K[redacted]M X-52239

NUMERICAL TECHNIQUES FOR THE SOLUTION OF SYMMETRIC POSITIVE LINEAR DIFFERENTIAL EQUATIONS

A Thesis Submitted to

Case Institute of Technology

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

by

Theodore Katsanis

June 1967

Thesis Advisor: Professor Milton Lees

18743

ABSTRACT

A finite difference method for the solution of symmetric positive linear differential equations is developed. The method is applicable to any region with piecewise smooth boundaries. Methods for solution of the finite difference equations are discussed. The finite difference solutions are shown to converge at essentially the rate $O(h^{1/2})$ as $h \to 0$, h being the maximum distance between adjacent mesh points.

An alternate finite difference method is given with the advantage that the finite difference equations can be solved iteratively. However, there are strong limitations on the mesh arrangements which can be used with this method.

The Tricomi equation can be expressed in symmetric positive form. Admissible boundary conditions for any region with piecewise smooth boundaries are given, with a wide choice of boundary conditions being possible.

A Tricomi equation with a known analytical solution is solved numerically as an illustration of the numerical results which can be obtained. There is strong convergence to the analytical solutions, but pointwise divergence. Smoothing of the solution reduces this, though, and satisfactory numerical results are obtained.

Author

ACKNOWLEDGMENTS

I would like to express my appreciation to Professor Milton Lees for his guidance and constructive criticism, and for his encouragement.

I also wish to thank Lewis Research Center of the National Aeronautics and Space Administration for direct support through its graduate study program.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

In the theory of partial differential equations there is a fundamental distinction between those of elliptic, hyperbolic and parabolic type. Generally each type of equation has different requirements as to the boundary or initial data which must be specified to assure existence and uniqueness of solutions, and to be well posed. These requirements are usually well-known for an equation of any particular type. Further, many analytical and numerical techniques have been developed for solving the various types of partial differential equations, subject to the proper boundary conditions, including even many nonlinear cases. However, for equations of mixed type much less is known, and it is usually difficult to know even what the proper boundary conditions are.

As a step toward overcoming this problem Friedrichs [1] has developed a theory of symmetric positive linear differential equations independent of type. Chu [2] has shown that this theory can be used to derive finite difference solutions in two-dimensions for rectangular regions, or more generally, by means of a transformation, for regions with four corners joined by smooth curves. In this paper a more general finite difference method for the solution of symmetric positive equations is presented. The only restriction on the shape of the region is that the boundary be piecewise smooth. It is proven that the finite difference solution converges to the solution of the differential equation at essentially the rate $O(h^{1/2})$ as $h \to 0$, h being the maximum distance between adjacent mesh points for a two-dimensional region. Also weak convergence to weak solutions is shown.

An alternate finite difference method is given for the two-dimensional case with the advantage that the finite difference equation can be solved iteratively. However, there are strong limitations on the mesh arrangements which can be used with this method.

As an example of the potential usefulness of the theory of symmetric positive equations, the Tricomi equation

$$y\varphi_{xx} - \varphi_{yy} = f(x,y)$$

can be expressed in symmetric positive form. It is shown that suitable boundary conditions can always be determined, regardless of the shape of the region. The problem in a practical case is to determine an "admissible" boundary condition which corresponds to available boundary information.

As an illustration of numerical results which can be obtained by the proposed finite difference scheme, a Tricomi equation with a known analytical solution is solved numerically. The results indicate that, although there is strong (i.e., L^2) convergence of the finite difference solution to the analytical solution, there is pointwise divergence along the boundary. However, smoothing the solution can eliminate this problem, and satisfactory numerical results are obtained, although rigorous mathematical justification of the smoothing process is not given.

CHAPTER I

SYMMETRIC POSITIVE LINEAR DIFFERENTIAL EQUATIONS

1.1 Basic Definitions

Let Ω be a bounded open set in the m-dimensional space of real numbers, R^m. The boundary of Ω will be denoted by $\partial\Omega$, and its closure by $\bar{\Omega}$. It is assumed that $\partial\Omega$ is piecewise smooth. A point in R^m is denoted by $x = (x_1, x_2, \ldots, x_m)$ and an r-dimensional vector valued function defined on Ω is given by $u = (u_1, u_2, \ldots, u_r)$. Also let $\alpha^1, \alpha^2, \ldots, \alpha^m$ and G be given $r \times r$ matrix-valued functions and $f = (f_1, f_2, \ldots, f_r)$ a given r dimensional vector-valued function, all defined on Ω (at least). It is assumed that the α^i are piecewise differentiable. For convenience, let $\alpha = (\alpha^1, \alpha^2, \ldots, \alpha^m)$, so that we can use expressions such as

$$\nabla \cdot (\alpha u) = \sum_{i=1}^{m} \frac{\partial}{\partial x_i} (\alpha^i u) \tag{1.1}$$

With this notation we can write the identity

$$\sum_{i=1}^{m} \frac{\partial}{\partial x_i} (\alpha^i u) = \sum_{i=1}^{m} \frac{\partial \alpha^i}{\partial x_i} u + \sum_{i=1}^{m} \alpha^i \frac{\partial u}{\partial x_i}$$

simply as

$$\nabla \cdot (\alpha u) = (\nabla \cdot \alpha)\, u + \alpha \cdot \nabla u \tag{1.2}$$

With this we can give the definitions for symmetric positive operators and admissible or semi-admissible boundary conditions which were introduced by Friedrichs [1].

Let K be the first order linear partial differential operator defined by

$$Ku = \alpha \cdot \nabla u + \nabla \cdot (\alpha u) + Gu \tag{1.3}$$

K is <u>symmetric positive</u> if each component, α^i, of α is symmetric and the symmetric part, $(G + G^*)/2$, of G is positive definite on on $\bar{\Omega}$.

For the purpose of giving suitable boundary conditions, a matrix, β, is defined (a.e.) on $\partial\Omega$ by

$$\beta = n \cdot \alpha \tag{1.4}$$

where $n = (n_1, n_2, \ldots, n_m)$ is defined to be the outer normal on $\partial\Omega$.

The boundary condition $Mu = 0$ on $\partial\Omega$ is <u>semi-admissible</u> if $M = \mu - \beta$, where μ is any matrix with non-negative definite symmetric part, $(\mu + \mu^*)/2$. If in addition, $\mathfrak{N}(\mu - \beta) \oplus \mathfrak{N}(\mu + \beta) = R^r$ on the boundary, $\partial\Omega$, the boundary condition is termed <u>admissible</u>. ($\mathfrak{N}(\mu - \beta)$ is the null space of the matrix $(\mu - \beta)$.)

The problem is to find a function u which satisfies

$$\left.\begin{aligned} Ku &= f \quad \text{on } \Omega \\ Mu &= 0 \quad \text{on } \partial\Omega \end{aligned}\right\} \tag{1.5}$$

where K is symmetric positive.

It turns out that many of the usual partial differential equations may be expressed in this symmetric positive form, with the standard boundary conditions also expressed as an admissible boundary condition. This includes equations of both hyperbolic and elliptic type. However, the greatest interest lies in the fact that the definitions are completely independent of type. An example of potentially great practical importance is the Tricomi equation which arises from the equations for transonic fluid flow. The Tricomi equation is of mixed type, i.e., it is hyperbolic in part of the region, elliptic in part, and is parabolic along the line between the two parts.

The significance of the semi-admissible boundary condition is that this insures the uniqueness of a classical solution to a symmetric positive equation. On the other hand, the stronger, admissible boundary condition is required for existence. The existence of a classical solution is generally difficult to prove for any particular case, and depends on properties at corners of the region. However, it is very easy to prove existence (but not uniqueness!) of weak solutions with only semi-admissible boundary conditions.

1.2 Basic Identities and Inequalities

Let $\mathcal{H}$ be the Hilbert space of all square integrable r-dimensional vector-valued functions defined on Ω. The inner product is given by

$$(u,v) = \int_\Omega u \cdot v \tag{1.6}$$

where

$$u \cdot v = \sum_{i=1}^{r} u_i v_i$$

and

$$\|u\|^2 = (u,u) \tag{1.7}$$

A boundary inner product is defined by

$$(u,v)_B = \int_{\partial\Omega} u \cdot v \tag{1.8}$$

with the corresponding norm

$$\|u\|_B^2 = (u,u)_B \tag{1.9}$$

We introduce now the adjoint operators K^* and M^*, which are defined by

$$K^*u = -\alpha \cdot \nabla u - \nabla \cdot (\alpha u) + G^*u \tag{1.10}$$

$$M^*u = (\mu^* + \beta)u \tag{1.11}$$

The relation between K and M and their adjoints is given by Friedrichs "first identity."

Lemma 1.1 If K is symmetric positive, then

$$(v,Ku) + (v,Mu)_B = (K^*v,u) + (M^*v,u)_B \tag{1.12}$$

Proof - The proof follows from Green's Theorem. By definition we have

$$(v,Ku) - (K^*v,u) = \int_\Omega v \cdot (\alpha \cdot \nabla u) + v \cdot (\nabla \cdot (\alpha u)) + v \cdot Gu$$

$$+ \int_\Omega (\alpha \cdot \nabla v) \cdot u + (\nabla \cdot (\alpha v)) \cdot u - G^*v \cdot u$$

$$= \int_\Omega \sum_{i=1}^{m} \left\{ v \cdot \left(\alpha^i \frac{\partial u}{\partial x_i}\right) + v \cdot \frac{\partial(\alpha^i u)}{\partial x_i} + \left(\alpha^i \frac{\partial v}{\partial x_i}\right) \cdot u + \frac{\partial(\alpha^i v)}{\partial x_i} \cdot u \right\}$$

$$= 2 \int_\Omega \sum_{i=1}^{m} \frac{\partial}{\partial x_i} (v \cdot \alpha^i u)$$

since the α^i are symmetric. Therefore

$$(v,Ku) - (K^*v,u) = 2 \int_{\partial\Omega} n \cdot (v \cdot \alpha u) = 2 \int_{\partial\Omega} v \cdot \beta u$$

by Green's Theorem, and since $\beta = n \cdot \alpha$. Finally

$$(v,Ku) - (K^*v,u) = \int_{\partial\Omega} (\mu^*v \cdot u + \beta v \cdot u - v \cdot \mu u + v \cdot \beta u)$$

$$= (M^*v,u)_B - (v,Mu)_B$$

which proves the lemma.

The "first identity" can now be used to obtain Friedrichs "second identity."

<u>Lemma 1.2</u> If K is symmetric positive, then

$$(u,Ku) + (u,Mu)_B = (u,Gu) + (u,\mu u)_B \qquad (1.13)$$

<u>Proof</u> - It follows from the definitions of K^* and M^* that $K + K^* = G + G^*$ and $M + M^* = \mu + \mu^*$; hence, letting $v = u$ in the "first identity," we obtain

$$(u,Ku) + (u,Mu)_B = \frac{1}{2}\left[(u,(K + K^*)u) + (u,(M + M^*)u)_B\right]$$
$$= \left(u, \frac{G + G^*}{2} u\right) + \left(u, \frac{\mu + \mu^*}{2} u\right)_B$$
$$= (u,Gu) + (u,\mu u)_B$$

The "second identity" immediately yields an inequality which will give us an a priori bound and insure uniqueness of any classical solution to a symmetric positive equation with semi-admissible boundary conditions.

Lemma 1.3 Suppose u is a solution to (1.5) where M is semi-admissible. Let λ_G be the smallest eigenvalue of $(G + G^*)/2$ in $\bar{\Omega}$. Then

$$\|u\| \le \frac{1}{\lambda_G} \|f\| \tag{1.14}$$

Proof - Since K is symmetric positive, $\lambda_G > 0$, and therefore $\|u\|^2 \le (u,Gu)/\lambda_G$. Using Lemma 1.2, since $\mu + \mu^*$ is non-negative definite by the assumption of the semi-admissible boundary condition, we have

$$\|u\|^2 \le \frac{1}{\lambda_G}\left[(u,Gu) + (u,\mu u)_B\right] = \frac{1}{\lambda_G}(u,Ku),$$

since $Mu = 0$, so that

$$\|u\|^2 \le \frac{1}{\lambda_G}\|u\|\ \|Ku\| = \frac{1}{\lambda_G}\|u\|\ \|f\|$$

One other inequality can be obtained for $\|u\|_B$ by assuming that $\mu + \mu^*$ is positive definite.

Lemma 1.4 Let u satisfy equation (1.5) where M is semi-admissible. Further, assume that $(\mu + \mu^*)/2$ is positive definite on $\partial\Omega$ with smallest eigenvalue λ_μ. Then

$$\|u\|_B = \frac{1}{\sqrt{\lambda_G\lambda_\mu}} \|f\| \tag{1.15}$$

Proof - From the hypothesis,

$$\|u\|_B^2 \leq \frac{1}{\lambda_\mu} (u,\mu u) \leq \frac{1}{\lambda_\mu} \left[(u,\mu u) + (u,Gu)\right] = \frac{1}{\lambda_\mu} (u,Ku)$$

$$\leq \frac{1}{\lambda_\mu} \|u\| \ \|Ku\| \leq \frac{1}{\lambda_\mu\lambda_G} \|f\|^2$$

by Lemma 1.3.

1.3 Uniqueness of a C_1 Solution

Lemma 1.3 insures the uniqueness of a classical solution, and also that it is well posed in L^2 for homogeneous boundary conditions.

Theorem 1.1 If $u \in C_1(\Omega)$ satisfies equation (1.5) where M is semi-admissible, then u is the unique solution to (1.5). Further (1.5) is well posed in the sense that for any $\epsilon > 0$ there exists a $\delta > 0$ such that if f is replaced by f_ϵ in (1.5) with $\|f_\epsilon - f\| < \delta$, and if a solution u_ϵ still exists, then $\|u_\epsilon - u\| < \epsilon$.

Proof - Suppose that $v \in C_1(\Omega)$ is any solution of (1.5), then $K(u - v) = 0$, $M(u - v) = 0$ is semi-admissible and by Lemma 1.3, $\|u - v\| = 0$. For the second part let $\delta = \lambda_G \epsilon$, then

$$K(u_\epsilon - u) = f_\epsilon - f, \ M(u_\epsilon - u) = 0,$$

hence

$$\|u_\epsilon - u\| \leq \frac{1}{\lambda_G} \|f_\epsilon - f\| < \epsilon$$

Actually piecewise differentiability of u is adequate for the above theorem provided u is continuous. This follows easily since, when Green's theorem is applied, the values of u along the discontinuities of the derivative will cancel, providing us with all the previous results.

1.4 Weak and Strong Solutions

By widening the class of solutions to (1.5) to include weak solutions it is quite easy to prove existence of a solution to a symmetric positive equation under only semi-admissible boundary conditions. We will use Friedrichs' definition of weak solution. Let $V = C_1(\Omega) \cap \{v | M^*v = 0 \text{ on } \partial\Omega\}$. A function $u \in \mathcal{H}$ (defined in section 1.2) is a weak solution of (1.5) if $f \in \mathcal{H}$ and for all $v \in V$

$$(v,f) = (K^*v,u) \tag{1.16}$$

It follows from the "first identity" (1.12) that a classical solution is also a weak solution.

Theorem 1.2 If M is semi-admissible, there exists a weak solution to (1.5).

Proof - Let $\mathscr{W}$ be the subspace of all functions w, where $w = K^*v$ with $v \in V$. Since K^* is symmetric positive and M^* is semi-admissible, Theorem 1.1 implies that v is unique for any given w. Hence, for any fixed $f \in \mathcal{H}$, we can define a linear functional L_f, defined on $\mathscr{W} \subset \mathcal{H}$ by

$$L_f(w) = (v,f).$$

This linear functional is bounded, since

$$|(v,f)| \leq \|v\| \, \|f\| \leq \frac{1}{\lambda_G} \|f\| \, \|w\|$$

by Lemma 1.3 applied to K^* and M^*. By the Hahn-Banach theorem

L_f can be extended to all of $\mathcal{H}$, and by the Riesz representation theorem there is a $u \in \mathcal{H}$ such that

$$(v,f) = (w,u)$$

which proves the theorem.

This only shows that $u \in \mathcal{H}$, however, if $u \in C_1(\Omega)$, we see from Lemma 1.1 that

$$\begin{aligned}(v,Ku) + (v,Mu)_B &= (K^*v,u) + (M^*v,u)\\ &= (v,f) \text{ for all } v \in V.\end{aligned}$$

Hence $(v,Ku - f) = 0$ if $v = 0$ on $\partial\Omega$, so that we must have $Ku = f$ in Ω. This in turn shows that $(v,Mu)_B$ must be zero. Friedrichs [1] shows that if, in addition, M is admissible, then $Mu = 0$. The conclusion then is that a weak solution which satisfied admissible boundary conditions and is continuously differentiaable is also a classical solution to (1.5).

A function $u \in \mathcal{H}$ is a <u>strong solution</u> to (1.5) if there exists a sequence $\{u^i\}$ of functions such that each $u^i \in C_1(\Omega)$ and

$$\lim_{i \to \infty} \left\{ \|u^i - u\| + \|Ku^i - f\| + \|Mu^i\|_B \right\} = 0$$

Variations of the definitions of weak and strong solutions are common (cf. Sarason [3]). In general it is not known whether a weak solution is differentiable; it is, however, possible, under certain additional hypotheses, to show that a weak solution is also a strong solution. One hypothesis used by Friedrichs [1] is that $\partial\Omega$ has a continuous normal. Sarason [3] considers the case where $\partial\Omega$ is of class C_2. Sarason also considers the two-dimensional case with corners, which requires special conditions to be satisfied at the corners. Other "weak=strong" theorems are given in Sarason [3], Lax and Phillips [4], and Phillips and Sarason [5].

1.5 A Simple Example

An illustration of the types of boundary conditions with more or less boundary data than usual can be given by means of a one-dimensional example. Suppose that

$$Ku = 2x \frac{du}{dx} + 2u = 0 \quad \text{for } -1 \leq x \leq 1 \tag{1.17}$$

If we write K in self adjoint form

$$Ku = x \frac{du}{dx} + \frac{d(xu)}{dx} + u$$

we have $\alpha = x$ and $G = 1$, so that K is positive symmetric. At $x = -1$, $\beta = n\alpha = -x$, and we can let $\mu = |\beta| = -x$. Hence $M = \mu - \beta = 0$ and no boundary condition is imposed at $x = -1$. At $x = 1$, $\beta = x$, and letting $\mu = |\beta|$, we have again that $M = 0$, and no boundary condition is necessary at the right end either. Thus, for equation (1.17), no boundary condition at all is an admissible boundary condition! To see that this is so, we can calculate the solution to (1.17). Since $Ku = 2\, d(xu)/dx = 0$, we have $xu = c$, as the general solution. However, the theory is concerned only with solutions in $L^2(-1,1)$, and $u = c/x$ is square integrable only for $c = 0$, so we do indeed have a unique solution in $L^2(-1,1)$ without specifying any boundary data at all.

A simple example can also be given of an ordinary differential equation which requires more boundary data than usual. For this let

$$Ku = -2x \frac{du}{dx} = -2 \tag{1.18}$$

In self adjoint form

$$Ku = -x \frac{du}{dx} - \frac{d(xu)}{dx} + u$$

so that $\alpha = -x$ and $G = 1$. In this case if we make $\mu = |\beta|$, we get $\mu = -\beta$, so that $M = \mu - \beta = 2$, at both $x = 1$, and $x = -1$. Hence, boundary data must be specified at both end points for admissible boundary conditions. Again, we can check this by solving the equation. The general solution to (1.18) is

$$u = \log|x| + c$$

Since $\int_0^1 \log^2 x < \infty$ we see that we have a valid solution for any c. Also, because of the singularity at $x = 0$, we can specify the value of u at both $x = 1$ and $x = -1$.

CHAPTER II

FINITE DIFFERENCE SOLUTION OF SYMMETRIC POSITIVE DIFFERENTIAL EQUATIONS

2.1 Finite Difference Approximation to K and M

First we will express K in a form slightly different from (1.3), by the use of (1.2). We have

$$
\begin{aligned}
Ku &= \alpha \cdot \nabla u + \nabla \cdot (\alpha u) + Gu \\
&= 2\nabla \cdot (\alpha u) - (\nabla \cdot \alpha)\, u + Gu \qquad (2.1)
\end{aligned}
$$

Using the concept of vectors whose components are themselves matrices or vectors leads to somewhat simpler notation for the application of Green's theorem.

Lemma 2.1 (Green's Theorem) Let g be a continuously differentiable m-dimensional vector-valued function defined on $\Omega \subset R^m$, with vector components in either R, R^r or $R^r \times R^r$. Then

$$\int_\Omega \nabla \cdot g = \int_{\partial\Omega} g \cdot n \qquad (2.2)$$

Proof - Consider the case when g has matrix components, i.e., $g = (g^1, g^2, \ldots, g^m)$ where $g^i = (g^i_{j,k})$ is an $r \times r$ matrix. Then

$$\int_\Omega \nabla \cdot g = \int_\Omega \sum_{i=1}^{m} \frac{\partial}{\partial x_i} (g^i)$$

is a matrix. Using the subscript j,k to indicate the element in the j^{th} row and k^{th} column, we have

$$\left(\int_\Omega \nabla \cdot g\right)_{j,k} = \int_\Omega \sum_{i=1}^{m} \frac{\partial}{\partial x_i}\left(g^i_{j,k}\right) = \int_\Omega (\nabla \cdot g_{j,k})$$

(using obvious notation); therefore

$$\int_\Omega \left(\nabla \cdot g\right)_{j,k} = \int_{\partial\Omega} g_{j,k} \cdot n = \left(\int_{\partial\Omega} g \cdot n\right)_{j,k}$$

Similarly, the result holds when g has vector components, so the lemma is proved.

We now integrate the equation $Ku = f$ over any region $P \subset \Omega$ using (2.1) and Green's theorem to obtain

$$\int_P Ku = \int_P \left[2\nabla \cdot (\alpha u) - (\nabla \cdot \alpha)u + Gu\right]$$

$$= 2\int_{\partial P} \alpha u \cdot n - \int_P (\nabla \cdot \alpha)u + \int_P Gu$$

$$= 2\int_{\partial P} \beta u - \int_P (\nabla \cdot \alpha)u + \int_P Gu = \int_P f \qquad (2.3)$$

By a suitable approximation to (2.3) the desired finite difference equations will be obtained.

Let H be a set of N mesh points for Ω. It is not required for the theory that the mesh points all lie in Ω. With each mesh point $x_j \epsilon H$ we identify a mesh region, $P_j \subset \Omega$ by

$$P_j = \left\{x \,\middle|\, |x - x_j| < |x - x_k|,\ \forall x_k \epsilon H,\ k \neq j;\ x \epsilon \Omega\right\}$$

If P_j is adjacent to P_k we say that x_j is connected to x_k (corresponding to the fact that the directed graph of the resulting matrix will have a directed path in both directions between j and k, see p. 16, [6]). Let $l_{j,k} = |x_j - x_k|$, where x_j is connected to x_k, and let $h = \max l_{j,k}$. Now define A_j to be the "volume" of P_j and $L_{j,k}$ to be the "area" of the $r - 1$ dimensional "surface" between P_j and P_k. We put $\Gamma_{j,k} = \overline{P}_j \cap \overline{P}_k$. Figure 1 illustrates mesh points and corresponding mesh regions for two dimensions. This concept of mesh regions is based on the suggestions of MacNeal [7]. We will always use the notation $\sum_j$ to indicate a sum over all points, x_j, in H, and $\sum_k$ to indicate a sum over points, x_k, which are connected to some one point, x_j.

The desired finite difference equation can now be obtained by a suitable approximation to equation (2.3). We use the symbol $\doteq$ to indicate the discrete approximation that will be used for each expression. First

$$\int_{\Gamma_{j,k}} \beta u \doteq L_{j,k}\beta_{j,k} \frac{u_j + u_k}{2} \tag{2.4}$$

where $u_j = u(x_j)$ and $\beta_{j,k}$ is the value of β for P_j at the center of $\Gamma_{j,k}$. (Note that $\beta_{j,k} = -\beta_{k,j}$). The approximation to the next term of equation (2.3) requires approximating u with u_j first, and then applying Green's theorem before approximating α. With this we obtain

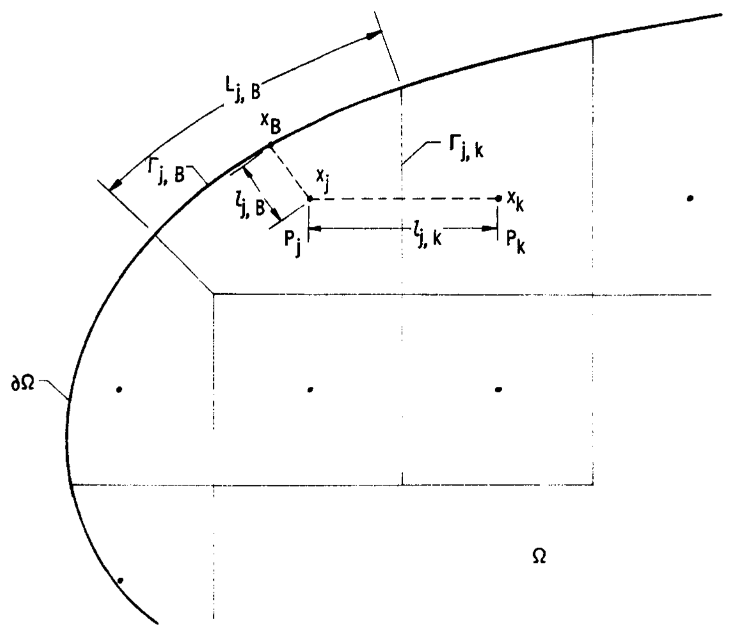


Figure 1. - Typical mesh regions in the two-dimensional case.

$$\int_{P_j} (\nabla \cdot \alpha)u \doteq \int_{P_j} (\nabla \cdot \alpha)u_j \doteq \int_{\partial P_j} \beta u_j \tag{2.5}$$

The final approximation is then

$$\int_{\Gamma_{j,k}} \beta u_j \doteq L_{j,k}\beta_{j,k}u_j \tag{2.6}$$

Equations (2.4) and (2.6) take care of the integration over the interface between any P_j and P_k. Now we need to make an approximation for the boundary sides. It will be convenient to be able to subdivide $\overline{P}_j \cap \partial\Omega$ into more than one piece. We will label each piece $\Gamma_{j,B}$ and we will use the convention that $\sum_B$ will mean a summation over the B for just one j. We use $l_{j,B}$ to denote the distance from x_j to x_B, where x_B is located at the "center" of $\Gamma_{j,B}$ and $L_{j,B}$ is used for the "area" of $\Gamma_{j,B}$. Also $\beta_{j,B} = \beta(x_B)$. This notation is indicated for the two-dimensional case in Figure 1. The desired approximations are now given by

$$\int_{\Gamma_{j,B}} \beta u \doteq L_{j,B}\beta_{j,B}u_B \tag{2.7}$$

$$\int_{\Gamma_{j,B}} \beta u_j \doteq L_{j,B}\beta_{j,B}u_j \tag{2.8}$$

Finally the remaining terms in equation (2.3) are approximated by

$$\int_{P_j} Gu \doteq A_jG_ju_j \tag{2.9}$$

$$\int_{P_j} f \doteq A_j f_j \tag{2.10}$$

where $G_j = G(x_j)$ and $f_j = f(x_j)$. Also we can approximate $\int Ku$ by

$$\int_{P_j} Ku \doteq A_j (K_h u)_j \tag{2.11}$$

where K_h is the finite difference operator to be defined and which will approximate K. Using approximations (2.4) to (2.11) in equation (2.3) we arrive at the following definition of K_h,

$$\begin{aligned} A_j(K_h u)_j &= \sum_k L_{j,k}\beta_{j,k}(u_j + u_k) + 2 \sum_B L_{j,B}\beta_{j,B}u_B \\ &\quad - \sum_k L_{j,k}\beta_{j,k}u_j - \sum_B L_{j,B}\beta_{j,B}u_j + A_j G_j u_j \\ &= \sum_k L_{j,k}\beta_{j,k}u_k + \sum_B L_{j,B}\beta_{j,B}(2u_B - u_j) + A_j G_j u_j \end{aligned} \tag{2.12}$$

where u here denotes a discrete function defined on $\overline{H} = H \cup \{x_B\}$, and $u_j = u(x_j)$. We will seek to find a function defined on $\overline{H}$ and satisfying $(K_h u)_j = f_j$ for every $x_j \epsilon H$. Of course the solution is not yet uniquely determined since there are more unknowns than equations. The boundary condition $Mu = 0$ will furnish us with the necessary information to determine u uniquely on H (but not necessarily on all of $\overline{H}$).

Using M_h to denote the boundary operator used to approximate M, we make the following definition

$$(M_h u)_{j,B} = \mu_{j,B}u_j - \beta_{j,B}(2u_B - u_j) \tag{2.13}$$

for all j where P_j is a boundary polygon, and for all boundary surfaces of P_j (each of which is associated with a point x_B). It is easily seen that M_h is consistent with M (i.e., $(M_h u)_{j,B} \to Mu(x_{j,B})$ as $h \to 0$ if u is continuous). The reason for this choice of M_h is that the condition $M_h u = 0$ can be used to eliminate u_B in $K_h u$ in a simple manner, and also we will be able to prove basic identities for the finite difference operators analogous to those for the continuous operators (eqs. (1.12) and (1.13)).

2.2 Basic Identities for the Finite Difference Operators

The existence and uniqueness of a solution to the finite difference equation and the convergence to a continuous solution as $h \to 0$ depends on proving the basic identities for the discrete operators. Let $\mathcal{H}_h$ be the finite dimensional Hilbert space of discrete functions defined on H. The inner product is given by

$$(u,v)_h = \sum_j A_j u_j \cdot v_j, x_j \epsilon H \tag{2.14}$$

and

$$\|u\|_h^2 = (u,u)_h \tag{2.15}$$

Also a "boundary" inner product is given by

$$(u,v)_{B_h} = \sum_j \sum_B L_{j,B} u_{j,B} \cdot v_{j,B} \tag{2.16}$$

for P_j a boundary mesh region, and

$$\|u\|_{B_h}^2 = (u,u)_{B_h} \tag{2.17}$$

The discrete adjoint operators K_h^* and M_h^* are defined in the obvious way,

$$A_j(K_h^*u)_j = -\sum_k L_{j,k}\beta_{j,k}u_k - \sum_B L_{j,B}\beta_{j,B}(2u_B - u_j) + A_jG_j^*u_j \tag{2.18}$$

$$(M_h^*u)_{j,B} = u_{j,B}^*u_j + \beta_{j,B}(2u_B - u_j) \tag{2.19}$$

We can now give the "first identity" for the discrete operators.

Lemma 2.2 If K is symmetric positive, then

$$(v,K_hu)_h + (v,M_hu)_{B_h} = (K_h^*v,u)_h + (M_h^*v,u)_{B_h} \tag{2.20}$$

for any functions u,v defined on $\bar{H}$.

Proof - Using the definitions, equations (2.12) and (2.18), we have

$$\begin{aligned}
(v,K_hu)_h - (K_h^*v,u)_h &= \sum_j \left[v_j \cdot A_j(K_hu)_j - A_j(K_h^*v)_j \cdot u_j\right] \\
&= \sum_j \Big[\sum_k L_{j,k}v_j \cdot \beta_{j,k}u_k \\
&\quad + \sum_B L_{j,B}v_j \cdot \beta_{j,B}(2u_B - u_j) + A_jv_j \cdot G_ju_j \\
&\quad + \sum_k L_{j,k}\beta_{j,k}v_k \cdot u_j \\
&\quad + \sum_B L_{j,B}\beta_{j,B}(2v_B - v_j) \cdot u_j - A_jG_j^*v_j \cdot u_j\Big]
\end{aligned}$$

By rearrangement, since $\beta_{j,k} = -\beta_{j,k}$, and since $\beta_{j,k}$ is symmetric we have

$$\sum_j \sum_k L_{j,k}\beta_{j,k}v_k \cdot u_j = \sum_j \sum_k L_{j,k}\beta_{k,j}v_j \cdot u_k = -\sum_j \sum_k L_{j,k}v_j \cdot \beta_{j,k}u_k$$

and we see that all terms cancel with the exception of the boundary terms, so that

$$(v,K_h u)_h - (K_h^* v,u)_h = \sum_j \sum_B L_{j,B}\Big(v_j \cdot \beta_{j,B}(2u_B - u_j) + \beta_{j,B}(2v_B - v_j) \cdot u_j\Big) \tag{2.21}$$

On the other hand, using equations (2.13) and (2.19)

$$(M_h^* v,u)_{B_h} - (v,M_h u)_{B_h} = \sum_j \sum_B L_{j,B}\Big(\mu_{j,B}^* v_j \cdot u_j + \beta_{j,B}(2v_B - v_j) \cdot u_j\Big) - \sum_j \sum_B L_{j,B}\Big(v_j \cdot \mu_{j,B}u_j - v_j \cdot \beta_{j,B}(2u_B - u_j)\Big)$$

which is the same as the right side of (2.21). Hence the "first identity" for the difference operators is proved.

The discrete operators have been defined so that $K_h + K_h^* = G + G^*$ and $M_h + M_h^* = \mu + \mu^*$. By letting $v = u$ in (2.20) we can prove the discrete "second identity" exactly as for the continuous case (Lemma 1.2).

<u>Lemma 2.3</u> If K is symmetric positive, then

$$(u,K_h u)_h + (u,M_h u)_{B_h} = (u,Gu)_h + (u,\mu u)_{B_h} \tag{2.22}$$

2.3 Existence of Solution to Finite Difference Equations

Using equation (2.13) and $M_h u = 0$ we can eliminate u_B from equation (2.12) so that the equation $K_h u = f$ can be reduced to

$$\sum_k L_{j,k}\beta_{j,k}u_k + \sum_B L_{j,B}\mu_{j,B}u_j + A_j G_j u_j = A_j f_j, \forall j \qquad (2.23)$$

If we consider the case when Ω is two-dimensional and rectangular, and the P_j are all equal rectangles, we can compare (2.23) with the finite difference equation obtained by Chu [2]. The equation obtained by Chu is the same as (2.23) for interior rectangles, but is different for boundary rectangles.

Let A be the $rN \times rN$ matrix of coefficients of (2.23). Letting $\langle u,v\rangle = \sum_j u_j \cdot v_j$, the ordinary vector inner product, we have

$$\langle u,Au\rangle = (u,K_h u)_h + (u,M_h u)_{B_h} \qquad (2.24)$$

Hence, by the "second identity" (2.22), A has positive definite symmetric part which shows that A is non-singular. We can also obtain an a priori bound for $\|u\|_h$ just as in the continuous case.

Lemma 2.4 Suppose u is a solution to

$$K_h u = f, \; M_h u = 0$$

where K is symmetric positive and M is semi-admissible. Then

$$\|u\|_h \leq \frac{1}{\lambda_G} \|f\|_h \qquad (2.25)$$

If in addition, $(\mu + \mu^*)$ is positive definite on $\partial\Omega$, then

$$\|u\|_{B_h} \leq \frac{1}{\sqrt{\lambda_G \lambda_u}} \|f\|_h \qquad (2.26)$$

Proof - The proof is identical to that for Lemmas 1.3 and 1.4, but using the h norms and inner products.

2.4 Convergence of the Finite Difference Solution to a Continuous Solution

It is possible to show that the solution of the finite difference equation (2.23) converges strongly to a continuously differentiable solution of equation (1.5), under the proper hypotheses. For simplicity we prove convergence only for the case when Ω is two-dimensional ($m = 2$). Extension to regions in higher dimensions, with the same rate of convergence, follows directly. To allow the type of comparison we wish to make we will define operators mapping $\mathcal{H}$ into $\mathcal{H}_h$ and vice versa. Let $r_h\colon \mathcal{H} \to \mathcal{H}_h$ be the projection defined by

$$(r_h u)_j = u(x_j) \text{ for all } x_j \in H \tag{2.27}$$

In the other direction, let $p_h\colon \mathcal{H}_h \to \mathcal{H}$ be an injection mapping defined by

$$p_h u_h(x) = (u_h)_j, \text{ for all } x \in P_j \tag{2.28}$$

We immediately have the following relations,

$$r_h p_h = I \tag{2.29}$$

$$\|p_h u_h\| = \|u_h\|_h \text{ for all } u_h \in \mathcal{H}_h \tag{2.30}$$

We can now state our basic convergence theorem for two-dimensional regions.

Theorem 2.1 Suppose that $u \in C^2(\overline{\Omega})$ satisfies

$$Ku = f \quad \text{on} \quad \Omega \subset R^2$$

$$Mu = 0 \quad \text{on} \quad \partial\Omega$$

where K is symmetric positive, and $\mu + \mu^*$ is positive definite on $\partial\Omega$. For any given $h > 0$, let H_h be a set of associated mesh points such that the maximum distance between connected nodes is less than h and also that $L_{j,k}$, $L_{j,B}$ and $|x - x_j|$ for $x \epsilon P_j$ are all less than h. It is assumed that the mesh is sufficiently regular so that h^2/A_j for each P_j is bounded independently of h by a constant $K_1 > 0$, which is possible for sufficiently nice regions. Also it is assumed that a uniform rectangular mesh is used for all P_j any point of which is at a distance greater than $K_2 h$ from $\partial\Omega$, where K_2 is a positive constant. It is assumed that $\alpha \epsilon C^2(\bar{\Omega})$.

Let $u_h \epsilon \mathcal{H}_h$ be the unique solution to

$$K_h u_h = r_h f \quad \text{on} \quad H_h$$

$$M_h u_h = 0$$

Then $\|p_h u_h - u\| = O(h^\nu)$ as $h \to 0$ for any positive $\nu < 1/2$.

Chu [2] proved convergence of his finite difference scheme, where Ω is a rectangle or a region with four corners, but the rate of convergence was not established.

<u>Proof</u> - Define $w_h = u_h - r_h u$. Let λ_G be the smallest eigenvalue of $(G + G^*)/2$ in $\bar{\Omega}$. Using the "second identity" (2.22), we have

$$\|w_h\|_h^2 \leq \frac{1}{\lambda_G} \left[(w_h, G w_h)_h + (w_h, \mu w_h)_{B_h}\right] = \frac{1}{\lambda_G} \left[(w_h, K_h w_h)_h + (w_h, M_h w_h)_{B_h}\right]$$

Using the Cauchy-Schwartz inequality, we have

$$\|w_h\|_h^2 \le \frac{1}{\lambda_G} (\|w_h\|_h \; \|K_h w_h\|_h + \|w_h\|_{B_h} \; \|M_h w_h\|_{B_h}) \tag{2.31}$$

We will show that $\|K_h w_h\|_h = O(h^{1/2})$ and $\|M_h w_h\|_{B_h} = O(h)$, as $h \to 0$. We shall need the following lemma.

Lemma 2.5 Let g be a function defined on a finite region $P \subset R^2$, and suppose that g satisfies a Lipschitz condition, i.e., there is a constant $K_3 > 0$ such that $|g(x) - g(y)| \le K_3 |x - y|$, for all $x, y \in P$. Then, if A_o is the area of P and $|x - x_o| \le h$ in P,

$$\left| g(x_o) - \frac{1}{A_o} \int_P g(x) \right| \le K_3 h$$

Proof - By direct calculation

$$\left| g(x_o) - \frac{1}{A_o} \int_P (g(x)) \right| = \frac{1}{A_o} \left| \int_P (g(x_o) - g(x)) \right| \le \frac{1}{A_o} \int_P K_3 |x - x_o| \le K_3 h$$

We proceed now with the proof of the theorem. Let Ω_1 denote that portion of Ω consisting of those P_j which are rectangular, and let Ω_2 denote the rest of the P_j. From the hypothesis we see that the area of Ω_2 is less than the length of $\partial\Omega$ times $K_2 h$. We have now that

$$\|K_h w_h\|_h^2 = \int_{\Omega_1} (p_h K_h w_h)^2 + \int_{\Omega_2} (p_h K_h w_h)^2$$

$$= \sum_{j \in J_1} \int_{P_j} (Ku(x_j) - (K_h r_h u)_j)^2 + \sum_{j \in J_2} \int_{P_j} (Ku(x_j) - (K_h r_h u)_j)^2 \tag{2.32}$$

where

$$J_i = \{ j \mid P_j \subset \Omega_i \}, \; i = 1,2$$

To simplify notation we will use u_j for $u(x_j)$ and u_B for $u(x_B)$. We now obtain a suitable bound for $|Ku(x_j) - (K_h r_h u)_j|$

$$
\begin{aligned}
|Ku(x_j) - (K_h r_h u)_j| &= \Big| 2\nabla\cdot(\alpha u)(x_j) - (\nabla\cdot\alpha)u(x_j) + G_j u_j \\
&\quad - \sum_k \frac{L_{j,k}}{A_j}\beta_{j,k}(u_j + u_k) - 2\sum_B \frac{L_{j,B}}{A_j}\beta_{j,B}u_B \\
&\quad + \sum_k \frac{L_{j,k}}{A_j}\beta_{j,k}u_j + \sum_B \frac{L_{j,B}}{A_j}\beta_{j,B}u_j - G_j u_j \Big| \\
&\le \Big| 2\nabla\cdot(\alpha u)(x_j) - \sum_k \frac{L_{j,k}}{A_j}\beta_{j,k}(u_j + u_k) - 2\sum_B \frac{L_{j,B}}{A_j}\beta_{j,B}u_B \Big| \\
&\quad + \Big| (\nabla\cdot\alpha)u(x_j) - \sum_k \frac{L_{j,k}}{A_j}\beta_{j,k}u_j - \sum_B \frac{L_{j,B}}{A_j}\beta_{j,B}u_j \Big|
\end{aligned}
\tag{2.33}
$$

Consider the first term in the last expression above

$$
\begin{aligned}
&\Big| 2\nabla\cdot(\alpha u)(x_j) - \sum_k \frac{L_{j,k}}{A_j}\beta_{j,k}(u_j + u_k) - 2\sum_B \frac{L_{j,B}}{A_j}\beta_{j,B}u_B \Big| \\
&\qquad \le \Big| 2\nabla\cdot(\alpha u)(x_j) - \frac{2}{A_j}\int_{P_j} \nabla\cdot(\alpha u) \Big| \\
&\qquad + \frac{1}{A_j}\Big| \sum_k \int_{\Gamma_{j,k}} 2(\beta u - (\beta u)_{j,k}) \\
&\qquad + \sum_B \int_{\Gamma_{j,B}} 2(\beta u - (\beta u)_{j,B}) \Big| \\
&\qquad + \frac{1}{A_j}\Big| \sum_k \int_{\Gamma_{j,k}} \beta_{j,k}(2u_{j,k} - (u_j + u_k)) \Big|
\end{aligned}
\tag{2.34}
$$

By Lemma 2.5, since α and $u \in C^2(\bar{\Omega})$ imply that their derivatives satisfy a Lipschitz condition,

$$\left| 2\nabla \cdot (\alpha u)(x_j) - \frac{2}{A_j} \int_{P_j} \nabla \cdot (\alpha u) \right| = O(h) \tag{2.35}$$

We consider now the case when $j \in J_1$, so that P_j is a rectangle with x_j at the center.

Since $u \in C^2(\Omega)$, we have

$$u_j = u_{j,k} - \frac{l_{j,k}}{2} u'_{j,k} + \frac{l_{j,k}^2}{(4)2} u''(\xi_1)$$

$$u_k = u_{j,k} + \frac{l_{j,k}}{2} u'_{j,k} + \frac{l_{j,k}^2}{(4)2} u''(\xi_2)$$

where the derivatives are directional derivatives in the direction $x_k - x_j$. Hence, if $|u''| < K_3$ in Ω, we have

$$|2u_{j,k} - (u_j + u_k)| < \frac{K_3}{4} h^2$$

This means that

$$\left| \int_{\Gamma_{j,k}} \beta_{j,k}(2u_{j,k} - (u_j + u_k)) \right| \leq L_{j,k} \, \|\beta_{j,k}\| \; |2u_{j,k} - (u_j + u_k)| = O(h^3) \tag{2.36}$$

when $j \in J_1$.

We now examine a Taylor series expansion for βu about the point $x_{j,k} = (x_j + x_k)/2$.

$$\left.\begin{aligned}\beta(x_{j,k}+tz)u(x_{j,k}+tz) &= (\beta u)_{j,k} + t\left(\frac{d}{dt}(\beta u)\right)_{j,k} + \frac{t^2}{2}g(\xi^1)\\ \beta(x_{j,k}-tz)u(x_{j,k}-tz) &= (\beta u)_{j,k} - t\left(\frac{d}{dt}(\beta u)\right)_{j,k} + \frac{t^2}{2}g(\xi^2)\end{aligned}\right\}\quad(2.37)$$

where z is a unit vector orthogonal to $x_j - x_k$, t is a scalar parameter, $g(\xi) = (g_1(\xi_1), g_2(\xi_2), \ldots, g_r(\xi_r))$, g_i is the i^{th} component of the vector $d^2/dt^2\ (\beta u)$, and ξ_i is a point on the straight line between $x_{j,k} + (L_{j,k}/2)z$ and $x_{j,k} - (L_{j,k}/2)z$. Using (2.37) we obtain the following bound,

$$\begin{aligned}\left|\int_{\Gamma_{j,k}} \beta u - (\beta u)_{j,k}\right| &= \left|\int_0^{L_{j,k}/2} (\beta(x_{j,k}+tz)u(x_{j,k}+tz)\right.\\ &\qquad \left. + \beta(x_{j,k}-tz)u(x_{j,k}-tz) - 2(\beta u)_{j,k})dt\right|\\ &\leq \int_0^{L_{j,k}/2} t^2K_4\ dt = O(h^3)\end{aligned}\qquad(2.38)$$

Now, using (2.35), (2.36) and (2.38) in (2.34) we obtain

$$|2\nabla\cdot(\alpha u)(x_j) - \sum_k \frac{L_{j,k}}{A_j}\beta_{j,k}(u_j + u_k)| = O(h)\qquad(2.39)$$

for all $j \in J_1$, since $h^2/A_j \leq K_1$ and the boundary terms are not present.

Consider now the second term on the right of (2.33):

$$\left|(\nabla \cdot \alpha)u(x_j) - \sum_k \frac{L_{j,k}}{A_j} \beta_{j,k}u_j - \sum_B \frac{L_{j,B}}{A_j} \beta_{j,B}u_j\right|$$

$$\leq \left|(\nabla \cdot \alpha)u(x_j) - \frac{1}{A_j} \int_{P_j} (\nabla \cdot \alpha)u\right| + \frac{1}{A_j}\left|\int_{P_j} (\nabla \cdot \alpha)(u - u_j)\right|$$

$$+ \frac{1}{A_j}\left|\sum_k \int_{\Gamma_{j,k}} (\beta - \beta_{j,k})u_j + \sum_B \int_{\Gamma_{j,B}} (\beta - \beta_{j,B})u_j\right| \qquad (2.40)$$

By Lemma 2.5

$$\left|(\nabla \cdot \alpha)u(x_j) - \frac{1}{A_j} \int_{P_j} (\nabla \cdot \alpha)u\right| = O(h) \qquad (2.41)$$

Next, since u satisfies a Lipschitz condition, $|x - x_j| < h$ for all $x \in P_j$, and since $\|\nabla \cdot \alpha\|$ is uniformly bounded in Ω, we have

$$\frac{1}{A_j}\left|\int_{P_j} (\nabla \cdot \alpha)(u - u_j)\right| = O(h) \qquad (2.42)$$

Finally, since $\beta_{j,k}$ and $\beta_{j,B}$ are each evaluated at the midpoint of $\Gamma_{j,k}$ or $\Gamma_{j,B}$, respectively, we can use a Taylor series analysis, as in deriving equation (2.38), to obtain

$$\frac{1}{A_j}\left|\sum_k \int_{\Gamma_{j,k}} (\beta - \beta_{j,k})u_j + \sum_B \int_{\Gamma_{j,B}} (\beta - \beta_{j,B})u_j\right| = O(h) \qquad (2.43)$$

Combining (2.41), (2.42), and (2.43) in (2.40) we obtain

$$\left|(\nabla \cdot \alpha)u(x_j) - \sum_k \frac{L_{j,k}}{A_j} \beta_{j,k}u_j - \sum_B \frac{L_{j,B}}{A_j} \beta_{j,B}u_j\right| = O(h) \qquad (2.44)$$

Note that (2.44) holds for all j, not just for $j \in J_1$.

We can now substitute (2.39) and (2.44) in (2.33) to obtain

$$|Ku(x_j) - (K_h r_h u)_j| = O(h) \quad \text{for all} \quad j \in J_1 \tag{2.45}$$

We cannot obtain as good a bound for $|Ku(x_j) - (K_h r_h u)_j|$ when $j \in J_2$, although (2.44) holds, since $\Gamma_{j,k}$ is not in general bisected by the line between x_j and x_k. However, we can show that $|Ku(x_j) - (K_h r_h u)_j|$ is uniformly bounded for $j \in J_2$, which is adequate since the area of Ω_2 is of order h. The two inequalities which must be re-examined are (2.36) and (2.38).

We now have, since u and (βu) satisfy Lipschitz conditions, that

$$\left| \int_{\Gamma_{j,k}} \beta_{j,k}(2u_{j,k} - (u_j + u_k)) \right| = O(h^2) \tag{2.46}$$

$$\left.\begin{aligned} \left| \int_{\Gamma_{j,k}} \beta u - (\beta u)_{j,k} \right| &= O(h^2) \\ \left| \int_{\Gamma_{j,B}} \beta u - (\beta u)_{j,B} \right| &= O(h^2) \end{aligned}\right\} \tag{2.47}$$

Using this, with the other results which still hold, we see that $|Ku(x_j) - (K_h r_h u)_j|$ is uniformly bounded for $j \in J_2$, as $h \to 0$. Using this, together with (2.45) in (2.32) we obtain

$$\|K_h w_h\|_h^2 = O(h^2) + O(h) \tag{2.48}$$

so that

$$\|K_h w_h\|_h = O(h^{1/2}) \tag{2.49}$$

The next step is to show that $\|M_h w_h\|_{B_h} = O(h)$. We have

$$\|M_h w_h\|_B = \|M_h u_h - M_h r_h u\|_{B_h} = \|M_h r_h u\|_{B_h}$$

since $M_h u_h = 0$. Now

$$\|M_h r_h u\|_{B_h} = \sum_j \sum_B L_{j,B}(M_h r_h u)^2_{j,B}$$

$$= \sum_j \sum_B L_{j,B}(\mu_{j,B}u_j - \beta_{j,B}(2u_B - u_j))^2$$

However, using the fact that $\beta_{j,B}u_B = \mu_{j,B}u_B$,

$$|\mu_{j,B}u_j - \beta_{j,B}(2u_B - u_j)| \le |\mu_{j,B}(u_j - u_B)| + |\beta_{j,B}(u_B - u_j)| = O(h)$$

since u is differentiable, and $\|\mu\|$ and $\|\beta\|$ are uniformly bounded. This shows that

$$\|M_h r_h u\|^2_{B_h} = O(h^2),$$

since $\sum_{j,B} L_{j,B}$ is simply the length of $\partial\Omega$. This proves that

$$\|M_h w_h\|_{B_h} = O(h) \tag{2.50}$$

Using (2.49) and (2.50) in (2.31), we see that

$$\|w_h\|^2_h = \|w_h\|_h O(h^{1/2}) + \|w_h\|_{B_h} O(h) \tag{2.51}$$

From Lemma 2.4, $\|w_h\|_{B_h}$ must be bounded, since

$$\|w_h\|_{B_h} \le \|u_h\|_{B_h} + \|r_h u\|_{B_h}$$

$$\le \frac{1}{\sqrt{\lambda_G \lambda_\mu}} \|r_h f\|_h + \|r_h u\|_{B_h}$$

which is certainly uniformly bounded as $h \to 0$. Likewise $\|w_h\|_h$ is bounded. So from (2.51) we have

$$\|w_h\|_h = O(h^{1/4}) \tag{2.52}$$

However, if we use (2.52) in (2.51) we get $\|w_h\|_h = O(h^{3/8})$, or by repeating this procedure enough times,

$$\|w_h\|_h = O(h^\nu), \text{ for any positive } \nu < 1/2 \tag{2.53}$$

Finally, we establish the convergence rate for $\|p_h u_h - u\|$. We have

$$\|p_h u_h - u\| \le \|p_h u_h - p_h r_h u\| + \|p_h r_h u - u\|$$

$$= \|w_h\|_h + \|p_h r_h u - u\| \tag{2.54}$$

The last term can be estimated by

$$\|p_h r_h u - u\|^2 = \sum_j \int_{P_j} (u_j - u)^2 = O(h^2) \tag{2.55}$$

Using (2.53) and (2.55) in (2.54) we get

$$\|p_h u_h - u\| = O(h^\nu) + O(h) = O(h^\nu), \text{ for any positive } \nu < 1/2 \tag{2.56}$$

This completes the proof of Theorem (2.1).

2.5 Solution of the Finite Difference Equation

For our method to be of practical use we must have some method for computing the solution to the finite difference equation (2.23). We will consider only the two-dimensional case here. In any case we can partition the matrix A so as to be block tridiagonal. For example, suppose that the mesh points H lie on lines such that the mesh points on any one line are connected only to points on the same line or adjacent lines. Then we can partition A into blocks corresponding to each line. The diagonal blocks will themselves be block tridiagonal with $r \times r$ blocks. The matrix equation can then be solved by the block tridiagonal algorithm ([8] and [6], p. 196). We suppose A to be written in the form,

$$A = \begin{pmatrix} B_1 & C_1 & & & \bigcirc \\ A_2 & B_2 & C_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & C_{NL-1} \\ \bigcirc & & & A_{NL} & B_{NL} \end{pmatrix} \tag{2.57}$$

where NL is the number of lines. Each B_i is an $rn \times rn$ block tridiagonal matrix, where n is the number of points on the i^{th} line. From equation (2.23) since $\beta_{j,k} = -\beta_{k,j}$ we see that $A_i = C^*_{i-1}$. Thus C_i need not be stored for a computer solution. The block tridiagonal algorithm is completely analogous to the ordinary tridiagonal algorithm. Suppose the equation to be solved is $Au = f$, where u and f are partitioned as required. A typical block equation is

$$A_i u_{i-1} + B_i u_i + C_i u_{i+1} = f_i$$

First let

$$W_1 = B_1$$

$$y_1 = f_1$$

The forward sweep is given by

$$\left.\begin{aligned} G_i &= A_i W_{i-1}^{-1} \\ y_i &= f_i - G_i y_{i-1} \\ W_i &= B_i - G_i C_{i-1} \end{aligned}\right\} \text{ for } i = 2,3, \ldots, NL$$

This is followed by the backward sweep. First,

$$u_{NL} = W_{NL}^{-1} y_{NL}$$

Then

$$u_i = W_i^{-1}(y_i - C_i u_{i+1}) \text{ for } i = NL - 1, NL - 2, \ldots, 1$$

Of course this algorithm will not work for every non-singular block tridiagonal matrix. However, Schecter [8], gives a sufficient condition for the validity of the algorithm, and that is simply that A has definite symmetric part. We have already shown that A has positive definite symmetric part. There is one real disadvantage to the method, however, and that is the fact that each W_i^{-1} is a full matrix and must be stored during the forward sweep for use on the backward sweep. This results in large computer storage requirements, and the use of tapes or disks for temporary storage for only a moderate number of mesh points. This, of course, is very time consuming. An alternate procedure is suggested by Schecter [8]. In Schecter's method only one matrix need be inverted and stored for a number of consecutive lines with an equal number of points per line. However, the matrix to be inverted may be ill-conditioned if too many lines are grouped in this way.

An alternate method of solution may be possible in some cases. Note that A may be decomposed as

$$A = D + S$$

where D is Hermitian and positive definite, and S is skew symmetric. The eigenvalues of D are usually easy to calculate since D is block diagonal with $r \times r$ blocks. If the smallest eigenvalue, λ_D, of D is larger than the spectral radius, $\rho(S)$, of S, we will have

$$\|D^{-1}S\| \le \|D^{-1}\| \; \|S\| = \frac{1}{\lambda_D}\,\rho(S) < 1$$

In this case we could use the following iterative method. Let $u^{(o)}$ be arbitrary, and define $u^{(i)}$ recursively by

$$Du^{(i)} = -Su^{(i-1)} + f$$

In this case $\lim_{i\to\infty} u^{(i)} = u$. In general, though, the eigenvalues of D will not all be sufficiently large for this simple method to work. However, the original finite difference equations can be modified in some cases by the addition of a "viscosity" term, so as obtain a convergent iterative procedure for the solution of the matrix equation. This will be discussed further in Chapter III.

2.6 Convergence to a Weak Solution

We can consider the discrete analogue of a weak solution. Let V_h be the set of discrete functions, v_h, defined on $\overline{\overline{H}}$ and satisfying $M_h^* v_h = 0$. For a discrete weak solution, u_h, we would then require that

$$(K_h^* v_h, u_h)_h = (v_h, r_h f)_h \quad \text{for all} \quad v \in V_h \tag{2.58}$$

Form the "first identity" (2.20) we have then

$$(v_h, r_h f)_h = (v_h, K_h u_h)_h + (v_h, M_h u_h)_{B_h} \quad \text{for all} \quad v \in V_h \tag{2.59}$$

We see from this that $(K_h u_h)_j = f_j$ for all P_j which are not on the boundary, by choosing $(v_h)_j = 1$, and $(v_h)_k = 0$ for $k \neq j$. Because of the discrete nature of the equations we are not assured of u_h satisfying the boundary conditions. However, conversely, if u_h satisfies $K_h u_h = r_h f$ and $M_h u_h = 0$ we see imeediately that (2.58) must be satisfied.

Chu [2] has shown weak convergence of his finite difference solution to a weak solution of a symmetric positive equation and Cea [9] has investigated generally the question of weak or strong convergence of approximate solutions to weak solutions of elliptic equations. Using these ideas, we can prove weak convergence of our finite difference solutions to weak solutions of symmetric positive equations.

Theorem 2.2 For any $h > 0$, let H_h be a set of mesh points satisfying the requirements of Theorem 2.1. It is assumed that $\alpha \epsilon C^2(\bar{\Omega})$. Let u_h be the unique solution to

$$K_h u_h = r_h f$$

$$M_h u_h = 0$$

If $\{h_i\}_{i=1}^{\infty}$ is a positive sequence converging to zero, then $\{p_{h_i} u_{h_i}\}_{i=1}^{\infty}$ has a subsequence which converges weakly in H to a weak solution, u, of equation (1.5), that is

$$(K^* v, u) = (v, f) \text{ for all } v \epsilon V$$

Furthermore, if u is a unique weak solution, then $\{p_{h_i} u_{h_i}\}_{i=1}^{\infty}$ converges weakly to u.

Proof - First we note that $\|p_h u_h\|$ is bounded, since $\|p_h u_h\| = \|u_h\|_h \leq \frac{1}{\lambda_G} \|r_h f\|_h$, by Lemma 2.4. Hence, there is a subsequence of $\{p_{h_i} u_{h_i}\}$ that converges weakly to some $u \in \mathcal{H}$. (See Theorem 4.41-B, Taylor [10].) For convenience of notation we will suppress the subscripts on the h.

We have, for all $v \in V$,

$$|(K_h^* r_h v, u_h)_h - (K^* v, p_h u_h)| \leq \|p_h K_h^* r_h v - K^* v\| \; \|p_h u_h\|$$

$$\leq \left(\|K_h^* r_h v - r_h K^* v\|_h + \|p_h r_h K^* v - K^* v\|\right)\|p_h u_h\| \qquad (2.60)$$

But $\|p_h r_h K^* v - K^* v\| \to 0$, and in Theorem 2.1 we can substitute K^* for K in equation (2.49) to show that $\|K_h^* r_h v - r_h K^* v\| \to 0$ (since $K_h w_h = r_h K u - K_h r_h u$). Since $\|p_h u_h\|$ is bounded,

$$\lim_{h \to 0} |(K_h^* r_h v, u_h)_h - (K^* v, p_h u_h)| = 0$$

However, since $K^* v \in \mathcal{H}$, we know that $\lim_{h \to 0} (K^* v, p_h u_h) = (K^* v, u)$

We have shown, then, that

$$\lim_{h \to 0} (K_h^* r_h v, u_h)_h = (K^* v, u), \quad \text{for all } v \in V. \qquad (2.61)$$

The discrete "first identity", equation (2.20), gives

$$(K_h^* r_h v, u_h)_h + (M^* r_h v, u_h)_{B_h} = (r_h v, K_h u_h)_h + (r_h v, M_h u_h)_{B_h} = (r_h v, r_h f)_h \qquad (2.62)$$

Hence

$$|(K_h^* r_h v, u_h)_h - (r_h v, r_h f)_h| \leq \|M^* r_h v\|_{B_h} \|u_h\|_{B_h} \qquad (2.63)$$

By Lemma 2.4 $\|u_h\|_{B_h} \leq \|r_h f\|_h / \sqrt{\lambda_G \lambda_\mu}$ which is bounded.

Also, the proof of equation (2.50) shows that $\lim_{h\to 0} \|M^* r_h v\|_{B_h} = 0$, for all $v \in V$, so that

$$\lim_{h\to 0} |(K_h^* r_h v, u_h)_h - (r_h v, r_h f)_h| = 0 \tag{2.64}$$

Further, it is obvious that

$$\lim_{h\to 0} (r_h v, r_h f)_h = (v,f) \tag{2.65}$$

Combining (2.61), (2.64) and (2.65) gives

$$(K^* v, u) = (v,f), \text{ for all } v \in V,$$

which completes the proof of the theorem.

CHAPTER III

SPECIAL FINITE DIFFERENCE SCHEME FOR ITERATIVE SOLUTION OF MATRIX EQUATION

3.1 Special Finite Difference Scheme

As pointed out in section 2.5, the matrix equation $Au = f$ can be solved by an iterative procedure if the eigenvalues of the diagonal coefficient matrix are sufficiently large compared to the eigenvalues of the off-diagonal coefficient matrix. Following the idea of Chu [2], we modify the finite difference equation by adding a "viscosity" term which will have a diminishing effect on the finite difference equations as $h \to 0$, and yet will assure the convergence of an iterative method. Unfortunately, the method is not applicable to every arrangement of mesh points. In fact there are rather severe restrictions which must be met. The first requirement is that the difference in areas of adjacent mesh regions be sufficiently small. This cannot be readily done along an irregular boundary, however, unless the boundary is modified. A problem arises if the boundary is modified. The boundary condition is given by $Mu = (\mu - \beta)u = 0$ on $\partial\Omega$. We need to extend M to be defined in a neighborhood of the boundary. It is possible to extend M continuously in a neighborhood of the boundary. However, if the direction of the boundary changes, β changes drastically, and we have no assurance that μ will be positive definite. The second requirement then is that M can be extended continuously over a neighborhood of the boundary, in such a way that μ will have positive definite symmetric part along the approximating boundary.

Let Ω_h be an approximation to Ω. Ω_h will have to meet several requirements to be specified later. H_h will denote a set of mesh points associated with Ω_h and with maximum distance h between connected nodes, and $\overline{H}_h$ will denote $H_h \cup \{x_B\}$. The discrete inner product is given by

$$(u_h, v_h) = \sum_j A_j (u_h)_j \cdot (v_h)_j \tag{3.1}$$

with the A_j being the area of $P_j \subset \Omega_h$. Simiarly, the "boundary" inner product is changed so that the lengths, $L_{j,B}$, are the lengths along $\partial\Omega_h$.

We define now two new finite difference operators, $\overline{K}_h$ and $\overline{M}_h$, by

$$(\overline{K}_h u)_j = (K_h u)_j + \sum_k \sigma \frac{u_j - u_k}{l_{j,k}} + \sum_B \sigma \frac{u_j - u_B}{l_{j,B}} \tag{3.2}$$

$$(\overline{M}_h u)_{j,B} = (M_h u)_{j,B} - \frac{\sigma A_j}{L_{j,B} l_{j,B}} (u_j - u_B) \tag{3.3}$$

where σ is a positive number which must satisfy requirements to be specified later.

It will be useful to prove a slightly different version of the "second identity".

<u>Lemma 3.1</u> If K is symmetric positive, then

$$(u_h,\bar{K}_h u_h)_h + (u_h,\bar{M}_h u_h)_{B_h} = (u_h,Gu_h)_h + (u_h,\mu u_h)_{B_h} + \sum_{(j,k)} \frac{\sigma A_j}{l_{j,k}} (u_j - u_k)^2 \tag{3.4}$$

where $\sum_{(j,k)}$ indicates a sum over every (j,k) pair where x_j is connected to x_k.

<u>Proof</u>: Using the "second identity" for K_h and M_h, equation (2.22), we have

$$(u_h,\bar{K}_h u_h)_h + (u_h,\bar{M}_h u_h)_{B_h} = (u_h,Gu_h)_h + (u_h,\mu u_h)_{B_h}$$

$$+ \sum_j \sum_k \frac{\sigma A_j}{l_{j,k}} u_j \cdot (u_j - u_k) + \sum_j \sum_B \frac{\sigma A_j}{l_{j,B}} u_j \cdot (u_j - u_B)$$

$$- \sum_j \sum_B \frac{\sigma A_j}{l_{j,B}} u_j \cdot (u_j - u_B)$$

The last two terms cancel. For the other term we have

$$\sum_j \sum_k \frac{\sigma A_j}{l_{j,k}} u_j \cdot (u_j - u_k) = \sum_{(j,k)} \left[\frac{\sigma A_j}{l_{j,k}}\Big((u_j \cdot (u_j - u_k) + u_k \cdot (u_k - u_j)\Big)\right]$$

$$= \sum_{(j,k)} \frac{\sigma A_j}{l_{j,k}} (u_j - u_k)^2$$

which completes the proof.

Lemma 3.1 immediately assures the existence and uniqueness of a solution for the special finite difference scheme. Using $\overline{M}_h u_h = 0$ to eliminate u_B from $\overline{K}_h u_h = r_h f$, we obtain

$$\sum_k \left(L_{j,k}\beta_{j,k} - \frac{\sigma A_j}{l_{j,k}} I \right) u_k + \left(A_j G_j + \sum_k \frac{\sigma A_j}{l_{j,k}} I + \sum_B L_{j,B}\mu_{j,B} \right) u_j = A_j f_j \tag{3.5}$$

for all $x_j \epsilon H_h$

Let A be the matrix of coefficients of (3.5).

<u>Lemma 3.2</u> If K is symmetric positive, then

$$\overline{K}_h u_h = r_h f$$

$$\overline{M}_h u_h = 0$$

has a unique solution on H_h.

<u>Proof</u>: The hypothesis implies that

$$\langle u, Au \rangle = (u_h, \overline{K}_h u_h)_h + (u_h, \overline{M}_h u_h)_{B_h}$$

By Lemma 3.1 A has positive definite symmetric part, and hence is non-singular. Thus (3.5) defines u_h uniquely on H_h.

Also it will be noted that the "second identity" of Lemma 3.1 will give the same a priori bounds for $\|u_h\|_h$ and $\|u_h\|_{B_h}$ as given by (2.25) and (2.26).

3.2 Convergence of Special Finite Difference Scheme

We will now show that the special finite difference scheme converges to a smooth solution, under a number of hypotheses given in the theorem. The theorem also includes all the hypotheses needed to assure convergence of the iterative matrix solution.

Though quite a number of requirements are given, there are only two essential restrictions, namely, that the areas A_j must be nearly uniform, and that M can be specified on a modified boundary in such a way that μ remains positive definite.

Theorem 3.1 Suppose that $u \in C^2(\overline{\Omega})$ satisfies

$$Ku = f \quad \text{on} \quad \Omega$$

$$Mu = 0 \quad \text{on} \quad \partial\Omega$$

where K is symmetric positive. For any $h > 0$, let Ω_h be an approximation to Ω, and let H_h be a corresponding set of mesh points with maximum distance h between connected nodes, and also with $L_{j,k}$, $L_{j,B}$, and $|x - x_j|$ for $x \in P_j$ all less than h. It is assumed that the following hypotheses are satisfied:

(i) There exists $K_1 > 0$, independent of h, such that for every P_j we have $h^2/A_j < K_1$.

(ii) There exists $K_2 > 0$, independent of h, such that all P_j with any point at a distance greater than $K_2 h$ from $\partial\Omega$ are equal rectangles.

(iii) There exists $K_3 > 0$, independent of h, such that for all $x \in \partial\Omega_h$, the distance from x to $\partial\Omega$ is less than $K_3 h$.

(iv) There exists $K_4 > 0$, such that M can be extended so as to satisfy a uniform Lipschitz condition at all points at a distance less than K_4 from $\partial\Omega$.

(v) Ω_h is such that $\mu = M + \beta$ has positive definite symmetric part on $\partial\Omega_h$.

(vi) Let W be the set of points that are a distance less than K_4 from $\partial\Omega$. Then α, G, and f are all extended to be defined on $\Omega \bigcup W$ with $\alpha \epsilon C^2(\Omega \bigcup W)$ and G positive definite on $\Omega \bigcup W$.

(vii) There exists $K_5 > 0$, independent of h, such that all points, x_j, associated with a boundary polygon, P_j, are in the polygon, and at a sufficient distance, $\ell_{j,B}$, from any boundary node, x_B, of P_j so that $A_j \leq K_5 L_{j,B} \ell_{j,B}$.

(viii) Either $\Omega_h \subset \Omega$ or else u can be extended so that $u \epsilon C^2(\overline{\Omega}_h)$.

(ix) $\sigma > \eta K_1 \rho_B + d$, where $d > 0$ and $\rho_B = \sup\limits_{x \epsilon \Omega \bigcup W} \rho(n \cdot \alpha(x))$, where n is any unit vector and η is the maximum number of nodes connected to any one node.

(x) $|A_j/A_k - 1| \leq d\lambda_G(h')^2/(\eta^2\sigma^2 h)$, for all connected nodes, x_j and x_k, where λ_G is the smallest eigenvalue of G in $\overline{\Omega}_h$, and $h' = \min(\ell_{j,k})$.

(xi) The length of $\partial\Omega_h$ is uniformly bounded.

Let u_h be the unique solution to

$$\overline{K}_h u_h = r_h f$$
$$\overline{M}_h u_h = 0$$

then

$$\|u_h - r_h u\| = O(h^\nu) \text{ as } h \to 0, \text{ for any positive } \nu < 1/2$$

Proof: Letting $w_h = u_h - r_h u$, and using the "second identity," (3.4), we see that the inequality (2.31) is still valid for $\overline{K}_h$ and $\overline{M}_h$,

$$\|w_h\|_h^2 \le \frac{1}{\lambda_G} \left(\|w_h\|_h \; \|\overline{K}_h w_h\|_h + \|w_h\|_{B_h} \; \|\overline{M}_h w_h\|_{B_h}\right) \tag{3.6}$$

We have

$$\overline{K}_h w_h = \overline{K}_h u_h - \overline{K}_h r_h u = r_h f - \overline{K}_h r_h u = r_h f - K_h r_h u + K_h r_h u_h - \overline{K}_h r_h u,$$

hence

$$\|\overline{K}_h w_h\|_h \le \|r_h K u - K_h r_h u\|_h + \|K_h r_h u - \overline{K}_h r_h u\|_h \tag{3.7}$$

In checking the proof of Theorem 2.1 we see that $r_h K u - K_h r_h u$ is the same as $K_h w_h$ (Theorem 2.1), hence the bound of (2.49) holds for this term;

$$\|r_h K u - K_h r_h u\|_h = O(h^{1/2}) \tag{3.8}$$

For the other term we have

$$\|(\overline{K}_h - K_h) r_h u\|_h^2 = \sum_j A_j \sigma^2 \left(\sum_k \frac{u_j - u_k}{l_{j,k}} + \sum_B \frac{u_j - u_B}{l_{j,B}} \right)^2 \tag{3.9}$$

Let J_1 denote the set of subscripts for those P_j which are equal rectangles, and let J_2 denote the rest of the subscripts. When $j \in J_1$ we have only the term $\sum_k (u_j - u_k)/l_{j,k}$ to consider. Because of the rectangular arrangement of points we can use a Taylor series analysis to show that

$$\left| \sum_k \frac{u_j - u_k}{l_{j,k}} \right| = O(h)$$

so that

$$\sum_{j\in J_1} A_j\sigma^2\left(\sum_k \frac{u_j - u_k}{l_{j,k}}\right)^2 = O(h^2) \tag{3.10}$$

On the other hand, when $j\in J_2$ we cannot do as well. However, we note that both $(u_j - u_k)/l_{j,k}$ and $(u_j - u_B)/l_{j,k}$ are uniformly bounded since u has a bounded derivative. Also, by hypothesis (ii), $\sum_{j\in J_2} A_j = O(h)$, so that

$$\sum_{j\in J_2} A_j\sigma^2\left(\sum_k \frac{u_j - u_k}{l_{j,k}} + \sum_B \frac{u_j - u_B}{l_{j,B}}\right)^2 = O(h) \tag{3.11}$$

It is assumed, of course, that the number of nodes connected to any one node is bounded as $h\to 0$.

Now, using (3.10) and (3.11) in (3.9) we have

$$\|(\overline{K}_h - K_h)r_h u\|_h = O(h^{1/2}) \tag{3.12}$$

Taking this together with (3.8) in (3.7) finally

$$\|\overline{K}_h w_h\|_h = O(h^{1/2}) \tag{3.13}$$

It is necessary now to obtain a bound for $\|\overline{M}_h w_h\|_{B_h}$. Since $\overline{M}_h w_h = \overline{M}_h u_h - \overline{M}_h r_h u = -\overline{M}_h r_h u$, we have

$$\|\overline{M}_h w_h\|_{B_h} \leq \|M_h r_h u\|_{B_h} + \|(\overline{M}_h - M_h)r_h u\|_{B_h} \tag{3.14}$$

We have

$$\|M_h r_h u\|_B^2 = \sum_j \sum_B L_{j,B}\,(\mu_{j,B} - \beta_{j,B}(2u_B - u_j))^2$$

We can establish a bound, since

$$|\mu_{j,B} - \beta_{j,B}(2u_B - u_j)| \le |\mu_{j,B}(u_j - u_B)| + |(\mu_{j,B} - \beta_{j,B})u_B| + |\beta_{j,B}(u_j - u_B)|$$

The first and last term on the right are of order h, since u is differentiable and $\|\mu\|$ and $\|\beta\|$ are bounded. By hypothesis (iv) M satisfies a Lipschitz condition, and so does u. Since the distance from x_B to $\partial\Omega$ is less than K_3h by (iii) and $Mu = 0$ on $\partial\Omega$, we see that $|(\mu_{j,B} - \beta_{j,B})u_B| = O(h)$. Since, by (xi), $\sum_j \sum_B L_{j,B}$ is uniformly bounded, we have

$$\|M_h r_h u\|_{B_h} = O(h) \tag{3.15}$$

Also

$$\|(\overline{M}_h - M_h)r_h u\|_B^2 = \sum_j \sum_B L_{j,B}\left(\frac{A_j\sigma}{L_{j,B}l_{j,B}}(u_j - u_B)\right)^2$$

$$\le \sum_j \sum_B L_{j,B}K_5^2\sigma^2(u_j - u_B)^2, \text{ by (vii)}$$

$$= O(h^2) \tag{3.16}$$

This shows that

$$\|\overline{M}w_h\|_{B_h} = O(h) \tag{3.17}$$

We check now to see that $\|w_h\|_h$ and $\|w_h\|_{B_h}$ are bounded. We have, using the a priori bound for $\|u_h\|_h$,

$$\|w_h\|_h \le \|u_h\|_h + \|r_h u\|_h \le \frac{1}{\lambda_G}\|r_h f\|_h + \|r_h u\|_h \tag{3.18}$$

which must be bounded since f and u are. In the same manner, $\|w_h\|_{B_h}$ must be bounded. Using this fact together with (3.13) and (3.17) in (3.6) we have

$$\|w_h\|_h = O(h^{1/4}) \tag{3.19}$$

Using now (3.19) in (3.6) we get $\|w_h\|_h = O(h^{3/8})$ and by repeating the process as many times as needed we get

$$\|w_h\|_h = O(h^{\nu}), \text{ for any positive } \nu < 1/2 \tag{3.20}$$

3.3 Convergence of the Matrix Iterative Solution

For the iterative solution of the matrix equation $Au = f$ we will split A into a block diagonal part D, and off diagonal part B. (We will suppress the subscript h on the finite difference solution u_h.) Thus, from (3.5), the j^{th} block of D is an $r \times r$ matrix,

$$D_j = A_j G_j + \sum_k \frac{\sigma A_j}{l_{j,k}} I + \sum_B L_{j,B} \mu_{j,B}$$

and a typical block element of B is

$$B_{j,k} = L_{j,k} \beta_{j,k} - \frac{\sigma A_j}{l_{j,k}} I$$

and $A = D + B$. The iterative method is given by

$$u^{(i+1)} = -D^{-1} B u^{(i)} + D^{-1} f$$

where $u^{(0)}$ is arbitrary. The hypotheses of Theorem 3.1 assure the convergence of $u^{(i)}$ to u.

<u>Theorem 3.2</u> For any $h > 0$, let Ω_h and H_h satisfy the hypotheses of Theorem 3.1. Let $u^{(0)}$ be an arbitrary vector defined on H_h, and let $\{u^{(i)}\}_{i=0}^{\infty}$ be a sequence defined recursively by

$$u^{(i+1)} = -D^{-1}Bu^{(i)} + D^{-1}f$$

Then $\lim_{i\to\infty} u^{(i)} = u$, where $Au = f$.

<u>Proof</u> - By the contraction mapping theorem it is sufficient to show that $\|D^{-1}B\| < 1$ for some matrix norm. Let v be an arbitrary vector defined on H_h, and let $w = D^{-1}Bv$. Since $Dw = Bv$, we have

$$\langle w,Dw\rangle = \langle w,Bv\rangle$$

or

$$\sum_j w_j \cdot \left(A_j G_j + \sum_k \frac{\sigma A_j}{l_{j,k}} I + \sum_B L_{j,B}\mu_{j,B}\right) w_j$$

$$= \sum_j \sum_k w_j \cdot \left(L_{j,k}\beta_{j,k} - \frac{\sigma A_j}{l_{j,k}} I\right) v_k$$

$$\leq \frac{1}{2} \sum_j \sum_k w_j \cdot \left(\frac{\sigma A_j}{l_{j,k}} I - L_{j,k}\beta_{j,k}\right) w_j$$

$$+ \frac{1}{2} \sum_j \sum_k v_k \cdot \left(\frac{\sigma A_j}{l_{j,k}} I - L_{j,k}\beta_{j,k}\right) v_k \qquad (3.21)$$

This last inequality follows from the fact that

$$\langle w,Hv\rangle \leq \frac{1}{2} \langle w,Hw\rangle + \frac{1}{2} \langle v,Hv\rangle$$

for any positive definite Hermitian matrix. We see that $(\sigma A_j)/(l_{j,k})\, I - L_{j,k}\beta_{j,k}$ is positive definite, since

$$\frac{\sigma A_j}{l_{j,k}} \geq \frac{\sigma A_j}{h} > \frac{\sigma}{K_1}\, h > h\rho(\beta_{j,k}) \geq L_{j,k}\rho(\beta_{j,k}) \qquad (3.22)$$

by (i) and (ix). By rearranging the terms of (3.21) so as to have all the w terms on the left and all the v terms to the right, we obtain

$$\sum_j w_j \cdot \left(A_j G_j + \sum_B L_{j,B}\mu_{j,B}\right) w_j + \frac{1}{2}\sum_j \sum_k w_j \cdot \left(\frac{\sigma A_j}{l_{j,k}}\, I + L_{j,k}\beta_{j,k}\right) w_j$$

$$\leq \frac{1}{2}\sum_j \sum_k v_j \cdot \left(\frac{\sigma A_k}{l_{j,k}}\, I + L_{j,k}\beta_{j,k}\right) v_j \qquad (3.23)$$

The last expression was obtained by interchanging j and k, since

$$l_{j,k} = l_{k,j},\quad L_{j,k} = L_{k,j},\quad \text{and} \quad \beta_{j,k} = -\beta_{k,j}$$

We can write (3.23) in the following form.

$$\sum_j w_j \cdot \left(A_j G_j + \sum_B L_{j,B}\mu_{j,B}\right) w_j + \frac{1}{2}\sum_j \sum_k w_j \cdot \left(\frac{\sigma A_j}{l_{j,k}}\, I + L_{j,k}\beta_{j,k}\right) w_j$$

$$\leq \frac{1}{2}\sum_j \sum_k v_j \cdot \left(\frac{\sigma A_j}{l_{j,k}}\, I + L_{j,k}\beta_{j,k}\right) v_j$$

$$+ \frac{1}{2}\sum_j \sum_k \frac{\sigma}{l_{j,k}}\, (A_k - A_j) v_j^2 \qquad (3.24)$$

or

$$\langle w,Xw\rangle + \langle w,Yw\rangle \leq \langle v,Yv\rangle + \langle v,Zv\rangle \tag{3.25}$$

where X, Y, and Z are matrices defined by (3.24).

We have already shown that Y is positive definite (using (3.22)); hence we can define a norm by

$$\|v\|_Y^2 = \langle v,Yv\rangle \tag{3.26}$$

We will show that $D^{-1}B$ is a strict contraction in the Y norm. First we will need some inequalities. We have

$$\langle w,Xw\rangle > \lambda_G\|w\|_h^2 \tag{3.27}$$

Next, we have

$$\sum_k\left(\frac{\sigma}{l_{j,k}} + \frac{L_{j,k}}{A_j}\rho(\beta_{j,k})\right) \leq \frac{\eta\sigma}{h'} + \frac{\eta K_1}{h}\left(\frac{\sigma}{K_1}\right) \leq 2\eta\,\frac{\sigma}{h'}$$

by (i) and (ix), so that

$$\langle w,Yw\rangle \leq \frac{\eta\sigma}{h'}\|w\|_h^2 \tag{3.28}$$

Also $\langle w,Yw\rangle$ can be bounded below, since

$$\rho\left(\sum_k\left[\frac{\sigma}{l_{j,k}} + \frac{L_{j,k}}{A_j}\beta_{j,k}\right]\right) \geq \frac{\sigma}{h} - \rho\left(\sum_k\frac{L_{j,k}}{A_j}\beta_{j,k}\right) \geq \frac{\sigma}{h} - \frac{\eta K_1}{h}\rho_B \geq \frac{d}{h}$$

by (i) and (ix). Thus, we have

$$\langle v,Yv\rangle \geq \frac{d}{2h}\|v\|_h^2 \tag{3.29}$$

Finally, since

$$\left|\frac{\sigma(A_k - A_j)}{A_j l_{j,k}}\right| \leq \left|\frac{A_k}{A_j} - 1\right|\frac{\sigma}{h'}$$

we have

$$\langle v, Zv\rangle \le \Lambda \frac{\eta\sigma}{2h'} \|v\|_h^2 \tag{3.30}$$

where $\Lambda = \max |A_k/A_j - 1|$, for all connected nodes, x_j and x_k. From the definition (3.26), and using (3.27) and (3.28) we have

$$\langle w, Xw\rangle + \langle w, Yw\rangle > \left(1 + \frac{\lambda_G h'}{\eta\sigma}\right) \|w\|_Y^2 \tag{3.31}$$

On the other hand from (3.29) and (3.30)

$$\langle v, Yv\rangle + \langle v, Zv\rangle \le \left[1 + \frac{\eta\sigma\Lambda}{d}\left(\frac{h}{h'}\right)\right] \|v\|_Y^2 \tag{3.32}$$

Substituting (3.31) and (3.32) in (3.25) we have

$$\|w\|_Y^2 < \left(\frac{1 + \frac{\eta\sigma\Lambda}{d}\left(\frac{h}{h'}\right)}{1 + \frac{\lambda_G h'}{\eta\sigma}}\right) \|v\|_Y^2 \tag{3.33}$$

Since $w = D^{-1}Bv$, and v is arbitrary, we see that $\|D^{-1}B\|_Y < 1$ since

$$\Lambda < \frac{d\lambda_G h'}{\eta^2\sigma^2}\left(\frac{h'}{h}\right) \tag{3.34}$$

by hypothesis (x). This completes the proof of Theorem 3.2.

Of course, if Ω_h can be selected so that all the A_j are equal, then hypothesis (x) is satisfied, and

$$\|D^{-1}B\|_Y < \frac{1}{\left(1 + \frac{\lambda_G h'}{\eta\sigma}\right)^{1/2}} \tag{3.35}$$

In the special case where all the P_j are equal rectangles, $\eta = 4$, so that

$$\|D^{-1}B\|_Y < \frac{1}{\left(1 + \frac{\lambda_G h'}{4\sigma}\right)^{1/2}} \tag{3.36}$$

CHAPTER IV

APPLICATION TO THE TRICOMI EQUATION

4.1 Transonic Gas Dynamics Problem

An example of a problem of physical significance can be drawn from the field of gas dynamics. A stream function is introduced such that derivatives of the stream function are velocities of the gas. The stream function satisfies a second order partial differential equation which is elliptic where the flow is subsonic, and hyperbolic where the flow is supersonic. The equation is of mixed type, then, for a transonic flow problem. When the equation is linearized by means of a hodagraph transformation, and after a further transformation, the Tricomi equation results,

$$F(y)\psi_{xx} - \psi_{yy} = 0 \tag{4.1}$$

where $F(y)$ is a continuous monotone function such that $yF(y) > 0$ for $y \neq 0$. Details of the derivation of (4.1) are given by Bers [11]. A solution for (4.1) for a region Ω which includes a portion of the x-axis is determined by proper boundary value data along portions of $\partial\Omega$. The proper boundary value data is known only for special cases. Usually the appropriate boundary data is the value of the function over part of the boundary. If the boundary data is sufficiently smooth, we can transform the homogeneous equation with non-homogeneous boundary conditions to a non-homogeneous equation satisfying homogeneous boundary conditions. The problem can be stated in the following form then,

$$F(y)\varphi_{xx} - \varphi_{yy} = f_1(x,y) \quad \text{on} \quad \Omega \tag{4.2}$$

$$a\frac{\partial\varphi}{\partial s} + b\frac{\partial\varphi}{\partial n} = 0 \quad \text{on} \quad \partial\Omega$$

where a or b or both may be zero on part of $\partial\Omega$. It is possible, also, that the stronger condition $\partial\varphi/\partial s = \partial\varphi/\partial n = 0$ may be imposed on some portion of $\partial\Omega$.

4.2 Tricomi Equation in Symmetric Positive Form

It is desired to express (4.2) as a system of first order differential equations. Following Friedrichs, we can do this by letting $u_1 = \varphi_x$, $u_2 = \varphi_y$, and $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$. Using the compatibility condition, $\varphi_{yx} = \varphi_{xy}$, we have

$$\begin{pmatrix} F(y) & 0 \\ 0 & 1 \end{pmatrix} \frac{\partial u}{\partial x} + \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \frac{\partial u}{\partial y} = \begin{pmatrix} f_1 \\ 0 \end{pmatrix} \tag{4.3}$$

This equation is symmetric, but not positive, since $G = 0$. To make (4.3) symmetric positive, we can multiply by a 2×2 matrix, B. In order to keep the coefficient matrices of $\partial u/\partial x$ and $\partial u/\partial y$ symmetric, B must be of the form

$$B = \begin{pmatrix} b & cF(y) \\ c & b \end{pmatrix} \tag{4.4}$$

where b and c are functions of x and y.

Equation (4.3) can now be expressed in symmetric positive form by

$$Ku = f \quad \text{on} \quad \Omega \tag{4.5}$$

where

$$\left.\begin{aligned}
Ku &= \alpha^x u_x + (\alpha^x u)_x + \alpha^y u_y + (\alpha^y u)_y + Gu \\
\alpha^x &= \frac{1}{2}\begin{pmatrix} bF(y) & cF(y) \\ cF(y) & b \end{pmatrix} \\
\alpha^y &= -\frac{1}{2}\begin{pmatrix} cF(y) & b \\ b & c \end{pmatrix} \\
G &= -\alpha^x_x - \alpha^y_y = \frac{1}{2}\begin{pmatrix} (c_y - b_x)F(y) + cF'(y) & b_y - c_x F(y) \\ b_y - c_x F(y) & c_y - b_x \end{pmatrix} \\
f &= \begin{pmatrix} bf_1 \\ cf_1 \end{pmatrix}
\end{aligned}\right\} \tag{4.6}$$

For the proper choice of functions b and c, G will be positive definite, resulting in symmetric positive K. The specific choice of b and c depends on the shape of Ω and on the boundary conditions which are specified.

It is possible that B may be singular in Ω, however, this does no harm if B is singular only along a line.

4.3 Admissible Boundary Conditions

Let $n = (n_x, n_y)$ be the outer normal along $\partial\Omega$. Then $\beta = n \cdot \alpha$ or

$$\beta = \frac{1}{2}\begin{pmatrix} F(y)(n_x b - n_y c) & n_x c F(y) - n_y b \\ n_x c F(y) - n_y b & n_x b - n_y c \end{pmatrix} \tag{4.7}$$

Friedrichs [1] noted that the quadratic form $u \cdot \beta u$ may be written

$$u \cdot \beta u = \frac{(b^2 - c^2 F(y))(n_y u_1 - n_x u_2)^2 - (n_y^2 - F(y) n_x^2)(b u_1 + c u_2)^2}{2(b n_x + c n_y)} \tag{4.8}$$

From this we can easily specify the boundary matrix, μ, so that admissible boundary conditions result. Let μ be defined so that

$$u \cdot \mu u = \frac{|b^2 - c^2 F(y)|(n_y u_1 - n_x u_2)^2 + |n_y^2 - F(y) n_x^2|(b u_1 + c u_2)^2}{2|b n_x + c n_y|} \tag{4.9}$$

Of course μ is non-negative definite. Also, $\eta(\mu - \beta) \oplus \eta(\mu + \beta) = R^2$, so that the boundary condition $Mu = (\mu - \beta)\, u = 0$ is admissible. Thus we can always obtain admissible boundary conditions. However, since b and c can be chosen subject only to the constraint that G is positive definite, we can have a wide variety of possible boundary conditions.

The actual boundary conditions for φ are determined by the signs of $b n_x + c n_y$, $b^2 - c^2 F(y)$, and $n_y^2 - F(y) n_x^2$. For the elliptic part of Ω, $y < 0$, we have $F(y) < 0$; hence both $b^2 - c^2 F(y) > 0$, and $n_y^2 - F(y) n_x^2 > 0$, so that the boundary condition is determined solely by the sign of $b n_x + c n_y$. Suppose that $b n_x + c n_y \lessgtr 0$. Then

$$u \cdot Mu = u \cdot (\mu - \beta) u = \frac{|b^2 - c^2 F(y)|(n_y u_1 - n_x u_2)^2}{|b n_x + c n_y|} \tag{4.10}$$

so that $Mu = 0$ implies $n_y u_1 = n_x u_2$. In terms of φ this means that $\varphi_x/\varphi_y = n_x/n_y = -(dy/ds)/(dx/ds)$, where $y = y(s)$, $x = x(s)$ are boundary coordinates as a function of arc length, s, along $\partial\Omega$. Hence

$$\frac{d\varphi}{ds} = \varphi_x \frac{dx}{ds} + \varphi_y \frac{dy}{ds} = 0 \tag{4.11}$$

so that φ is constant along $\partial\Omega$ when $bn_x + cn_y < 0$ in the elliptic region. On the other hand, if $bn_x + cn_y > 0$, we have

$$u \cdot Mu = \frac{|n_y^2 - F(y)n_x^2|(bu_1 + cu_2)^2}{|bn_x + cn_y|} \tag{4.12}$$

so that $Mu = 0$ implies that $bu_1 + cu_2 = 0$, so that $d\varphi/dp = 0$, where p is in some non-tangential direction. Thus, for the elliptic region we generally have a single boundary condition corresponding to the usual elliptic type of boundary condition.

In the hyperbolic region the boundary conditions depend on whether the magnitude of the boundary slope is greater than, less than, or equal to the magnitude of the slope of the characteristic curve. For equation (4.2), the characteristics satisfy the differential equation

$$\frac{dy}{dx} = \pm \frac{1}{\sqrt{F(y)}} \tag{4.13}$$

Suppose that the boundary is tangent to a characteristic, then

$$\frac{n_x}{n_y} = -\frac{dy}{dx}$$

so that

$$n_y^2 = F(y)n_x^2 \tag{4.14}$$

Suppose that a portion of $\partial\Omega$ is a left running characteristic, so that $n_y = \sqrt{F(y)}\, n_x$. Then, from (4.8)

$$u \cdot \beta u = \frac{(b + c\sqrt{F(y)})(b - c\sqrt{F(y)})(n_y u_1 - n_x u_2)^2}{2(b + c\sqrt{F(y)})n_x}$$

$$= \frac{(b - c\sqrt{F(y)})(n_y u_1 - n_x u_2)^2}{2n_x} \tag{4.15}$$

Suppose that this portion of $\partial\Omega$ is a right boundary, so that $n_x > 0$, then, if $b < c\sqrt{F(y)}$, equation (4.11) holds and $d\varphi/ds = 0$, but if $b > c\sqrt{F(y)}$, $\mu = \beta$ and no boundary condition is imposed. Similarly, along a right running characteristic, the same types of boundary conditions are determined by the sign of $b + c\sqrt{F(y)}$.

If the boundary is not characteristic, we have $n_y^2 > F(y)n_x^2$ if the magnitude of the boundary slope is less than the magnitude of the characteristic slope, and vice versa. Thus, the particular boundary condition is determined by the signs of all three terms, $n_y^2 - F(y)n_x^2$, $b^2 - c^2F(y)$, and $bn_x + cn_y$.

The various boundary conditions implied by a choice of b and c are summarized in Table I. It can be seen that there is considerable choice in the type of boundary which can be specified by the proper choice of the functions b and c.

TABLE I. - SUMMARY OF BOUNDARY CONDITIONS

Elliptic part of $\Omega (y < 0)$	
Boundary condition	Condition on b and c
$\frac{d\varphi}{ds} = 0$	$bn_x + cn_y < 0$
$\frac{d\varphi}{dp} = 0$	$bn_x + cn_y > 0$

Hyperbolic part of $\hat{\Omega} (y > 0)$		
Boundary condition	Type of boundary	Conditions on b and c
$\frac{d\varphi}{ds} = 0$	$n_y = \sqrt{F(y)}n_x, n_x > 0$	$b \gtrless c\sqrt{F(y)}$
none	$n_y = \sqrt{F(y)}n_x, n_x > 0$	$b > c\sqrt{F(y)}$
none	$n_y = -\sqrt{F(y)}n_x, n_x < 0$	$b < -c\sqrt{F(y)}$
$\frac{d\varphi}{ds} = 0$	$n_y = -\sqrt{F(y)}n_x, n_x < 0$	$b > -c\sqrt{F(y)}$
$\frac{d\varphi}{dp} = 0$	$n_y^2 > F(y)n_x^2$	$b^2 > c^2F(y)$ and $bn_x + cn_y > 0$
$\frac{d\varphi}{ds} = 0$	$n_y^2 > F(y)n_x^2$	$b^2 > c^2F(y)$ and $bn_x + cn_y < 0$
$\frac{d\varphi}{ds} = \frac{d\varphi}{dn} = 0$	$n_y^2 > F(y)n_x^2$	$b^2 < c^2F(y)$ and $bn_x + cn_y > 0$
none	$n_y^2 > F(y)n_x^2$	$b^2 < c^2F(y)$ and $bn_x + cn_y < 0$
none	$n_y^2 < F(y)n_x^2$	$b^2 > c^2F(y)$ and $bn_x + cn_y > 0$
$\frac{d\varphi}{ds} = \frac{d\varphi}{dn} = 0$	$n_y^2 < F(y)n_x^2$	$b^2 > c^2F(y)$ and $bn_x + cn_y < 0$
$\frac{d\varphi}{ds} = 0$	$n_y^2 < F(y)n_x^2$	$b^2 < c^2F(y)$ and $bn_x + cn_y > 0$
$\frac{d\varphi}{dp} = 0$	$n_y^2 < F(y)n_x^2$	$b^2 < c^2F(y)$ and $bn_x + cn_y < 0$

Note: p denotes the distance in some non-tangential direction.

4.4 Sample Problem

A simple choice of b and c which will result in G being positive definite in Ω, if $F'(y) > 0$, is

$$\left.\begin{aligned} b &= -b_0 - b_1 x,\ b_1 > 0 \\ c &= c_0, \text{ where } c_0 > -\frac{b_1 F(y)}{F'(y)} \text{ in } \Omega \end{aligned}\right\} \tag{4.16}$$

Then

$$G = \frac{1}{2}\begin{pmatrix} b_1 F(y) + c_0 F'(y) & 0 \\ 0 & b_1 \end{pmatrix} \tag{4.17}$$

which is obviously positive definite.

To show the type of boundary conditions which may result, consider the case $F(y) = y$, so that

$$G = \frac{1}{2}\begin{pmatrix} b_1 y + c_0 & 0 \\ 0 & b_1 \end{pmatrix} \tag{4.18}$$

The characteristics in this case satisfy one of the equations

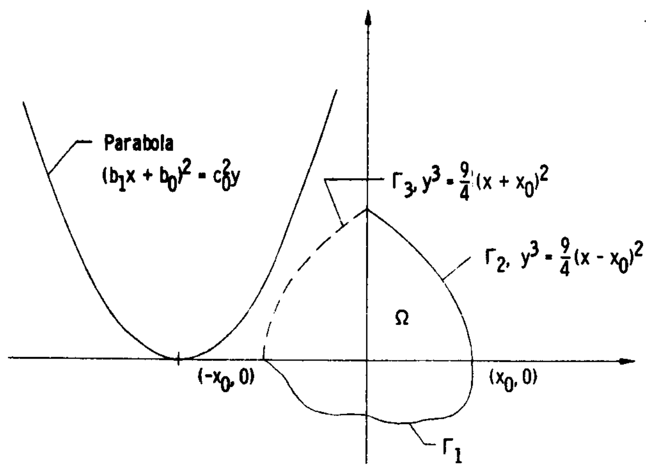
$$\frac{dy}{dx} = \pm \frac{1}{\sqrt{y}}$$

which can be solved to obtain the characteristic equation,

$$y^3 = \frac{9}{4}(x - x_0)^2 \tag{4.19}$$

where x_0 is the point on the x-axis intersected by the characteristic.

As an illustration, suppose that Ω is the region shown in figure 2, which is bounded by two characteristics in the hyperbolic region and by a curve satisfying $n_y < [(b_0 + b_1 x)/c_0]\, n_x$



	Type of boundary	Boundary condition
Γ_1	$n_y < \frac{(b_0 + b_1x)}{c_0} n_x$	$\frac{d\varphi}{ds} = 0$
Γ_2	$n_y = \sqrt{y}\, n_x$	$\frac{d\varphi}{ds} = 0$
Γ_3	$n_y = -\sqrt{y}\, n_x$	None

Figure 2. - Region, Ω, for a Tricomi problem.

in the elliptic region. It is assumed that b_0/b_1 is chosen large enough so that the parabola $(b_1x + b_0)^2 = c_0^2y$ lies entirely to the left of Ω, as indicated in figure 2. The boundary condition is $d\varphi/ds = 0$ for the elliptic portion, Γ_1, of $\partial\Omega$, and for one characteristic, Γ_2, with no boundary condition on the other characteristic, Γ_3. This is known as a Tricomi problem. Variations are possible with Γ_2 and Γ_3 replaced by several characteristics. This type of problem is discussed by Bers [11], p. 88.

It is worthwhile noting that the solution obtained by the finite difference solution of the symmetric positive form of the Tricomi equation consists of derivatives of the stream function, which corresponds to velocities in the physical problem. Hence, even though we have a convergence rate which is less than $O(h^{1/2})$, it is essentially equivalent to a convergence rate of $O(h^{3/2})$ if the original second order equation were solved directly for the stream function.

CHAPTER V

A NUMERICAL EXAMPLE

5.1 Description of Problem

A numerical solution to a Tricomi equation was calculated using the finite difference scheme of Chapter II. The accuracy of the solution was checked by using a problem for which an analytical solution is known.

The Tricomi equation can be put in symmetric positive form as indicated in the last chapter, as given by equations (4.5) and (4.6). The region Ω chosen is indicated in figure 3. The choice of b and c are

$$\left.\begin{aligned} b &= -3 - x \\ c &= 2 \end{aligned}\right\} \tag{5.1}$$

which gives $b_x = -1$, and $b_y = c_x = c_y = 0$. We choose $F(y) = y$. Using this in (4.6) we have

$$G = \begin{pmatrix} 1 + \frac{y}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \tag{5.2}$$

which is positive definite in Ω.

We now check to see what the admissible boundary conditions are from Table I. For the hyperbolic part of $\Omega (y > 0)$, we need

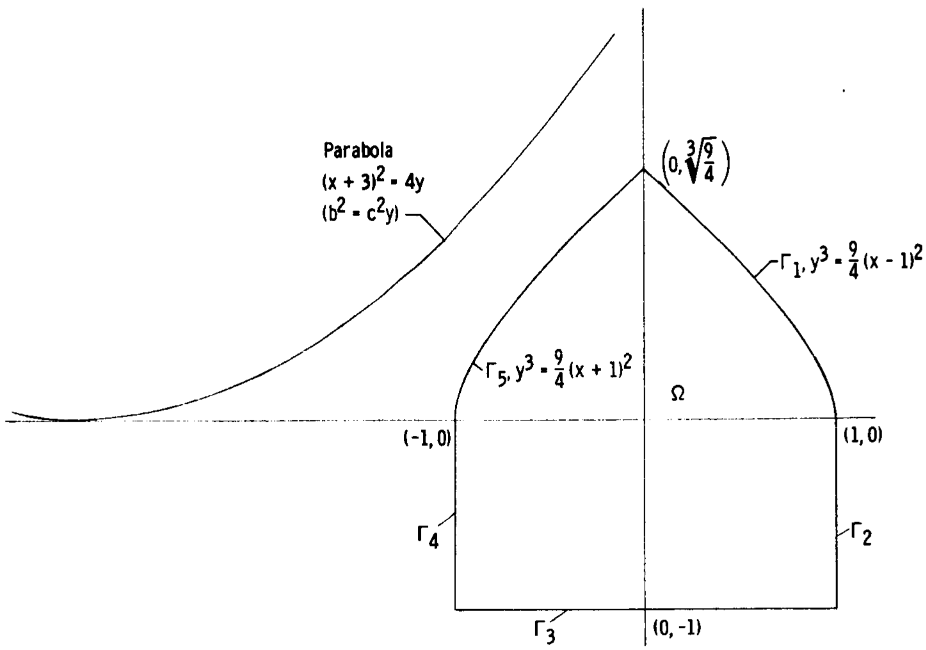


Figure 3. - Region for numerical example.

to know the sign of $b^2 - c^2y = (x + 3)^2 - 4y$. From figure 3 we see that $b^2 > c^2y$ in Ω, hence, since $b < 0$ in Ω, we have $\pm c\sqrt{y} < -b$ in Ω. From Table I, we have $d\varphi/ds = 0$ on Γ_1 and no boundary conditions on Γ_5. For the elliptic part of $\Omega(y < 0)$ we need to check the sign of $bn_x + cn_y$. Along Γ_2 we have $n_x = 1$, $n_y = 0$, so that $bn_x + cn_y = -(x + 3) < 0$ along Γ_2. Hence, the admissible boundary condition along Γ_2 is $d\varphi/ds = 0$. Next we check Γ_3. Then $n_x = 0$ and $n_y = -1$, and $bn_x + cn_y = -2 < 0$, so that $d\varphi/ds = 0$ along Γ_3. Finally, along Γ_4, $n_x = -1$, $n_y = 0$, giving $bn_x + cn_y = x + 3 = 2 > 0$, since $x = -1$. Hence $d\varphi/dp = 0$ along Γ_4, where p is some non-tangential direction. To find the specific direction, we go back to equation (4.12) which holds in this case. We see that $Mu = 0$ implies that $bu_1 + cu_2 = -2\varphi_x + 2\varphi_y = 0$ or $\varphi_x = \varphi_y$. Hence p is in a direction sloping downward at 45°. We summarize the boundary conditions:

	Boundary Condition
Γ_1	$\frac{d\varphi}{ds} = 0$
Γ_2	$\frac{d\varphi}{ds} = 0$
Γ_3	$\frac{d\varphi}{ds} = 0$
Γ_4	$\varphi_x = \varphi_y$
Γ_5	None

A simple, but non-trivial function satisfying these boundary conditions is easily obtained by choosing a function which is zero along Γ_1, Γ_2, Γ_3, and Γ_4, with the normal derivative also zero along Γ_4. These requirements are met by

$$\varphi(x,y) = (x + 1)^2(x - 1)(y + 1)(4y^3 - 9(x - 1)^2) \tag{5.3}$$

The function f_1 is determined by calculating $y\varphi_{xx} - \varphi_{yy} = f_1$, which gives

$$\begin{aligned} f_1(x,y) = y(y+1)[(4y^3 - 9(x-1)^2(6x+2) - 18(x^2-1)(7x-1)] \\ - 24(x+1)^2(x-1)y(2y+1) \end{aligned} \tag{5.4}$$

The functions for which we are solving are then

$$\left.\begin{aligned} \varphi_x &= (x+1)(y+1)[(4y^3 - 9(x-1)^2)(3x-1) - 18(x+1)(x-1)^2] \\ \varphi_y &= (x+1)^2(x-1)[16y^3 + 12y^2 - 9(x-1)^2] \end{aligned}\right\} \tag{5.5}$$

and from (4.6) we have

$$\left.\begin{aligned} \alpha^x &= \begin{pmatrix} -\frac{x+3}{2}y & y \\ y & -\frac{x+3}{2} \end{pmatrix} \\ \alpha^y &= \begin{pmatrix} -y & \frac{x+3}{2} \\ \frac{x+3}{2} & -1 \end{pmatrix} \end{aligned}\right\} \tag{5.6}$$

$$f = \begin{pmatrix} -(x+3)f_1 \\ 2f_1 \end{pmatrix} \tag{5.7}$$

We need to evaluate the matrix μ along all boundaries, with μ defined by equation (4.9). A straightforward calculation gives the following values for μ.

Boundary segment	μ
Γ_1	$\frac{x + 3 + 2\sqrt{y}}{2\sqrt{1+y}} \begin{pmatrix} y & -\sqrt{y} \\ -\sqrt{y} & 1 \end{pmatrix}$
Γ_2	$\begin{pmatrix} -2y & y \\ y & 2 - y \end{pmatrix}$
Γ_3	$\frac{1}{2} \begin{pmatrix} (x+3)^2 + 2 & -(x+3) \\ -(x+3) & 2 \end{pmatrix}$
Γ_4	$\begin{pmatrix} -y & y \\ y & 1 - 2y \end{pmatrix}$
Γ_5	$\frac{x + 3 - 2\sqrt{y}}{2\sqrt{1+y}} \begin{pmatrix} y & \sqrt{y} \\ \sqrt{y} & 1 \end{pmatrix}$

This gives the information necessary to calculate the coefficients of the finite difference equation, which is

$$\sum_k L_{j,k}\beta_{j,k}u_k + \sum_B L_{j,B}\mu_{j,B}u_j + A_jG_ju_j = A_jf_j \tag{5.8}$$

Equation (5.8) holds for every mesh point, x_j, in the set of mesh points. For simplicity a uniform mesh was used, as indicated in figure 4. It will be noted that mesh points outside of Ω were used. A solution was calculated for two different mesh spacings, $h = 0.2$ and $h = 0.1$. The finite difference equation was solved in each case by the block tridiagonal algorithm mentioned in section 2.5. Since the analytical solution, u, is given by (5.5) we can calculate $\|u_h - r_hu\|_h$, as well as the maximum value over all mesh points of the maximum component of the error.

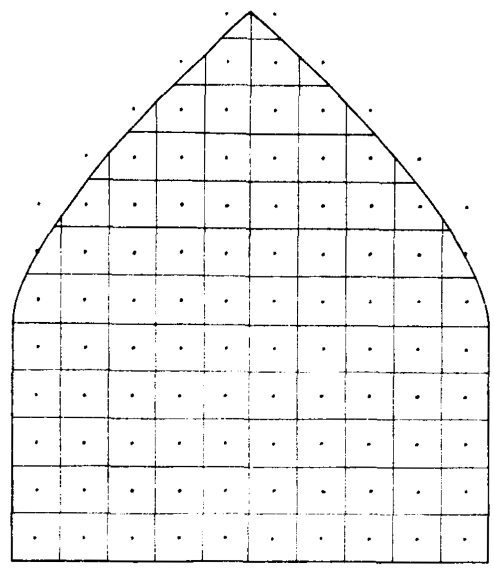
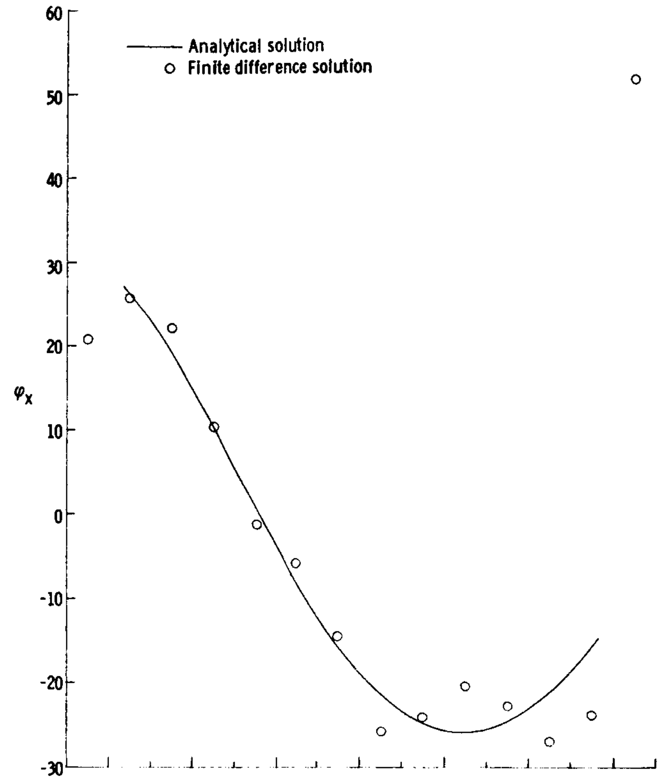


Figure 4. - Mesh point arrangement for numerical example.

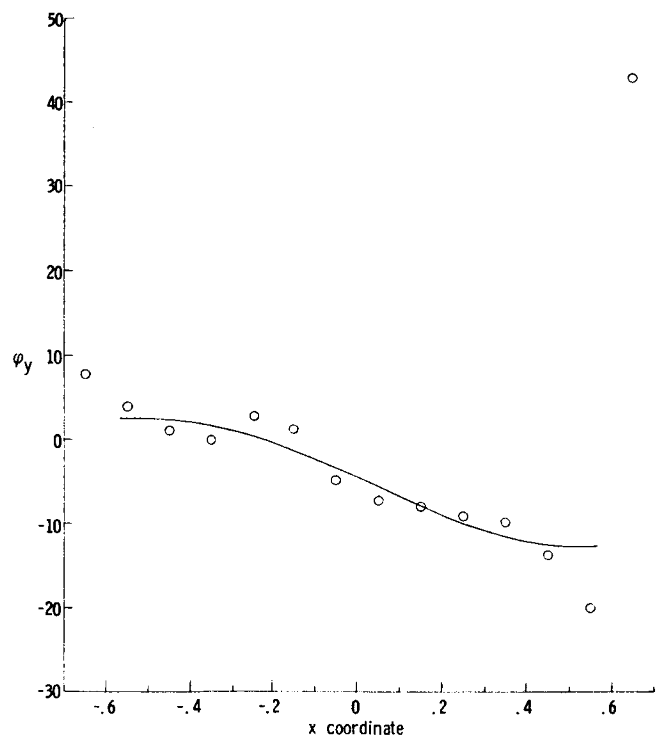
5.2 Description of Numerical Results

Theorem 2.1 assures us of essentially $O(h^{1/2})$ convergence in the L^2 norm. Unfortunately this does not assure us of pointwise convergence. As indicated in the proof of Theorem 2.1, the finite difference equations can be expected to be less accurate when the polygons, P_j, are not uniform rectangles. This was the case in the numerical example. The result was poor accuracy near the hyperbolic boundary segments, Γ_1 and Γ_5. In going from the coarse mesh (h = 0.2) to the fine mesh (h = 0.1), the L^2 error was reduced from 6.06 to 5.30 which is not unreasonable with the $O(h^{1/2})$ convergence rate. However, the maximum error actually increased from 33.5 to 60.9 indicating pointwise divergence. The horizontal line (y = 0.75) along which the finite difference solution for the finer mesh has the poorest agreement with the analytical solution is plotted in figure 5. It is seen that the finite difference solution has large oscillations with a "wild" point at the end of the line.

All this is not quite as bad as it seems, though, since L^2 convergence with pointwise divergence means that the divergent points will occur as sharp peaks. Therefore, it can be expected that a smoothing operation would give great improvement in the results. With this in mind, a simple smoothing procedure was tried. Since most smoothing methods are for one-dimensional functions, the solution was smoothed by lines, first along vertical



(a) φ_x.



(b) φ_y.

Figure 5. - Analytical and finite difference solutions for $y = 0.75$.

lines and then along horizontal lines. The method of smoothing used is similar to a method suggested by Hamming, p. 314, [12]. If it is desired to smooth equally spaced data, $\{y_k\}_{k=1}^{n}$, we can define the smoothed data, $\{\bar{y}\}_k {}_{k=1}^{n}$, by

$$\bar{y}_0 = \frac{y_0 + 2y_1 - y_2}{2}$$

$$\bar{y}_k = \frac{y_{k-1} + 2y_k + y_{k+1}}{4}, \text{ for } k = 2,3, \ldots, n-1$$

$$\bar{y}_n = \frac{-y_{n-2} + 2y_{n-1} + y_n}{2}$$

The result of applying this smoothing procedure to the solution based on the finer grid ($h = 0.1$) was to reduce the L^2 error from 5.30 to 2.07. The maximum error was reduced from 60.9 to 13.8. This maximum error was at a point lying outside of Ω, the maximum error for a mesh point in Ω was 6.4. The improvement obtained by this smoothing procedure is indicated by figure 6, which shows the horizontal line with poorest agreement after smoothing. The solution after smoothing along a more typical horizontal line is shown in figure 7.

It should be emphasized that the smoothing procedure used here was very simple and that most likely better results could be obtained with other smoothing methods. For example, Lanczos [13], gives several smoothing methods, both local and global (through the use of truncated Fourier series).

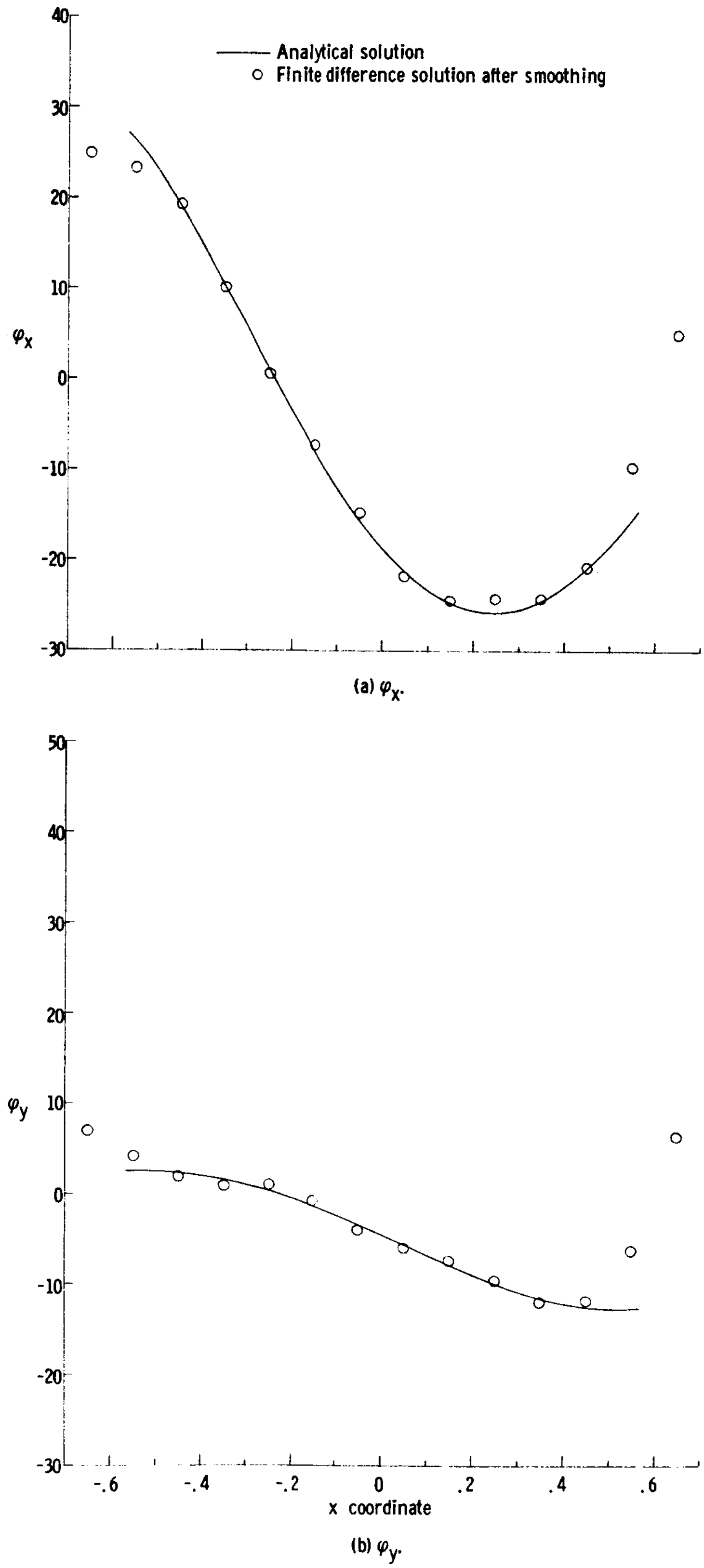
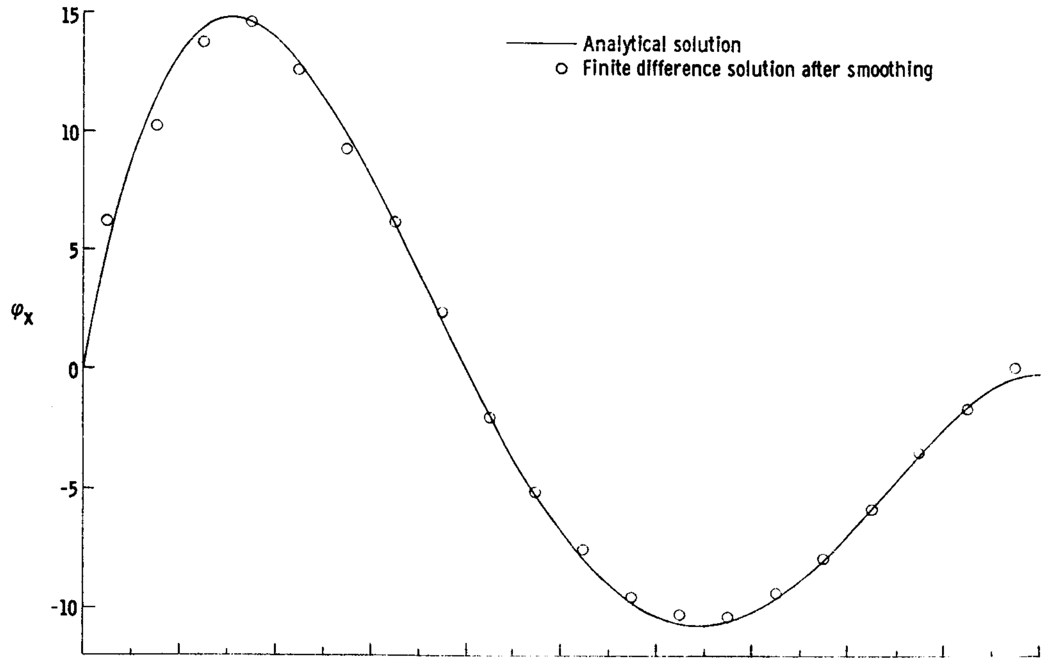
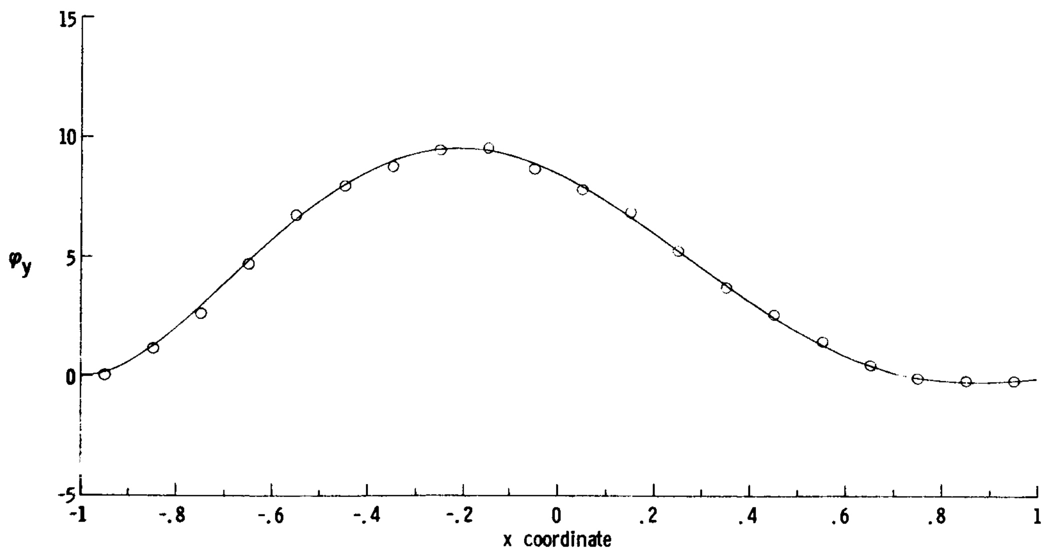


Figure 6. - Analytical and smoothed finite difference solutions for $y = 0.75$.



(a) φ_x.



(b) φ_y.

Figure 7. - Analytical and smoothed finite difference solutions for $y = -0.25$.

REFERENCES

1. Friedrichs, K. O., "Symmetric Positive Linear Differential Equations", Comm. Pure Appl. Math., Vol. 11, 1958, pp. 333-418.
2. Chu, C. K., "Type-Insensitive Finite Difference Schemes" Ph.D. Thesis, New York University, 1958.
3. Sarason, L., "On Weak and Strong Solutions of Boundary Value Problems", Comm. Pure Appl. Math., Vol. 15, 1962, pp. 237-288.
4. Lax, P. D., and Phillips, R. S., "Local Boundary Conditions for Dissipative Symmetric Linear Differential Operators", Comm. Pure Appl. Math., Vol. 13, 1960, pp. 427-455.
5. Phillips, R. S., and Sarason, L., "Singular Symmetric Positive First Order Differential Operators", J. Math. and Mech., Vol. 8, 1966, pp. 235-272.
6. Varga, R. S., "Matrix Iterative Analysis", Prentice-Hall, Inc., 1962.
7. MacNeal, R. H., "An Asymetrical Finite Difference Network", Quart. Appl. Math., Vol. 11, 1953, pp. 295-310.
8. Schecter, S., "Quasi-Tridiagonal Matrices and Type-Insensitive Difference Equations", Quart. Appl. Math., Vol. 18, 1960, pp. 285-295.

9. Cea, J., "Approximation Variationelle des Problemes aux Limites", Ann. Inst. Fourier, Vol. 14, 1964, pp. 345-444.

10. Taylor, A. E., "Introduction to Functional Analysis", John Wiley and Sons, Inc., 1958.

11. Bers. L., "Mathematical Aspects of Subsonic and Transonic Gas Dynamics", John Wiley and Sons, Inc., 1958.

12. Hamming, R. W., "Numerical Methods for Scientists and Engineers", McGraw-Hill Book Co., Inc., 1962.

13. Lanczos, C., "Applied Analysis", Prentice Hall, Inc., 1956.