trying to find the zero of the vector function $f(x) = Ax-b$. We can convert this to a fixed point problem by defining the function $g(x)$ by $g(x) = x-f(x) = (I-A)x+b$. Then we are looking for fixed points; that is, vectors x such that $x = g(x)$. The easiest iterative method for this problem is to choose an initial guess $x^{(0)}$ for the vector and then define $x^{(\nu+1)} = g(x^{(\nu)})$ for $0 \leq \nu$. Under the proper conditions the vectors $x^{(\nu)}$ will converge to the solution x. An introductory course in numerical analysis will usually consider conditions under which this iterative process will converge for scaler functions $g(x)$ of a single unknown x. We need to consider this question for matrix equations in the form

$$x = Mx + b$$

Here M is a matrix of order n, x and b vectors of dimension n. The iterative process is

$$x^{(\nu+1)} = Mx^{(\nu)} + b, \qquad x^{(0)} \text{ given} \tag{6.1-1}$$

If we denote the error $x-x^{(\nu)}$ by $e^{(\nu)}$, then by subtraction of the above equations we obtain

$$e^{(\nu+1)} = Me^{(\nu)}$$

From this equation we have $e^{(1)} = Me^{(0)}$, $e^{(2)} = Me^{(1)} = M^2e^{(0)}$, and by induction we can prove

$$e^{(\nu)} = M^{\nu}e^{(0)}.$$

Thus convergence is dependent on the ν^{th} powers of M. We need to find conditions under which M^{ν} will approach zero. If M is in some sense small, we would expect convergence. The following problem verifies this.

Problem 6.1-1. If the absolute row sums of M are strictly less than one, then the iterative process of equation (6.1-1) will converge. The condition is $\sum_{j=1}^{n} |m_{ij}| < 1$ for $1 \le i \le n$. Hint: Let $\epsilon_{\nu} = \max_{1\le i\le n} |e_i|$ where $\{e_i^{(\nu)}\}$ is the error vector, $e_i^{(\nu)} = x_i - x_i^{(\nu)}$. Then show $\epsilon_{\nu+1} < \epsilon_{\nu}$. This implies $\lim_{\nu\to\infty} e_{\nu} = 0$.

6.1.1 The convergence rate of the matrix iteration. We say that a matrix M is convergent if $\lim_{\nu\to\infty} M^\nu = 0$ (the zero on the right side denotes a matrix whose elements are all zero). There is a close relation between this property and the spectral radius $\sigma(M)$ of the matrix. (The spectral radius is defined by $\sigma(M) = \max_i |\lambda_i|$ where λ_i are the eigenvalues of M.) In fact, a matrix M is convergent if and only if $\sigma(M) < 1$. We refer the reader to the book by Isaacson and Keller [1966, p. 14] for a proof. A special case of this result is the following.

Problem 6.1-2. Suppose a matrix M of order n has n linearly independent eigenvectors. Show that M is convergent if and only if $\sigma(M) < 1$. Hint: Given any vector x we have $x = \alpha_1 v^{(1)} + \ldots + \alpha_n v^{(n)}$ where $v^{(j)}$ are the eigenvectors. Use this to show that $M^n x \to 0$ for all x. Then show $\lim_{\nu\to\infty} M^n = 0$.

Problem 6.1-3. Given an integer m and positive number ϵ, define a (2×2) matrix M and a vector x such that $\|x\|_2 \le 2$, $\|M^\nu x\|_2 \ge 1$ for $\nu \le m$ and $|\lambda_i|^m < \epsilon$, $1 \le i \le 2$, $\lambda_1 \ne \lambda_2$, where λ_i are the eigenvalues of M. Hint: Find a matrix M whose eigenvectors are $(1,0)$ and $(1,\epsilon)$.

This problem shows us that even if a matrix M is convergent and has n independent eigenvectors, the powers M^ν may grow quite large before beginning to decay to zero. The condition $\sigma(M) < 1$ on the spectral radius insures convergence, but it does not tell us how many iterations will be required to reduce the error $M^\nu e^{(0)}$ to a certain level. In the case of a symmetric or Hermitian matrix the spectral radius does give us an upper bound for the norm $\|M^\nu x\|$. This is due to the fact that the eigenvectors of a symmetric matrix form an orthogonal basis.

Problem 6.1-4. Let M be a symmetric matrix. Show that $\|M^{\nu}x\|_2 \leq (\sigma(M))^{\nu}\|x\|_2$ for any vector x. Here we use the Euclidean norm $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

The spectral radius can also fail to provide a good indication of the initial rate of convergence if there is an "eigenvector deficiency." By this we mean that there are fewer than n eigenvectors where n is the order of the matrix. The following problem will illustrate this case.

Problem 6.1-5. Let the matrix M be defined by

$$M = \begin{vmatrix} \rho & 1 \\ 0 & \rho \end{vmatrix}$$

Show that

$$M^{\nu} = \begin{vmatrix} \rho^{\nu} & \nu\rho^{\nu-1} \\ 0 & \rho^{\nu} \end{vmatrix}$$

Also show that given any $\epsilon > 0$ and any $R > 0$ it is possible to choose $\rho > 0$ such that for some ν, $\rho^{\nu} < \epsilon$ and $\|M^{\nu}\|_2 > R$.

Even though the spectral radius is not an ideal indicator of the convergence rate, it is usually the best we can do and generally provides reasonable results. Most analyses of iterative methods are based on the spectral radius.

We can define an average convergence rate as follows:

$$R_{\nu} \equiv R_{\nu}(M) \equiv -\frac{\ln\|M^{\nu}\|_2}{\nu}$$

Then

$$\|M^{\nu}x\|_2 \le \|M^{\nu}\|_2 \ \|x\|_2 \le e^{-\nu R_{\nu}} \ \|x\|_2 \tag{6.1-2}$$

The larger R_{ν}, the smaller the error. Note that $-\ln\|M\|_2 \le R_{\nu}$ for all ν, thus we have a lower bound for R_{ν}. This may not be too useful, since we can have $\|M\|_2 > 1$ even though M is convergent. If R_{ν} were bounded below by R $R_{\nu} \ge R$, then 1/R would bound the number of steps required to reduce the norm by e^{-1}. This follows from inequality (6.1-2). If $\nu \ge 1/R$, then

$$\|M^{\nu}x\| \le e^{-\nu R_{\nu}} \ \|x\| \le e^{-1} \ \|x\| \ .$$

The spectral radius can be related to the rate of convergence by means of the following results [for a proof see Varga, 1962, p. 67]. We define R_{∞} by $R_{\infty} = -\ln(\sigma(M))$. Then we have

$$\lim_{\nu\to\infty} R_{\nu}(M) = R_{\infty} \equiv -\ln(\sigma(M)) \tag{6.1-3}$$

$$R_{\nu} \le R_{\infty} \qquad \text{for all } \nu \tag{6.1-4}$$

Thus the convergence rate based on the spectral radius gives an optimistic estimate of the norm $\|M^{\nu}\|$. We have $\|M^{\nu}\| = e^{-\nu R_{\nu}} \ge e^{-\nu R_{\infty}}$. However, note that the error for a given vector may be better than indicated by the spectral radius; we may have

$$\|M^{\nu}x\| < e^{-\nu R_{\infty}} \ \|x\|$$

Problem 6.1-6. Provide an example where the inequality above holds.

We will not prove equation (6.1-3) but will give some indication why it is true.

Problem 6.1-7. If the matrix M has n independent eigenvectors, then show that equation (6.1-3) holds. Hint: In this case the matrix S of eigenvectors reduces M to diagonal form, $S^{-1}MS = D$, where $D = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ with λ_i the eigenvalues of M. Then $M^\nu = SD^\nu S^{-1}$. Now use $\|M^\nu\| \le \|S\|\|S^{-1}\|\|D^\nu\|$.

Problem 6.1-8. Prove relation (6.1-4).

6.1.2 <u>Two iterative methods--Jacobi and Gauss-Seidel</u>. In this section we will return to our original linear system $Ax = b$ and define some iterative schemes for the solution of this problem. First we will consider the Jacobi iteration. We write $A = D - E - F$ where D is a diagonal matrix consisting of the diagonal elements a_{ii} of A, E is a lower triangular matrix of the elements $-a_{ij}$ $(i > j)$ and F an upper triangular matrix of the elements $-a_{ij}$ $(i < j)$. We assume that the diagonal elements of A are nonzero. The Jacobi iteration is defined by

$$Dx^{\nu+1} = (E+F)x^\nu + b \tag{6.1-5}$$

We can write this in the form

$$x^{\nu+1} = (L+U)x^\nu + D^{-1}b$$

where $L = D^{-1}E$, $U = D^{-1}F$. Note that equation (6.1-5) is effectively explicit. The formula for a component of $x^{\nu+1}$ is

$$x_i^{\nu+1} = \frac{1}{a_{ii}}\left(\sum_{j\neq i} -a_{ij}\, x_j^\nu + b_i\right)$$

If A is strictly diagonally dominant, then the Jacobi method will converge.

Problem 6.1-9. If $a_{ii} > \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}|$ for $1 \leq i \leq n$, then the Jacobi method will converge.

We can modify the Jacobi method by including more of the matrix A on the left side. We then obtain the Gauss-Seidel method defined as follows:

$$(D-E)x^{\nu+1} = Fx^{\nu} + b$$

Since $a_{ii} \neq 0$ and E is lower triangular, we know that (D-E) is nonsingular. This scheme is effectively explicit if we solve for the components of $x^{\nu+1}$ in ascending order. Suppose we have already computed $x_i^{\nu+1}$ for $i < k$. Then $x_k^{\nu+1}$ is obtained from the relation

$$x_k^{\nu+1} = \frac{1}{a_{kk}} \left(- \sum_{1 \leq j < k} a_{kj}\, x_j^{\nu+1} - \sum_{k<j} a_{kj}\, x_j^{\nu} + b_k \right)$$

We can write the Gauss-Seidel iteration in the form

$$x^{\nu+1} = Lx^{\nu+1} + Ux^{\nu} + D^{-1}b$$

where $L = D^{-1}E$, $U = D^{-1}F$.

The convergence of the Jacobi iteration is determined by the spectral radius $\sigma(B)$ where $B = L + U$ since $x^{\nu+1} = Bx^{\nu} + D^{-1}b$. The convergence of the Gauss-Seidel iteration is determined by the spectral radius $\sigma(M_1)$ of $M_1 = (I-L)^{-1}U$ since $x^{\nu+1} = M_1x^{\nu} + (I-L)^{-1}D^{-1}b$. The following theorem relates the convergence of the two methods in the case where $B = L + U$ is non-negative. This is frequently the case for matrices A arising from elliptic PDE problems. Note that B is non-negative if the original matrix A satisfies the condition $a_{ij}/a_{ii} \leq 0$ for all $i \neq j$. We refer the reader to Varga [1962, p. 70] for a proof of the following statement. The proof is based on the Perron-Frobenius theory of non-negative matrices.

Assume the Jacobi matrix $B = L + U$ is non-negative. Let M_1 be the matrix which defines the Gauss-Seidel iteration. Then one and only one of the following relations hold:

a) $\sigma(B) = \sigma(M_1) = 0$

b) $0 < \sigma(M_1) < \sigma(B) < 1$

c) $\sigma(M_1) = \sigma(B) = 1$

d) $1 < \sigma(B) < \sigma(M_1)$

This theorem implies that if one of these methods converges, then the other must also converge. Also, the asymptotic convergence rate of the Gauss-Seidel iteration ($R_\infty(M_1) = -\ln \sigma(M_1)$) is larger than that of the Jacobi iteration.

Problem 6.1-10. Assume A is strictly diagonally dominant, that is $\max_i \sum_{j\neq i} \left|\frac{a_{ij}}{a_{ii}}\right| = r < 1$. Show that the Gauss-Seidel iteration will converge.
Hint: Use induction to show that each component of the error vector satisfies $|e_k^{(\nu+1)}| \leq r \max_i |e_i^{(\nu)}|$. Here $x^{\nu+1} = Lx^{\nu+1} + Ux^{\nu} + D^{-1}b$, $Ax = b$, $e^{\nu} = x - x^{\nu}$.

6.1.3 <u>Acceleration of convergence by means of successive overrelaxation (SOR)</u>. We will first treat the general case where we have an iterative method defined by a "splitting" of the matrix A. That is, we wish to solve $Ax = b$ where A can be "split" into $A = N-P$. We define an iterative scheme by $Nx^{\nu+1} = Px^{\nu} + b$. We assume that N is non-singular. In order that the scheme be usable, we must also be able to solve for $x^{\nu+1}$ without too much trouble.

The convergence of this iterative method is determined by the spectral radius of $M = N^{-1}P$ since we can rewrite the scheme in the form $x^{\nu+1} = Mx^{\nu} + N^{-1}b$. It is sometimes possible to reduce the spectral radius by means of the following "overrelaxation" procedure. We define first $\hat{x}^{\nu+1}$ by $N\hat{x}^{\nu+1} = Px^{\nu} + b$, then $x^{\nu+1}$ by $x^{\nu+1} = x^{\nu} + \omega(\hat{x}^{\nu+1} - x^{\nu})$. Here ω is a real parameter. If $\omega > 1$, then we have overrelaxation; if $\omega < 1$, underrelaxation. We will now show how this procedure may reduce the spectral radius. We can write this procedure in the form

$$Nx^{\nu+1} = [(1-\omega)N + \omega P]x^{\nu} + \omega b$$

or

$$x^{\nu+1} = M_{\omega}x^{\nu} + \omega N^{-1}b , \qquad M_{\omega} = N^{-1}[(1-\omega)N + \omega P]$$

The original iteration which does not use overrelaxation is determined by the matrix $M_1 = N^{-1}P$. The matrix for the overrelaxation process is $M_{\omega} = (1-\omega)I + \omega M_1$. If we denote the eigenvalues of the matrix M_1 by $\{\mu_1,\ldots,\mu_n\}$, then the eigenvalues of M_{ω} are $\lambda_i = 1-\omega + \omega\mu_i$. We can use this relation to choose ω so that the spectral radius of M_{ω} is made a minimum; that is, choose ω_0 so that $\sigma(M_{\omega_0}) \le \sigma(M_{\omega})$ for all ω.

Problem 6.1-11. Assume that $-1 < \mu_1 < \mu_2 < \ldots < \mu_n < 1$. Show that the best choice of ω is $\omega_0 = \dfrac{2}{2 + \mu_1 + \mu_n}$, then $\sigma(M_{\omega_0}) \le \sigma(M_{\omega})$. Also show that $\sigma(M_{\omega_0}) < \sigma(M_1)$ if $\mu_1 + \mu_n \neq 0$. If $\mu_1 = -1 + 3\epsilon$, $\mu_n = 1 - \epsilon$ where $0 < \epsilon \ll 1$, then what improvement in the asymptotic convergence rate R_{∞} would you expect? If $\mu_1 = 0$, $\mu_n = 1 - \epsilon$, then what improvement would you expect? Assuming the asymptotic convergence rate to be a good indication of the actual convergence rate, how much computer time would you expect to save by use of $\omega = \omega_0$ rather than $\omega = 1$?

Overrelaxation for the Gauss-Seidel iteration is defined somewhat differently from what we did above. The objective is the same--to reduce the spectral radius, thus increasing the convergence rate. In component form, the Gauss-Seidel overrelaxation is

$$a_{kk}\hat{x}_k^{\nu+1} = -\sum_{j<k} a_{kj}x_j^{\nu+1} - \sum_{j>k} a_{kj}x_j^{\nu} + b_k \qquad (6.1\text{-}6)$$

$$x_k^{\nu+1} = x_k^{\nu} + \omega\left(\hat{x}_k^{\nu+1} - x_k^{\nu}\right)$$

We relax each component $\hat{x}_k^{\nu+1}$ immediately after computing $\hat{x}_k^{\nu+1}$ instead of waiting until the vector $\hat{x}^{\nu+1}$ is computed.

Problem 6.1-12. Show that the Gauss-Seidel overrelaxation defined by equations (6.1-6) can be written in matrix form as

$$x^{\nu+1} = (I - \omega L)^{-1}\,[\omega U + (1-\omega)I]x^{\nu} + \omega D^{-1}b$$

where the matrices L, U, and D are defined in equation (6.1-5).

This procedure is frequently called successive overrelaxation or SOR. Our problem is to choose ω so as to minimize the spectral radius of $M_\omega = (I - \omega L)^{-1}\,[\omega U + (1-\omega)I]$.

6.2 The theory of successive overrelaxation.

In this section we will analyze the successive overrelaxation iteration. Our main objective is to obtain an expression for the spectral radius of the matrix M_ω which defines the SOR iteration. The spectral radius of SOR is related to that of the Jacobi iteration. This relation is very useful (from a practical as well as a theoretical point of view) for the determination of an optimal ω for the SOR iteration.

6.2.1 The convergence of SOR--the Kahan and Ostrowski-Reich theorems.

We will first give a result due to Kahan [1958]. Let M_ω be the matrix which defines the Gauss-Seidel iteration for an arbitrary matrix A (arbitrary, except $a_{ii} \neq 0$), namely $M_\omega = (I - \omega L)^{-1} (\omega U + (1-\omega)I)$. Then the spectral radius of M_ω satisfies $\sigma(M_\omega) \geq |\omega-1|$.

Problem 6.2-1. Prove the above theorem by Kahan. Hint: Let $\varphi(\lambda) = \det(\lambda I - M_\omega)$. Show that $\varphi(\lambda) = \det((\lambda+\omega-1)I - \omega\lambda L - \omega U)$. Note that $\det(I - \omega L) = 1$. If $\lambda_i(\omega)$ are the roots of $\varphi(\lambda) = 0$, then $\varphi(0) = (-1)^n \prod_i^n \lambda_i(\omega)$. Show that $\varphi(0) = (\omega-1)^n$. The result follows from this.

Next we will prove a theorem due to Ostrowski [1954]. The conditions of this theorem are frequently satisfied by the difference schemes used for elliptic PDE problems.

We assume A and D are Hermitian, D positive definite, with $A = D-E-E^*$, where E is lower triangular ($e_{ij} = 0$, if $j \geq i$). Let the iteration be defined by $(D - \omega E)x^{\nu+1} = (\omega E^* + (1-\omega)D)x^{\nu} + \omega b$ so that the iteration matrix is $M_\omega = (D - \omega E)^{-1} (\omega E^* + (1-\omega)D)$. Then $\sigma(M_\omega) < 1$ if and only if A is positive definite and $0 < \omega < 2$.

Note that if $A = D-E-E^*$ and A is Hermitian, D diagonal and positive definite, then $a_{ii} > 0$ for $1 \leq i \leq n$. Then $D - \omega E$ is nonsingular since E is lower triangular. Thus the Gauss-Seidel iteration satisfies the conditions of the theorem provided A is Hermitian with positive diagonal elements.
We will only outline the proof, leaving most of it as a problem. Let e_m be defined by $e_{m+1} = M_\omega e_m$ for $m \geq 1$, with e_0 given. Let $\delta_m = e_m - e_{m+1}$, then

$$(D - \omega E)\delta_m = \omega A e_m \tag{6.2-1}$$

and

$$\omega A e_{m+1} = (1-\omega)D\delta_m + \omega E^* \delta_m \tag{6.2-2}$$

Also,

$$(2-\omega)\delta_m^* D\delta_m = \omega(e_m^* A e_m - e_{m+1}^* A e_{m+1}) \tag{6.2-3}$$

Now assume A is positive definite and $0 < \omega < 2$. Let e_0 be an eigenvector of M_ω with eigenvalue λ. Then

$$\frac{2-\omega}{\omega} \; |1-\lambda|^2 \, e_0^* D e_0 = \left(1 - |\lambda|^2\right) e_0^* A e_0 \tag{6.2-4}$$

If $\lambda = 1$, then $e_1 = Me_0 = e_0$ and $\delta_0 = 0$ which implies $Ae_0 = 0$, but A is positive definite. Therefore $\lambda \neq 1$. Since A and D are positive definite, $0 < \omega < 2$, we have $1 - |\lambda|^2 > 0$, or $|\lambda| < 1$. This completes a proof of the first half of the theorem.

If M_ω is convergent, then $0 < \omega < 2$ from Kahan's theorem. Also for any initial error e_0, the sequence e_m converges to zero. Since the matrix M_ω is convergent, it must not have unity as an eigenvalue. Therefore $e_0 - e_1 = \delta_0 = (I-M_\omega)e_0 \neq 0$. Therefore

$$e_1^* Ae_1 < e_0^* Ae_0 \tag{6.2-5}$$

We can also show

$$e_{m+1}^* Ae_{m+1} \leq e_m^* Ae_m. \tag{6.2-6}$$

If A is not positive definite, then $e_0^* Ae_0 \leq 0$ for some nonzero vector e_0. But then the two equations above imply

$$e_m^* Ae_m < e_1^* Ae_1 < 0$$

for all m. Therefore e_m does not converge to zero, which contradicts the assumption that M_ω is convergent.

Problem 6.2-2. Show that equations (6.2-1) through (6.2-6) are valid.

6.2.2 The iteration matrix for SOR in a special case--the Dirichlet problem on a rectangle. The proofs we will give here apply only to Laplace's equation on a rectangle with Dirichlet boundary conditions. However, the results will apply to somewhat more general problems. We refer the reader to Varga [1962], Wachspress [1966], and Forsythe and Wasow [1960] for the proofs. The main result concerns the choice of the optimum overrelaxation parameter ω. We have from the Reich-Ostrowski theorem that SOR for the problem $Ax = b$ will converge if $0 < \omega < 2$ for reasonably general matrices A. Young [see Forsythe and Wasow, 1960] in 1950 developed a theory which permits the calculation of an optimum ω for a wide class of matrices A. We will give another development based on a paper by Keller [1958].

First we will review the description of the finite difference scheme given in section 5.3.2. The matrix A has the form

$$A = I - \theta_x(C_x + C_x^T) - \theta_y(C_y + C_y^T) \qquad (6.2\text{-}7)$$

where C_x and C_y are the lower triangular matrices of order $J*K = N$ given below. Note that the mesh points are defined by (x_j, y_k) where $x_j = ja/(J+1)$, $y_k = kb/(K+1)$, $0 \le j \le J+1$, $0 \le k \le K+1$.

$$C_x = \left|\begin{array}{cccccccc}
0&0&0&\cdots&&&&0\\
1&0&0&\cdots&&&&\\
0&1&0&\cdots&&&&\\
0&0&1&\cdots&&&&\\
\cdot&&&&&&&\\
\cdot&&&&&&&\\
\cdot&&&&&&&\\
\cdot&&&&1&0&0&0\\
\cdot&&&&0&1&0&0\\
0&0&0&\cdots&0&0&1&0
\end{array}\right|
\qquad
C_y = \begin{array}{cccccc}
1&2&3&&&K\\
0&0&0&\cdots&&0\\
I_J&0&0&\cdots&&\\
0&I_J&0&\cdots&&\\
\cdot&&&&&\\
\cdot&&&&&\\
\cdot&&&&&\\
\cdot&&&&&\\
\cdot&&&&&\\
\cdot&&&&&\\
0\;0&\cdots&&&I_J&0
\end{array}$$

Note that C_y is written in (K×K) block form where the blocks are either zero, or the identity matrix I_J of order J.

In section 6.1.2 we defined the Jacobi iteration. There we wrote the matrix A as A = D-E-F. With A as given in equation (6.2-7) above we have $D = I$, $E = \theta_x C_x + \theta_y C_y$, $F = \theta_x C_x^T + \theta_y C_y^T = E^T$. Therefore the Jacobi iteration becomes $x^{\nu+1} = (L + L^T)x^{\nu} + b$ where $L = E$. The iteration matrix is $B = L + L^T = I-A$. If we denote the eigenvalues of A by ξ_r, then the eigenvalues of B are $\lambda_r = 1 - \xi_r$, $1 \le r \le N = J*K$. The eigenvalues ξ_r were found in section 5.3.2 to be

$$\xi_r = 4\theta_x \sin^2 \frac{\pi p}{2(J+1)} + 4\theta_y \sin^2 \frac{\pi q}{2(K+1)} \tag{6.2-8}$$

where $r = p + J(q-1)$, $1 \le p \le J$, $1 \le q \le K$. Note that $\theta_x + \theta_y = \frac{1}{2}$ and therefore $0 < \xi_r < 2$, and $-1 < \lambda_r < 1$.

Problem 6.2-3. Show that the nonzero eigenvalues λ_r for the Jacobi iteration occur in pairs $\pm\lambda_r$; that is, to each $\lambda_r \neq 0$ there is a single $\lambda_{r'}$ such that $\lambda_{r'} = -\lambda_r$.

Note that the eigenvalues λ_r of B are the roots of the polynomial $\psi(\lambda)$ defined by

$$\psi(\lambda) = \det(\lambda I-B) = \det(\lambda I - \theta_x(C_x+C_x^T) - \theta_y(C_y + C_y^T)) \tag{6.2-9}$$

The matrix which defines successive overrelaxation is given in problem 6.1-12, namely

$$M_\omega = (I - \omega L)^{-1} [\omega L^T + (1-\omega)I]$$

The eigenvalues of M_ω are the roots of the equation $\varphi(\eta)$ where

$$\varphi(\eta) = \det\left(\eta I - \eta\omega(\theta_x C_x + \theta_y C_y) - \omega(\theta_x C_x^T + \theta_y C_y^T) + (\omega-1)I\right)$$

or

$$\varphi(\eta) = \det\left[(\eta+\omega-1)I - \eta\omega(\theta_x C_x + \theta_y C_y) - \omega(\theta_x C_x^T + \theta_y C_y^T)\right] \qquad (6.2\text{-}10)$$

6.2.3 The relation between the Jacobi and SOR methods. Our next objective is to relate the roots of $\varphi(\eta) = 0$ to those of $\psi(\lambda) = 0$. That is, we wish to relate the eigenvalues of the matrix of the Jacobi iteration to those of the SOR method. To do this we will need the following result which is the basic tool used by Keller [1958] (see Isaacson and Keller, 1966, p. 465).

Let α and β be nonzero scalers. Let γ_1, γ_2, γ_3, γ_4, γ_5 be arbitrary scalers. Then

$$\det\left[\gamma_1 I - \theta_x\left(\alpha\gamma_2 C_x + \alpha^{-1}\gamma_3 C_x^T\right) - \theta_y\left(\beta\gamma_4 C_y + \beta^{-1}\gamma_5 C_y^T\right)\right] \qquad (6.2\text{-}11)$$

is independent of α and β.

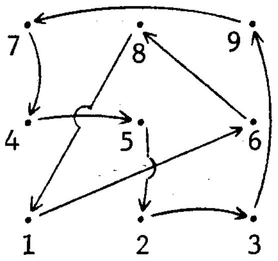
To prove this result we will need to relate the determinant to the geometry of the mesh. We are dealing with matrices of order N = J*K where the mesh points are given by (x_j, y_k), $x_j = ja/(J+1)$, $y_k = kb/(K+1)$. The unknowns $U_{j,k}$ are sought for $1 \le j \le J$, $1 \le k \le K$. The matrix whose determinant (equation (6.1-2)) we must evaluate is of order N = J*K. If we denote this matrix by G, then

$$\det(G) = \sum_{\pi} \operatorname{sgn}(\pi)\, g_{1,\pi(1)}\, g_{2,\pi(2)} \cdots g_{N,\pi(N)} \qquad (6.2\text{-}12)$$

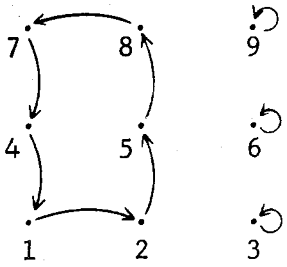
Here π denotes a permutation of the integers 1 through N. Thus the above sum contains N! terms. A permutation π is a one-to-one mapping of the set

$\{1,2,\ldots,N\}$ onto itself, thus $\pi(r)$ is integer, $1 \le \pi(r) \le N$, and $\pi(r) = \pi(s)$ only if $r = s$. To each point (x_j, y_k) in the mesh we assign an integer $r = j + J(k-1)$. This defines a one-to-one mapping of the mesh onto the integers $\{1,2,\ldots,N\}$. With each permutation we can associate a set of directed curves running from the mesh point corresponding to r to the mesh point corresponding to $\pi(r)$. The figures below indicate two examples of the curves associated with a permutation when $J = K = 3$. The permutation is defined by the columns of numbers.

r	$\pi(r)$
1	6
2	3
3	9
4	5
5	2
6	8
7	4
8	1
9	7



r	$\pi(r)$
1	2
2	5
3	3
4	1
5	8
6	6
7	4
8	7
9	9



Since each mesh point or node (denoted by r) appears exactly once in each column defining $\pi(r)$, each mesh point has exactly one curve leading into the point and one curve leaving it. Thus these curves form a set of disjoint cycles. By a cycle, we mean a sequence $r_1, r_2, r_3, \ldots, r_k$, $k \ge 2$, where $r_1 = r_k$ and $\pi(r_j) = r_{j+1}$ for $1 \le j \le k-1$. Note that the figure on the left above contains two cycles, and the one on the right four cycles.

The matrix G can have a nonzero element g_{rs} only if at least one of the matrices I, C_x, C_y, C_x^T, or C_y^T has a nonzero element in this position. Thus $g_{rs} \ne 0$ only if $s = r \pm 1$, or $s = r \pm J$. Thus the curve segment corresponding to $g_{rs} \ne 0$ is a horizontal or vertical segment of length ± 1. If the term $g_{1,\pi(1)}\, g_{2,\pi(2)} \cdots g_{N,\pi(N)}$ in the evaluation of the determinant is nonzero, then the corresponding cycles in the graph are made up of vertical and horizontal segments of length ±1. Since the cycle returns to its starting point, there must be exactly the same number of upward-directed vertical segments as there are downward-directed segments. Each upward-directed segment is associated with a term in the C_y^T matrix and each downward segment with a term in the C_y matrix. Therefore β and β^{-1} appear in multiplicative pairs in equation (6.2-12). Therefore they cancel and the determinant is thus independent of β. Similarly, the determinant is independent of α which proves the desired result.

Now we will use this result to obtain a relation between the polynomials $\psi(\lambda)$ and $\varphi(\eta)$ of equations (6.2-9) and (6.2-10). If we use $\alpha = \beta = \eta^{-\frac{1}{2}}$ in equation (6.2-11) and note that the determinant is independent of α and β, we obtain

$$\varphi(\eta) = \eta^{\frac{N}{2}} \det\left[\frac{\eta + \omega - 1}{\eta^{\frac{1}{2}}} - \omega(\theta_x C_x + \theta_y C_y) - \omega(\theta_x C_x^T + \theta_y C_y^T)\right]$$

$$\varphi(\eta) = \eta^{\frac{N}{2}} \;\psi\left(\frac{\eta + \omega - 1}{\eta^{\frac{1}{2}}}\right) \qquad (6.2\text{-}13)$$

We have assumed $\eta \neq 0$.

6.2.3 The choice of the optimum ω. We know from the Reich-Ostrowski theorem that the optimum ω must lie between zero and two. The derivation of the optimum ω will depend only on equation (6.2-13) and the fact that the nonzero roots of $\psi(\lambda) = 0$ occur in pairs with opposite sign. Equation (6.2-13) can be obtained in a more general way than we have done it here.

For example, it may hold for a nonrectangular mesh. We give the derivation in the form of a problem with several parts.

Problem 6.2-4. (1) Show that the nonzero eigenvalues η of M_ω (see equation (6.2-10) are related to the eigenvalues λ_i of B (see equation 6.2-9) by

$$\eta_i = \eta_i^{\pm} = \left[\frac{\omega\lambda_i}{2} \pm \sqrt{\left(\frac{\omega\lambda_i}{2}\right)^2 + 1 - \omega}\right]^2 \tag{6.2-14}$$

(2) Show that we need consider only non-negative λ_i (note that the λ_i are real and $-1 < \lambda_i < 1$). (3) If η_i is complex, then $\omega > 1$ and $|\eta_i| = \omega - 1$. (4) Suppose the λ_i are ordered so that $\lambda_1 \geq \lambda_i$ for all i. Let

$$\eta_1(\omega) = \left[\frac{\omega\lambda_1}{2} + \sqrt{\left(\frac{\omega\lambda_1}{2}\right)^2 + 1 - \omega}\right]^2 \tag{6.2-15}$$

Then $\frac{d\eta_1}{d\omega} \leq 0$ if $\eta_1(\omega)$ is real. (5) Let $\omega_{opt} = \frac{2}{1 + \sqrt{1 - \lambda_1^2}}$. Then show the spectral radius of M_ω is given by

$$\sigma(M_\omega) = \begin{cases} \eta_1(\omega) & \omega \leq \omega_{opt} \\ \omega - 1 & \omega \geq \omega_{opt} \end{cases} \tag{6.2-16}$$

Also show that $\sigma(M_\omega) \geq \sigma(M_{\omega_{opt}})$ for $0 < \omega < 2$.

We have thus found an optimum value for ω, that is, one which minimizes the spectral radius $\sigma(M_\omega)$, namely

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \lambda_1^2}} \qquad \text{where } \lambda_1 = 1 - 4\theta_x \sin^2 \frac{\pi}{2(J+1)} - 4\theta_y \sin^2 \frac{\pi}{2(K+1)} \tag{6.2-17}$$

Problem 6.2-5. Show that the asymptotic (i.e., R_∞) convergence rate for Gauss-Seidel ($\omega = 1$) is

$$R_{GS} = 2\delta^2\pi^2\left(\frac{1}{a^2} + \frac{1}{b^2}\right) + 0(\delta^4)$$

where $\delta^2 = \dfrac{\Delta x^2\,\Delta y^2}{2[\Delta x^2 + \Delta y^2]}$, $\Delta x = \dfrac{a}{J+1}$, $\Delta y = \dfrac{b}{K+1}$. Show that the rate of convergence for SOR with the optimum value of ω is

$$R_{SOR} = 2\delta\pi\sqrt{2\left(\frac{1}{a^2} + \frac{1}{b^2}\right)} + 0(\delta^2)$$

Let $a = b = 1$, $J = K = 49$. Estimate the number of iterations required to reduce the error by a factor of 10^{-4} for Gauss-Seidel and optimum SOR.

The effect of ω upon the spectral radius of the matrix M_ω for SOR is given in equations (6.2-15) and (6.2-16). A plot of these curves for $\lambda_1 = 0.99$ and $\lambda_1 = 0.9$ is given in figure 6.2-1 below. It is clear from these curves that if we are uncertain about the exact value of the optimum ω (ω_{opt} of equation (6.2-17)), then we do better with an overestimate rather than an underestimate.

For each pair of eigenvalues of the Jacobi iteration $\pm\lambda_i$, we will in general have two eigenvalues for SOR, namely those given by equation (6.2-14). However, if $\omega = \omega_{opt}$, then the square root in equation (6.2-14) is zero. A careful check will show that there are two eigenvalues equal to the spectral radius $\sigma(M_{\omega_{opt}})$. Furthermore, there is an eigenvector deficiency for this eigenvalue pair (see Wachspress [1966, p. 114]). Therefore we might expect the error to decay like $\nu\eta_1^{\nu-1}$ rather than η_1^ν where ν is the number of iterations and η_1 the spectral radius (see section 6.1.3).

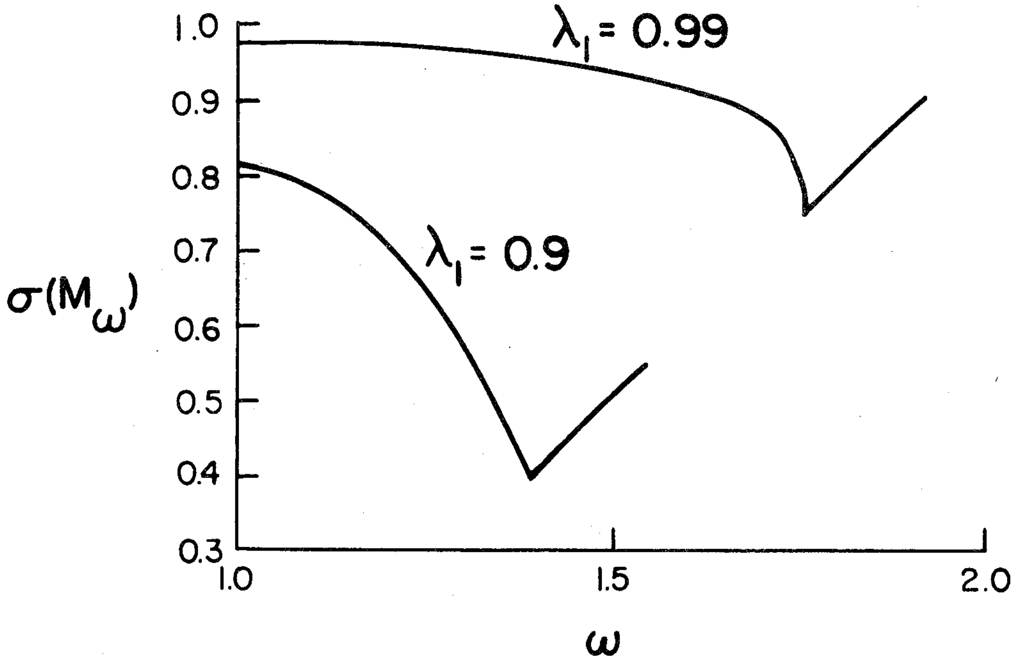


Figure 6.2-1

Spectral radius $\sigma(M_\omega)$ v.s. ω

The problem of finding the optimum value of ω is a difficult one. Equation (6.2-17) holds under rather general conditions so that this problem reduces to that of finding the spectral radius for the Jacobi iteration. For the simple case discussed above (rectangle with Dirichlet boundary conditions) we can compute the spectral radius λ_1. In more practical situations we must estimate λ_1. This can be done in a variety of ways. In problem 6.2-6 we reference some of the literature on this question. We may be able to choose ω_{opt} so that the asymptotic convergence rate R_∞ is optimized; however, we are usually more interested in the average convergence rate taken over the number of iterations we actually use, namely (see section 6.1.1)

$$R_\nu = -\frac{\ln\|M^\nu\|_2}{\nu}$$

Furthermore, we are most interested in a convergence rate based on the actual error rather than the norm of the iteration matrix, namely:

$$-\frac{\ln(\|e^\nu\|/\|e^0\|)}{\nu}$$

It would be very difficult to base a theory for the selection of ω_{opt} on this definition of the convergence rate.

Problem 6.2-6. Devise a computational procedure to choose an optimal value of ω for SOR. Base your choice on study of at least one of the following articles: Cairé [1961], Forsythe and Ortega [1960], Garabedian [1956], Kulsrud [1961], Wachspress [1966], Varga [1962, chapter 9], Forsythe and Wasow [1960, p. 368], or Hageman and Kellog [1968].

6.2.4 <u>The Young-Frankel theory for SOR</u>. In this section we will briefly describe some of the results due to Young [1954] and also Frankel [1950] concerning the convergence of SOR. In the preceding section we proved the main result (equation (6.2-10)) for the case of a rectangle with Dirichlet boundary conditions. Young's results apply to more general problems. Varga [1962] has further extended these results. We will not give any of the proofs. The books by Forsythe and Wasow [1960], Varga [1962] or Wachspress [1966] give a good account. We first need some definitions.

A matrix is said to be m-block tridiagonal ($m \geq 2$) if it exists in the form

$$\begin{vmatrix} D_1 & F_1 & 0 & \cdots & & 0 \\ E_2 & D_2 & F_2 & 0 & \cdots & 0 \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & \cdots & & 0 & E_m & D_m \end{vmatrix} \qquad (6.2\text{-}18)$$

We say that such a matrix is diagonally m-block tridiagonal if each matrix D_i is diagonal.

We say that a square matrix A has property (A) if there exists a permutation matrix P such that PAP^T is diagonally m-block tridiagonal.

In a system of linear equations $AU = B$ we say that the components U_i and U_j are coupled if $a_{ij} \neq 0$ or $a_{ji} \neq 0$.

Note that if we permute (that is, reorder) the components of U, where U is a solution of $AU = B$, then we have the equation $\hat{A}\hat{U} = \hat{B}$ where $\hat{A} = PAP^T$

and P is the permutation of U. Given the m-block tridiagonal form of A, of equation (6.2-18), we let S_i denote those components of U "corresponding" to D_i. The successive overrelaxation method requires that we solve for the components $U_j^{\nu+1}$ of the new iterate $U^{\nu+1}$ in some order.

We assume that A has property (A). We say that an order of solving the equations AU = B is consistent with this diagonally m-block tridiagonal representation of A if each component U_k of S_{i-1} is computed before any components of S_i with which U_k is coupled. An order of solving the equations is consistent if there is some diagonally m-block tridiagonal representation of A with which it is consistent.

These definitions will seem somewhat mysterious until one reads the proofs. Young's theory assumes that we are solving the system AU = B where A has property A and we are using a successive overrelaxation which is consistently ordered.

The first result of the theory is to show that the nonzero eigenvalues of the Jacobi iteration for A occur in pairs $\pm\lambda_i$. The next result is to show that the roots of the Jacobi method and the SOR method are related by equation (6.2-14). The selection of ω_{opt} then proceeds as above.

6.3 <u>The alternating-direction-implicit methods (ADI)</u>. In chapter 3 we discussed the use of the ADI technique for the solution of the parabolic heat equation. This technique can also be applied to elliptic equations. In fact, ADI is frequently one of the best methods for the solution of elliptic equations. We will introduce the ADI method for elliptic equations by relating it to the ADI for the heat equation which is described in chapter 3.

We will start with the heat equation in two dimensions, namely

$$\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} - h , \qquad w = w(x,y,t) \qquad (6.3\text{-}1)$$

$$h = h(x,y)$$

where (x,y) lies in domain R. The boundary condition is $w(x,y,t) = g(x,y)$ for (x,y) on ∂R. The initial condition is $w(x,y,0) = f(x,y)$. Since the data for this problem (h and g) are independent of the time, the solution w will approach the solution of the following elliptic problem for large time (independent of the initial function f); that is, $\lim_{t\to\infty} w(x,y,t) = u(x,y)$ where u satisfies the Poisson equation with Dirichlet boundary conditions.

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = h \qquad \text{if } (x,y) \in R \qquad (6.3\text{-}2)$$

$$u(x,y) = g(x,y) \qquad \text{if } (x,y) \in \partial R$$

Note that the function $v = w - u$ satisfies the equation

$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \qquad \text{on } R$$

$$v = 0 \qquad \text{on } \partial R$$

$$v(x,y,0) = f(x,y) \qquad \text{on } R, \text{ at } t = 0$$

Then it is possible to show that $\lim_{t\to\infty} v(x,y,t) = 0$, and in fact the solution will decay to zero at an exponential rate in t, that is $|v| \le ce^{-Kt}$ for large t. This suggests that we might be able to approximate the solution of the elliptic equation (6.3-2) by running a marching method for the time-dependent equation (6.3-1) out to a sufficiently large time.

We will next write down an ADI method for the time-dependent equation (6.3-1). We will assume the region R is a rectangle in order to simplify the description. We define the operators δ_x^2 and δ_y^2 as follows:

$$\delta_x^2 w = (w_{j+1,k} - 2w_{j,k} + w_{j-1,k})/\Delta x^2$$

$$\delta_y^2 w = (w_{j,k+1} - 2w_{j,k} + w_{j,k-1})/\Delta y^2$$

where j, k, $w_{j,k}$, Δx, and Δy have the usual meaning (see section 6.2.2). Then the ADI scheme for equation (6.3-1) is the following:

$$\begin{aligned} w_{j,k}^{n+\frac{1}{2}} &= w_{j,k}^{n} + \frac{\Delta t}{2}\delta_x^2 w_{j,k}^{n+\frac{1}{2}} + \frac{\Delta t}{2}\delta_y^2 w_{j,k}^{n} - \frac{\Delta t}{2} h_{j,k} \\ w_{j,k}^{n+1} &= w_{j,k}^{n+\frac{1}{2}} + \frac{\Delta t}{2}\delta_x^2 w_{j,k}^{n+\frac{1}{2}} + \frac{\Delta t}{2}\delta_y^2 w_{j,k}^{n+1} - \frac{\Delta t}{2} h_{j,k} \end{aligned} \qquad (6.3\text{-}3)$$

This is the Peaceman-Rachford version of ADI. It is somewhat different from the Douglas-Rachford version given in section 3.5. Note that the truncation error is $O(\Delta t + \Delta x^2 + \Delta y^2)$ rather than $O(\Delta t^2 + \Delta x^2 + \Delta y^2)$ as in section 3.5. In that section we were approximating time-dependent solutions whereas we are now interested in steady-state solutions, so we are no longer concerned about the $O(\Delta t)$ rather than $O(\Delta t^2)$ error term. What we now want is the fastest possible convergence to the steady-state solution.

Problem 6.3-1. Suppose we have found a steady-state solution of equation (6.3-3); that is, $w^n = w^{n+\frac{1}{2}} = w^{n+1} = u$. Show that u satisfies equation (5.3-3); that is, $Au = b$ where $A = I - \theta_x(C_x + C_x^T) - \theta_y(C_y + C_y^T)$ is defined by equation (6.2-1). Determine the right side b in terms of the boundary function g and the function h. Show that after multiplication by a suitable scaler, equations (6.3-3) can be written in matrix form as

$$\begin{aligned}(rI + H)w^{n+\frac{1}{2}} &= (rI - V)w^{n} + b\\ (rI + V)w^{n+1} &= (rI - H)w^{n+\frac{1}{2}} + b\end{aligned} \qquad (6.3\text{-}4)$$

where $A = H + V$. Determine H and V in terms of the matrices C_x and C_y of section 6.2.2. Show that H and V are symmetric and positive definite. Find a permutation of the elements of w so that V is tridiagonal (see section 3.5). Determine r in terms of Δx, Δy, and Δt.

Problem 6.3-2. The Douglas-Rachford [1955] method for the heat equation is (see Richtmyer and Morton, section 8.8)

$$\hat{w}^{n+1} = w^{n} + \Delta t\ \delta_x^2\ w^{n+1} + \Delta t\ \delta_y^2\ w^{n} - \Delta t h$$

$$w^{n+1} = w^{n} + \Delta t\ \delta_x^2\ \hat{w}^{n+1} + \Delta t\ \delta_y^2\ w^{n+1} - \Delta t h$$

Write this scheme in matrix form in terms of matrices H and V similar to equation (6.3-4).

We can now see that our derivation based on the ADI scheme for the heat equation has led to an iterative scheme for the solution of $Au = b$ where $A = H + V$, namely equations (6.3-4). This is an ADI method for the elliptic equation (6.3-2). The solution of equations (6.3-4) for w^{n+1} requires the inversion of a tridiagonal matrix on each "sweep" through the mesh; first a sweep involving the horizontal mesh lines and the inversion of $rI + H$ and then a sweep involving the vertical lines and inversion of $rI + V$. Note that these matrices are nonsingular if $r > 0$ even for Neumann boundary conditions (why?). The convergence rate of the ADI method is greatly improved when we let the iteration parameter r change with n, that is replace r by r_n. However, we will first analyze the case when r is a constant.

6.3.1. <u>ADI with a single iteration parameter</u>. The iteration scheme of equation (6.3-4) can be written $w^{n+1} = M_r w^n + c$ where

$$M_r = (rI + V)^{-1} (rI - H) (rI + H)^{-1} (rI - V)$$

$$c = (rI + V)^{-1} b + (rI + V)^{-1} (rI - H) (rI + H)^{-1} b \tag{6.3-5}$$

To prove that this scheme is convergent, we must show that the spectral radius $\sigma(M_r)$ satisfies $\sigma(M_r) < 1$.

Problem 6.3-3. If H is symmetric and positive definite, and $r > 0$, then show that the eigenvalues λ of $C = (rI - H)(rI + H)^{-1}$ satisfy $|\lambda| < 1$. Show that $\|C\|_2 < 1$ where $\|C\|_2$ is the matrix norm induced by the usual Euclidean vector norm. Show that $\|M_r\|_2 < 1$ if H and V are symmetric and positive definite and $r > 0$. (M_r is defined in equation (6.3-5)).
Hint: Use the fact that M_r is similar to the matrix $(rI-H)(rI+H)^{-1}(rI-V)(rI+V)^{-1}$. Would your proof work if the original differential equation (6.3-2) had Neumann rather than Dirichlet boundary conditions?

We will next compute the convergence rate for the single iteration parameter ADI method. This will apply to the case discussed in section 6.2.2, namely Laplace's equation with Dirichlet boundary conditions on a rectangle, except we will require an equally spaced mesh on a square; that is, $\Delta x = \Delta y$ and $J = K$. We do this for simplicity. We will assume that equations (6.3-4) are written in the form given in problem 6.3-1, namely $Au = b$ where $A = H + V$

with $H = \theta_x(-C_x - C_x^T + 2I)$ and $V = \theta_y(-C_y - C_y^T + 2I)$ where C_x and C_y are the matrices defined in section 6.2.2.

Problem 6.3-4. Verify that the eigenvectors of H and V are given by

$$u_{j,k}^{(p,q)} = \sin\left(\frac{jp\pi}{J+1}\right) \sin\left(\frac{kq\pi}{J+1}\right) \qquad \text{for } 1 \le j,k,p,q \le J = K$$

Show that the eigenvalues are $\lambda_s = \sin^2\left(\frac{\pi s}{2(J+1)}\right)$, $1 \le s \le J$

Problem 6.3-5. Show that the spectral radius of the iteration matrix M_r of equation (6.3-5) is

$$\sigma(M_r) = \max_{1 \le p \le J} \left(\frac{r-\lambda_p}{r+\lambda_p}\right)^2, \qquad \lambda_p = \sin^2\left(\frac{\pi p}{2(J+1)}\right)$$

Hint: If u is an eigenvector of both H and V, then u is also an eigenvector of HV, $H^{-1}V$, VH, and $V^{-1}H$.

Problem 6.3-6. If $f(r,x) = (r-x)/(r+x)$, then show that

$$\min_{0 \le r} \left[\max_{0<\alpha \le x \le \beta} |f(r,x)| \right] = \frac{1 - \sqrt{\alpha/\beta}}{1 + \sqrt{\alpha/\beta}}$$

and the minimum is assumed for $r = \sqrt{\alpha\beta}$. Hint: $f(r,x)$ is a monotone function for fixed r, hence it assumes its maximum (and minimum) at $x = \alpha$ and $x = \beta$.

Problem 6.3-7. Using equation (6.2-2) show that the spectral radius of the Jacobi iteration is $\cos\left(\pi/(J+1)\right)$ in case $J = K$ and $\Delta x = \Delta y$. Using equations (6.2-10) and (6.2-11) show that the spectral radius for SOR is $\frac{\left(1 - \sqrt{1 - \rho^2}\right)^2}{\rho^2}$ where $\rho = \cos\left(\pi/(J+1)\right)$. Show that the eigenvalues λ_p of problem 6.3-5 satisfy $\sin^2(\theta/2) \le \lambda_p \le \cos^2(\theta/2)$ where $\theta = \pi/(J+1)$ and therefore if we choose r according to problem 6.3-6, namely $r = \sin(\theta/2)\cos(\theta/2)$, then the spectral radius $\sigma(M_r)$ is given by

$$\sigma(M_r) = \left(\frac{1 - \tan(\theta/2)}{1 + \tan(\theta/2)}\right)^2 = \left(\frac{1 + \cos\theta - \sin\theta}{1 + \cos\theta + \sin\theta}\right)^2 = \frac{\left(1 - \sqrt{1 - \rho^2}\right)^2}{\rho^2}$$

where $\rho = \cos\theta$.

This result shows that optimized single parameter ADI and optimized successive overrelaxation have the same asymptotic convergence rate. Therefore to make any improvement over SOR for this problem, we must use more than one iteration parameter in the ADI method.

6.3.2 Convergence for the multiparameter ADI - the model problem. We can write the ADI iteration in the form of equations (6.3-5), namely

$$w^{n+1} = M(r)w^n + c(r)$$

The iteration matrix and the vector c are both functions of r. The multiparameter ADI uses a sequence of parameters $r_1,\dots,r_m$ and repeats the use of the sequence so that we have a cyclic process with period m. Starting with w^k we compute $w^{k+1},\dots,w^{k+m}$ as follows:

$$\begin{aligned} w^{k+1} &= M(r_1)w^k + c(r_1) \\ w^{k+2} &= M(r_2)w^{k+1} + c(r_2) \\ &\vdots \\ w^{k+m} &= M(r_m)w^{k+m-1} + c(r_m) \end{aligned} \qquad (6.3\text{-}6)$$

We then repeat the cycle, thus $w^{k+m+1} = M(r_1)w^{k+m} + c(r_1)$. We can write this procedure in the form $w^{k+m} = M_m w^k + c_m$ where

$$M_m = M(r_1)\ M(r_2)\ \ldots\ M(r_m). \tag{6.3-7}$$

The convergence of the multiparameter ADI is thus governed by the spectral radius of M_m.

Problem 6.3-8. Assume that H and V are symmetric non-negative definite matrices with at least one of them positive definite. Assume each r_j $1 \le j \le m$ is positive. Then show the multiparameter iteration is convergent, that is the spectral radius satisfies $\sigma(M_m) < 1$.

In order to obtain an optimal sequence of iteration parameters r_j we must make an additional assumption concerning the matrices H and V. We assume that H and V have a common set of eigenvectors. As before, we also assume that H and V are symmetric non-negative definite with at least one of them positive definite. If the system $Au = b$ $(A = H + V)$ satisfies these conditions, then it is called a model problem. For a model problem we have $Hu^{(k)} = \xi_k u^{(k)}$ and $Vu^{(k)} = \mu_k u^{(k)}$ where $u^{(k)}$ is the common set of eigenvectors. Note that the set $\{u^{(k)}\}$ forms an orthogonal basis. Also note that the problem of section 5.3 (and 6.2.2), namely Laplace's equation with Dirichlet boundary conditions on a rectangle, is a model problem.

The condition that H and V have a common set of eigenvectors is equivalent to the requirement that H and V commute; that is, $HV = VH$. We will refer the reader to Varga [1962, p. 221] for a proof. We offer the easy half of this proof as a problem.

Problem 6.3-9. Assume the matrices A and B (of order N) have a common set of eigenvectors which form a basis; that is, $Ax^{(i)} = \lambda_i x^{(i)}$, $Bx^{(i)} = \mu_i x^{(i)}$, $1 \le i \le N$ and the $x^{(i)}$ are independent. Show that AB = BA.

For the model problem we can compute the eigenvalues of the iteration matrix M_m of equation (6.3-7). Note that the common set of eigenvectors of H and V is also the set of eigenvectors for M_m.

Problem 6.3-10. If ξ_k and μ_k are the eigenvalues of H and V, then the eigenvalues of M_m are given by

$$\lambda_k = \prod_{i=1}^{m} \frac{(r_i - \xi_k)}{(r_i + \xi_k)} \frac{(r_i - \mu_k)}{(r_i + \mu_k)} \qquad 1 \le k \le N \qquad (6.3\text{-}8)$$

6.3.3 <u>Determination of a near optimal iteration parameter.</u> In order to minimize the spectral radius of M_m we must choose the r_i to minimize the above rational function of r_i. The book by Wachspress [1966, p. 178] deals with this question in considerable detail. Usually we will not know the eigenvalues $\{\xi_k, \mu_k\}$, but we may be able to determine an interval in which they lie. Suppose we have $0 \le a \le \xi_k \le b$, $0 \le c \le \mu_k \le d$, $a+c > 0$. We do not know enough about ξ_k and μ_k to minimize the spectral radius $\sigma(M_m)$ but we can choose r_i to minimize the following function

$$\min_{\{r_i\}} \max_{\substack{a\le\xi\le b \\ c\le\mu\le d}} \prod_{i=1}^{m} \frac{|(r_i-\xi)\ (r_i-\mu)|}{(r_i+\xi)\ (r_i+\mu)} = \min_{\{r_i\}} \max_{\substack{a\le\xi\le b \\ c\le\mu\le d}} |G(r_1,\ldots r_m,\xi,\mu)| \qquad (6.3\text{-}9)$$

It is possible to prove the existence of a unique minimizing set $\{r_i\}$ along with a method to construct the set $\{r_i\}$. However, the method is very complicated [Wachspress, 1966]. In the case m is a power of 2, $m = 2^t$, there is a simple method to compute the optimum set $\{r_i\}$ [Wachspress 1966, p. 194] or [Varga, 1962, p. 223]. We will not describe this, but will instead give a procedure which appeared in the original paper of Peaceman and Rachford [1955]. This produces a good, but not necessarily optimal set $\{r_i\}$ for any m.

We will proceed to determine the set $\{r_i\}$. We assume that $0 < \alpha \le \min(a,c)$ and $\beta \ge \max(b,d)$. We define

$$F_i(z) = \frac{r_i - z}{r_i + z}, \qquad F(z) = \prod_{i=1}^{m} F_i(z) .$$

Then the discussion preceding equation (6.3-9) shows that the spectral radius $\sigma(M_m)$ is bounded by

$$\sigma(M_m) \le \max_{\alpha \le z \le \beta} F^2(z)$$

Problem 6.3-11. Determine a sequence $\{\alpha_i\}$ such that $\alpha = \alpha_0 < \alpha_1 < \dots < \alpha_m = \beta$ and α_{i-1}/α_i is a constant independent of i. Show that

$$\max_{\alpha_{i-1} \le z \le \alpha_i} F(z) \le \max_{\alpha_{i-1} \le z \le \alpha_i} F_i(z)$$

Use problem 6.3-6 to show that

$$\max_{\alpha_{i-1} \le z \le \alpha_i} F_i(z) \le \frac{\gamma-1}{\gamma+1} \qquad \text{where } \gamma = \left(\frac{\beta}{\alpha}\right)^{1/2m}$$

if r_i is set equal to $\sqrt{\alpha_{i-1}\,\alpha_i}$. Note that this value of r_i will minimize the function $\max_{\alpha_{i-1} \le z \le \alpha_i} F_i(z)$, regarding the latter as a function of r_i.

Then show that

$$\sigma(M_m) \leq \left(\frac{\gamma-1}{\gamma+1}\right)^2$$

Show that the average convergence rate $R_m = -\frac{\ln \sigma(M_m)}{m}$ satisfies

$$R_m \geq \frac{4 \ln\gamma}{\ln(\beta/\alpha)} \ln\left(\frac{\gamma+1}{\gamma-1}\right) \qquad (6.3\text{-}10)$$

Note that $\gamma = (\beta/\sigma)^{1/2m}$.

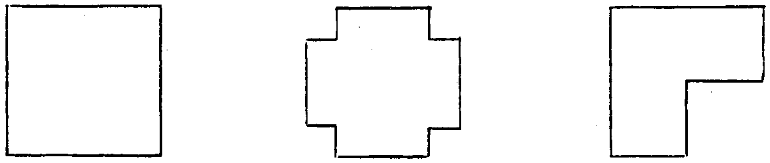
Problem 6.3-12. For the problem described in section 5.3 (also 6.2.2) we may use $\alpha = \sin^2(\pi/2N)$ and $\beta = \cos^2(\pi/2N)$ where $N = J+1 = K+1$. Show that for large N the convergence rate of ADI satisfies $R \cong 0\left(\frac{1}{N^{1/m}}\right)$.

For optimal SOR the convergence rate satisfies $R_m \cong 0\left(\frac{1}{N}\right)$. Thus if $m > 1$ we can obtain a large improvement through the use of ADI. For Laplace's equation on a square with Dirichlet boundary conditions, we can choose m so that the convergence rate satisfies the inequality below for sufficiently large N [Varga, 1962, p. 227].

$$R_m > \frac{3.107}{1.386 + 2 \ln(N/\pi)}$$

This again shows the convergence rate of ADI to be superior to that for SOR.

6.3.4 Comparison of SOR and ADI. In this section we will describe some numerical experiments intended to illustrate the previous discussion. We have taken these results from the paper of Birkhoff, Varga and Young [1962]. The experiments apply to Laplace's equation $\nabla^2 u = 0$ with Dirichlet boundary conditions, $u = 0$ on the boundary. The solution is therefore $U = 0$ at all mesh points. Three subsets of the unit square were used; the unit square itself (I), the unit square with the four corner squares of side length 1/5 removed (II), and the unit square with the lower right corner square of side length 1/2 removed (III). The figures below show these three regions.



I II III

Figure 6.3-1.

The initial guess for the solution U was taken to be unity at all mesh points. Then the iterative scheme was used until the value of U at each mesh point was less than 10^{-6}. In this case the value of U is equal to the error since the solution is $U \equiv 0$. The number of iterations required to achieve this error level is denoted by N^{α} where α denotes the case. When $\alpha = T$, the ADI iteration was used with the method of Wachspress used to compute the iteration parameters $r_1, \ldots, r_m$ [Wachspress, 1966, p. 194]. This produces a true minimum in equation (6.3-9) but requires that $m = 2^t$. An approximate value for N^T can be computed from the bound for the spectral radius given in equation (6.3-9) at least for case I. This bound was evaluated numerically to four-digit accuracy. In the tables below we refer to this as the calculated N_c^T. When $\alpha = A$, the approximate method of problem 6.3-11 was used to compute the values $\{r_1, \ldots, r_m\}$. Equation (6.3-10) was used to approximate the number of iterations required to reduce the error to 10^{-6}.

This number is denoted by N_c^A. The SOR method was also applied to these problems; the observed number of iterations is denoted by N^S. A predicted number of iterations for SOR in case I was obtained by solving $4N(\omega-1)^{N-1} = 10^{-6}$. It is denoted by N_c^S. Here ω is the optimum iteration parameter. The factor 4N appears because of the eigenvector deficiency in the iteration matrix for optimum SOR (see section 6.2.3); the Jordan canonical form of the iteration matrix is not diagonal. In section 6.3.1 we showed that ADI with $m = 1$ and SOR have the same asymptotic convergence rates. The values of $N_c^T(m = 1)$ and N_c^S are different because of this eigenvector deficiency. The results of Birkhoff, Varga, and Young are listed below. The parameter h denotes the mesh spacing, $\Delta x = \Delta y = h$. In comparing ADI and SOR one must remember that a single ADI iteration requires slightly more than twice the computer arithmetic than a single SOR sweep. Thus if $N^S = 2N^T$, then the computing time should be about equal.

Case I

	ADI						SOR	
	Optimal r_i			Approximate r_i				
Mesh h^{-1}	Obs m=4 N^T	Cal m=4 N_c^T	Obs m=1 N^T	Obs m=4 N^A	Cal m=4 N_c^A	Obs m=1 N^A	Obs N^S	Cal N_c^S
10	11	9	23	9	11	23	28	32
20	12	12	46	11	15	46	53	66
40	18	15	91	15	20	91	117	136
80	20	19	183	21	25	183	236	292

	Case II			Case III		
	ADI Approximate r_i		SOR	ADI Approximate r_i		SOR
Mesh h^{-1}	Obs m=4 N^A	Obs m=1 N^A	Obs N^S	Obs m=4 N^A	Obs m=1 N^A	Obs N^S
10	16	19	26	13	17	20
20	20	36	51	19	37	41
40	23	75	108	22	75	85
80	25	150	---	27	162	---

Additional comparisons of methods by experimental computations can be found in Wachspress [1966, chapter 8], Birkhoff, Varga and Young [1962], Price and Varga [1962], Young and Ehrlich [1960].

REFERENCES

G. Birkhoff, R. Varga, and D. Young (1962), "Alternating Direction Implicit Methods," in "Advances in Computers," Vol. 3, F. Alt and M. Rubinoff (ed), Academic Press, New York, pp. 190-275.

S. Burstein (1965), "Finite Difference Calculations for Hydrodynamic Flows Containing Discontinuities," Report NYO 1480-33, Courant Institute, New York University.

B. Carré (1961), "The Determination of the Optimum Accelerating Factor for Successive Over-relaxation," Computer J., 4, pp. 73-78.

J. Douglas and H. Rachford (1956), "On the Numerical Solution of Heat Conduction Problems in Two or Three Space Variables," Trans. Amer. Math. Soc., 82, pp. 421-439.

G. Forsythe and J. Ortega (1960), "Attempts to Determine the Optimum Factor for Successive Overrelaxation," Proc. Int. Conf. Infor. Processing, Unesco, Paris, p. 110.

__________ and W. Wasow (1960), "Finite Difference Methods for Partial Differential Equations," Wiley, New York.

S. Frankel (1950), "Convergence Rates of Iterative Treatments of Partial Differential Equations," Math Tables Aids Comput, 4, pp. 65-75.

P. Garabedian (1956), "Estimation of the Relaxation Factor for Small Mesh Size," Math Tables Aids Comput., 10, pp. 183-185.

J. Gary (1964), "On Certain Finite Difference Schemes for Hyperbolic Systems," Math Comput., 18, pp. 1-18.

_______ (1966), "A Generalization of the Lax-Richtmyer Theorem on Finite Difference Schemes," J. SIAM Numer. Anal., 3, pp. 467-473.

Gourlay and Mitchell (1966), "ADI Methods for Hyperbolic Systems," Numer. Math, 8, p. 137.

D. Greenspan (1965), "Introductory Numerical Analysis of Elliptic Boundary Value Problems," Harper and Row, New York.

L. Hageman and R. Kellogg (1968), "Estimating Optimum Overrelaxation Parameters," Math Comput., 22, pp. 60-69.

P. Henrici (1962), "Discrete Variable Methods in Ordinary Differential Equations," Wiley, New York.

E. Isaacson and H. Keller (1966), "Analysis of Numerical Methods," Wiley, New York.

REFERENCES (cont'd)

W. Kahan (1958), "Gauss-Seidel Methods of Solving Large Systems of Linear Equations," Doctoral Thesis, Univ. of Toronto.

H. Keller (1958), "On some Iterative Methods for Solving Elliptic Difference Equations," Quart, Appl. Math.,16, pp. 209-226.

H. Kreiss (1964), "On Difference Approximations of the Dissipative Type for Hyperbolic Differential Equations," Comm. Pure Appl. Math., 17, p. 335.

H. Kulsrud (1961), "A Practical Technique for the Determination of the Optimum Relaxation Factor of the Successive Over-relaxation Method," Comm. Assoc. Comput. Mach., 4, pp. 184-187.

A. Ostrowski (1954), "On the Linear Iteration Procedures for Symmetric Matrices," Rend Mat e Appl., 14, pp. 140-163.

D. Peaceman and H. Rachford (1955), "The Numerical Solution of Parabolic and Elliptic Differential Equations," J. Soc. Indust. Appl. Math., 3, pp. 28-41.

H. Price and R. Varga (1962), "Recent Numerical Experiments Comparing Successive Overrelaxation Iterative Methods with Implicit Alternating Direction Methods," Gulf Research and Development Co. Report, Harmanville, Pa.

G. Shortley and R. Weller (1938), "The Numerical Solution of Laplace's Equation," J. Appl. Phys., 9, pp. 334-348.

J. Spanier (1967), "Alternating Direction Methods Applied to Heat Conduction Problems," in "Mathematical Methods for Digital Computers," Vol. 2, A. Ralston and H. Wilf (ed), Wiley, New York.

H. Stetter (1965), "Stability and Convergence of Nonlinear Discretization Algorithms," Symposium on the Numerical Solution of Partial Differential Equations, Univ. of Maryland, May 1964.

R. Varga (1962), "Matrix Iterative Analysis," Prentice Hall, Englewood Cliffs, New Jersey.

E. Wachspress (1966), "Iterative Solution of Elliptic Systems," Prentice Hall, Englewood Cliffs, New Jersey.

J. Wilkinson (1963), "Rounding Errors in Algebraic Processes," Prentice Hall, Englewood Cliffs, New Jersey.

REFERENCES (cont'd)

D. Young (1954), "Iterative Methods for Solving Partial Difference Equations of Elliptic Type," Trans. Amer. Math Soc., 76, pp. 92-111 (also Harvard Doctoral Thesis in 1950).

________ and L. Ehrlich (1960), "Some Numerical Studies of Iterative Methods for Solving Elliptic Difference Equations," in "Boundary Problems in Differential Equations," U. of Wisconsin Press, Madison, pp. 143-162.