MATHEMATICS OF COMPUTATION
VOLUME 62, NUMBER 205
JANUARY 1994, PAGES 105–118

A FINITE DIFFERENCE METHOD FOR SYMMETRIC POSITIVE DIFFERENTIAL EQUATIONS

JINN-LIANG LIU

ABSTRACT. A finite difference method is developed for solving symmetric positive differential equations in the sense of Friedrichs. The method is applicable to partial differential equations of mixed type with more general boundary conditions. The method is shown to have a convergence rate of $O(h^{1/2})$, h being the size of mesh grid. Some numerical results are presented for a model problem of forward-backward heat equations.

1. INTRODUCTION

In the theory of partial differential equations there is a fundamental distinction between those of elliptic, hyperbolic, and parabolic types. Friedrichs [5] developed a theory of symmetric positive linear differential equations independent of type. The theory of Friedrichs's systems has been shown to be very useful in theoretical analysis for mixed-type problems such as the Tricomi problem and forward-backward heat equations. Furthermore, it also gives a simple and unified numerical treatment for these problems (see e.g. [1, 7, 9]). Otherwise, if a numerical method applies directly to a PDE of mixed type, the treatment of the interface on which the PDE changes type is in general very difficult to handle. For example, Vanaja and Kellogg [12] used an iterative method to solve discrete approximations of a forward-backward heat equation which involve three different systems, i.e., forward, backward, and interface finite difference systems. The method requires the solution of the equation to be more regular than that of the unified method proposed here. The unified method, on the other hand, not only requires less regularity for the solution but also applies to a more general setting of the problem; by this we mean less restriction on the assumption of coefficient functions that cause the equation to change type.

Several numerical methods have been developed for Friedrichs's systems [7, 9, 10]. Friedrichs [5] was the first to propose a finite difference procedure for the numerical solutions of symmetric positive systems in rectangular regions. Chu [4] further studied this method and extended it to curvilinear rectangular domains, but the rate of convergence was not established. Katsanis [7] gave a finite difference method for the Tricomi problem, using symmetric positive systems, which is applicable to any region with piecewise smooth boundaries,

Received by the editor April 16, 1992 and, in revised form, November 24, 1992 and February 16, 1993.

1991 *Mathematics Subject Classification.* Primary 65N30, 35K20.

Key words and phrases. Finite difference method, Friedrichs's positive systems, error estimates.

This work was supported in part by NSC-grant 81-0208-M-009-513, Taiwan, R.O.C.

0025-5718/94 $1.00 + $.25 per page

and showed that the rate of convergence is $O(h^{1/2})$, where h is the size of a regular mesh. However, he imposed on the system an extra constraint that the boundary matrix should be positive definite. This appears to be somewhat restrictive for the application of Friedrichs's theory. In fact, this work is motivated by forward-backward heat equations which do not reduce to symmetric positive systems having this property, and we find that the constraint is actually not required for our method. We show that for a rectangular domain the rate of convergence is also $O(h^{1/2})$. The main difference between our method and the method of Katsanis is that the approximation in Katsanis's scheme does not necessarily go beyond the boundary, whereas our method does. The boundary condition of the system is handled differently as well. It is shown in [4] that there exists a transformation of the dependent variables, coefficient matrices, the differential equations, and the boundary conditions such that when the domain is mapped from a curvilinear rectangle to a rectangle, the symmetric positive character of the equation is preserved. We confine our considerations to rectangular domains.

Since the development of the proposed method is primarily motivated by forward-backward heat equations, we stress further the main results of both the iterative method in [12] and our method. For problems in the x-y plane, if the solution has continuous derivatives of order 4 in x and order 2 in y, the rate of convergence of the discretization error for the iterative method is $O(h^2+k)$, where h and k are mesh sizes in x and y, respectively. The iterative process may be affected by different h and k and hence by the interface system. Our method instead requires the solution to have smoothness of order 3 in x and 2 in y, and gives an $O(h^{1/2})$ convergence. Since the solution is obtained by reducing the original second-order equation into a first-order system, it is essentially equivalent to a convergence rate of $O(h^{3/2})$, at least in the x-direction if the original equation were solved directly for the unknown function.

2. Symmetric positive systems

Let Ω be a bounded open set in $\mathbf{R}^m$, with a piecewise continuously differentiable boundary $\partial\Omega$. A point in $\mathbf{R}^m$ is denoted by $x = (x_1, x_2, \ldots, x_m)$ and an unknown r-dimensional vector-valued function defined on Ω is given by $\mathbf{u} = (u_1, u_2, \ldots, u_r)$. Let $\alpha^1, \alpha^2, \ldots, \alpha^m$ be symmetric $r \times r$ matrix-valued functions, G an $r \times r$ matrix-valued function, and $\mathbf{f} = (f_1, f_2, \ldots, f_r)$ a given r-dimensional vector-valued function, all defined on Ω. 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

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

With this notation we can write the identity

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

simply as

$$\nabla \cdot (\alpha \mathbf{u}) = (\nabla \cdot \alpha)\mathbf{u} + \alpha \cdot \nabla \mathbf{u}. \tag{2.1}$$

A symmetric system of linear differential equations with boundary conditions can then be written in the following generic form:

$$K\mathbf{u} := \alpha \cdot \nabla \mathbf{u} + \nabla \cdot (\alpha \mathbf{u}) + G\mathbf{u} = \mathbf{f}(x) \quad \text{in } \Omega, \tag{2.2}$$

$$M\mathbf{u} := (\mu(x) - \beta(x))\mathbf{u}(x) = 0 \quad \text{on } \partial\Omega. \tag{2.3}$$

The matrix β is defined (almost everywhere) on $\partial\Omega$ by $\beta = \mathbf{n} \cdot \alpha$, where $\mathbf{n} = (n_1, \ldots, n_m)$ is the outer normal on $\partial\Omega$. The matrix μ is defined on $\partial\Omega$ so that the boundary condition (2.3) is *admissible* and the operator K is *positive* in the sense of Friedrichs [5], i.e.,

(i) $\frac{1}{2}(\mu(x) + \mu^*(x))$ is positive semidefinite on $\partial\Omega$,
(ii) $\ker(\mu - \beta) \oplus \ker(\mu + \beta) = \mathbf{R}^r$ on $\partial\Omega$, and
(iii) $\frac{1}{2}(G + G^*)$ is positive definite in $\overline{\Omega}$.

The adjoint operators K^* and M^* of K and M, respectively, are formally defined by

$$K^*\mathbf{v} = -\alpha \cdot \nabla \mathbf{v} - \nabla \cdot (\alpha \mathbf{v}) + G^*\mathbf{v},$$
$$M^*\mathbf{v} = (\mu^* + \beta)\mathbf{v}.$$

Let $(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{u} \cdot \mathbf{v}\, dx$, and $\langle \mathbf{u}, \mathbf{v} \rangle = \int_{\partial\Omega} \mathbf{u} \cdot \mathbf{v}\, ds$. Let $H^m(\Omega)$, $m \geq 0$, be the usual, in vector form, Hilbert spaces equipped with the norm $\| \cdot \|_m$. The following results are given in [5]:

Lemma 2.1 (First Identity). *If K is symmetric positive, then for any $\mathbf{u}, \mathbf{v} \in H^1(\Omega)$,*

$$(K\mathbf{u}, \mathbf{v}) + \langle M\mathbf{u}, \mathbf{v} \rangle = (\mathbf{u}, K^*\mathbf{v}) + \langle \mathbf{u}, M^*\mathbf{v} \rangle. \tag{2.4}$$

Lemma 2.2 (Second Identity). *If K is symmetric positive, then for any $\mathbf{u} \in H^1(\Omega)$,*

$$(K\mathbf{u}, \mathbf{u}) + \langle M\mathbf{u}, \mathbf{u} \rangle = (G\mathbf{u}, \mathbf{u}) + \langle \mu\mathbf{u}, \mathbf{u} \rangle. \tag{2.5}$$

Let $V = C^1(\Omega) \cap \{v : M^*v = 0 \text{ on } \partial\Omega\}$. We shall say that $\mathbf{u} \in L^2(\Omega)$ is a *weak solution* of problem (2.2), if for all $\mathbf{v} \in V$

$$(\mathbf{u}, K^*\mathbf{v}) = (\mathbf{f}, \mathbf{v}).$$

The existence of weak solutions of (2.2), (2.3) is guaranteed if M is semi-admissible. Uniqueness is insured if we look for solutions in $H^1(\Omega)$. If, in addition, M is admissible and a weak solution is continuously differentiable, then it must also be a classical solution. It follows from the First Identity (2.4) that a classical solution is also a weak solution [5].

3. Finite difference method

We describe a finite difference approximation of (2.2) and (2.3) for a rectangular domain in the x-y plane. Extension to rectangular domains in higher dimensions is immediate. Let Ω be the rectangle centered at the origin, with boundaries $x = x_-$, $x = x_+$, $y = y_-$, and $y = y_+$. Let Ω be partitioned in a square grid of width h; a grid point is denoted by a pair of integers (i, j) with $(-I - 2, -J - 2) \leq (i, j) \leq (I + 2, J + 2)$; the step h is selected so that I and J are even integers. The grid points with $|i| = I$ or $|j| = J$ are called

boundary points (including the four corner points); those with $I < |i| \leq I+2$ or $J < |j| \leq J+2$ are extensions of the domain beyond the boundary and shall be called extension points.

We introduce the shift operators

$$S^x u(i, j) = u(i+1, j), \quad S^y u(i, j) = u(i, j+1),$$
$$S^{-x} u(i, j) = u(i-1, j), \quad S^{-y} u(i, j) = u(i, j-1).$$

We have $S^{2x} \equiv S^x S^x$, $S^{-2x} \equiv S^{-x} S^{-x}$ and similarly for S^{2y} and S^{-2y}.

The boundary operators B^0 and B^1 are defined such that $B^0 u$ is the value of u on the boundary and $B^1 u$ is the value of u one row beyond (into extension) the boundary. The operator B^{-1} is defined so that $B^{-1} u$ is the value one row within (into interior) the boundary, and B^2, B^{-2} etc. are similarly defined. (E.g. $B^1 = S^x B^0$ at $x = x_+$, $B^1 = S^{-x} B^0$ at $x = x_-$, etc.)

For each interior and boundary point, we define the finite difference operator K_h, an approximation to the differential operator K:

$$K_h \mathbf{u} = \frac{(S^x \alpha^1) S^{2x} \mathbf{u} - (S^{-x} \alpha^1) S^{-2x} \mathbf{u}}{2h} + \frac{(S^y \alpha^2) S^{2y} \mathbf{u} - (S^{-y} \alpha^2) S^{-2y} \mathbf{u}}{2h} + G\mathbf{u}.$$

The difference equation is written for each even interior point and boundary point:

$$K_h \mathbf{u} = \mathbf{f}. \tag{3.1}$$

To each boundary point, we assign the approximate boundary condition

$$M_h \mathbf{u} = (B^1 \mu) B^0 \mathbf{u} - (B^1 \beta) B^2 \mathbf{u} = 0, \tag{3.2}$$

where $B^1 \beta$ is defined as $n_x B^1 \alpha^1 + n_y B^1 \alpha^2$ at each row grid B^1, n_x and n_y being well defined for each boundary. Note that, at the corner points, n_x and n_y are not unique but will be well defined if we consider that (3.2) is computed piecewise on the boundary, i.e., piece by piece parallel to row B^1. In fact, this will give a unique representation of each unknown into the extension in terms of the unknown on the boundary including the four corner points. We define $B^1 \mu$ as

$$B^1 \mu = B^1 \beta + M. \tag{3.3}$$

Clearly, the difference equation and the approximate boundary condition indeed approximate the differential equation and the given boundary condition; i.e., if $\mathbf{u}(x, y)$ is a continuously differentiable function defined in the closed rectangle, we see immediately that $K_h \mathbf{u}$ converges pointwise to $K\mathbf{u}$, and $M_h \mathbf{u}$ converges pointwise to $M\mathbf{u}$, as h tends to 0. The values of the function at the boundary, and even in the extension, are all solved for as unknowns. This differs from usual finite difference procedures for boundary value problems, where the values on the boundary are known data. The finite difference method of Katsanis [7] is based on the formulation obtained by applying Green's theorem in any region in Ω centered at each grid point. We shall see there exists a unique solution of (3.1), (3.2).

We define the inner product and the norm respectively by

$$(\mathbf{u}, \mathbf{v})_h = \sum_e \mathbf{u} \cdot \mathbf{v} 4h^2,$$

where $\sum_e$ indicates summation over all even points (i, j) with $|i| \le I$, $|j| \le J$, and $\|\mathbf{u}\|_h + (\mathbf{u}, \mathbf{u})_h^{1/2}$. These can be interpreted as ordinary Hilbert space inner product and norm, if the values of $\mathbf{u}$ defined at even grid points are interpreted as values of piecewise constant step functions, constant in each $2h \times 2h$ square centered around each grid point. Let $L_h^2(\Omega)$ denote the space of these discrete functions.

Similarly, we define $\langle \mathbf{u}, \mathbf{v} \rangle_h = \sum_{B_e} \mathbf{u} \cdot \mathbf{v} 2h$, where $\sum_{B_e}$ indicates summation over all even boundary points with $|i| = I$ or $|j| = J$, and $|\mathbf{u}|_h = \langle \mathbf{u}, \mathbf{u} \rangle_h^{1/2}$. The interpretation is completely similar.

From the definitions of $\|\cdot\|_h$ and $|\cdot|_h$, we have the inequality

$$|\mathbf{u}|_h \le \frac{1}{\sqrt{2}} h^{-1/2} \|\mathbf{u}\|_h. \tag{3.4}$$

Now, we define the consistent finite difference operator and approximate boundary operator for the adjoint operators K^* and M^*, respectively:

$$K_h^* \mathbf{v} = -\frac{(S^x \alpha^1) S^{2x} \mathbf{v} - (S^{-x} \alpha^1) S^{-2x} \mathbf{v}}{2h} - \frac{(S^y \alpha^2) S^{2y} \mathbf{v} - (S^{-y} \alpha^2) S^{-2y} \mathbf{v}}{2h} + G^* \mathbf{v},$$

$$M_h^* \mathbf{v} = (B^1 \mu)^* B^0 \mathbf{v} + (B^1 \beta) B^2 \mathbf{v}.$$

Let H_I be the set of all even interior grid points in Ω, and H_B the set of all even boundary grid points in $\partial\Omega$. Let $H = H_I \cup H_B$. With each grid point $x_j \in H$ we identify a $2h \times 2h$ mesh region P_j. If P_j is adjacent to P_k we say that x_j is connected to x_k. Now define A_j to be the area of P_j, which is $4h^2$, and $L_{j,k}$ to be the length of the line segment between P_j and P_k. We denote $\Gamma_{j,k} = \overline{P}_j \cap \overline{P}_k$. We use the notation $\sum_k$ to indicate a sum over points, x_k, which are connected to some point, x_j.

We now show the discrete version of (2.4).

Lemma 3.1 (First Discrete Identity). *If* $\mathbf{v}$ *and* $\mathbf{u}$ *are functions defined at even grid points, then*

$$(K_h \mathbf{u}, \mathbf{v})_h + \langle M_h \mathbf{u}, \mathbf{v} \rangle_h = (\mathbf{u}, K_h^* \mathbf{v})_h + \langle \mathbf{u}, M_h^* \mathbf{v} \rangle_h. \tag{3.5}$$

Proof. At each grid point $x_j \in H$, we write

$$\frac{1}{A_j} \sum_k L_{j,k} \gamma_{j,k} \mathbf{u}_k$$

$$:= \frac{1}{2h} ((S^x \alpha^1) S^{2x} \mathbf{u} - (S^{-x} \alpha^1) S^{-2x} \mathbf{u} + (S^y \alpha^2) S^{2y} \mathbf{u} - (S^{-y} \alpha^2) S^{-2y} \mathbf{u}),$$

where $\mathbf{u}_k = \mathbf{u}(x_k)$, and $\gamma_{j,k}$ denote, in order, the values of α^1, $-\alpha^1$, α^2, and $-\alpha^2$ at the centers of $\Gamma_{j,k}$ which correspond to the east, west, north, and south sides of x_j, respectively. Note that if $\Gamma_{j,k}$ is beyond the boundary, we have $\gamma_{j,k} = \beta_{j,k}$. Hence, at $x_j \in H$,

$$A_j (K_h \mathbf{u})_j = \sum_k L_{j,k} \gamma_{j,k} \mathbf{u}_k + A_j G \mathbf{u}_j.$$

Similarly,

$$A_j (K_h^* \mathbf{v})_j = \sum_k L_{j,k} (-\gamma_{j,k}) \mathbf{v}_k + A_j G^* \mathbf{v}_j.$$

We then have

$$(K_h\mathbf{u}, \mathbf{v})_h - (\mathbf{u}, K_h^*\mathbf{v})_h$$
$$= \sum_j \left[\sum_k L_{j,k}(\gamma_{j,k}\mathbf{u}_k \cdot \mathbf{v}_j + \gamma_{j,k}\mathbf{v}_k \cdot \mathbf{u}_j) + A_j(G\mathbf{u}_j \cdot \mathbf{v}_j - G^*\mathbf{v}_j \cdot \mathbf{u}_j) \right].$$

Since $\gamma_{J,k} = -\gamma_{k,j}$, and since α^1 and α^2 are symmetric, we have, by rearrangement,

$$\sum_j \sum_k L_{j,k}\gamma_{j,k}\mathbf{v}_k \cdot \mathbf{u}_j = -\sum_j \sum_k L_{j,k}\mathbf{v}_j \cdot \gamma_{j,k}\mathbf{u}_k,$$

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

$$(K_h\mathbf{u}, \mathbf{v})_h - (\mathbf{u}, K_h^*\mathbf{v})_h = \sum_{J\in B} \left[\sum_{k\notin I} L_{j,k}(\beta_{j,k}\mathbf{u}_k \cdot \mathbf{v}_j + \beta_{j,k}\mathbf{v}_k \cdot \mathbf{u}_j) \right]. \tag{3.6}$$

On the other hand, at each $x_j \in H_B$, we have

$$(M_h\mathbf{u})_j = (B^1\beta + \mu - \beta)_j(B^0\mathbf{u})_j - (B^1\beta)_j(B^2\mathbf{u})_j$$
$$= (\beta_{j,k} + \mu_j - \beta_j)\mathbf{u}_j - \beta_{j,k}\mathbf{u}_k,$$

and

$$(M_h^*\mathbf{v})_j = (B^1\beta + \mu - \beta)_j^*(B^0\mathbf{v})_j + (B^1\beta)_j(B^2\mathbf{v})_j$$
$$= (\beta_{j,k}^* + \mu_j^* - \beta_j^*)\mathbf{v}_j + \beta_{j,k}\mathbf{v}_k,$$

where $k \notin I$. Hence,

$$\langle \mathbf{u}, M_h^*\mathbf{v}\rangle_h - \langle M_h\mathbf{u}, \mathbf{v}\rangle_h = \sum_{j\in B} \left[\sum_{k\notin I} L_{j,k}(\beta_{j,k}\mathbf{u}_k \cdot \mathbf{v}_j + \beta_{j,k}\mathbf{v}_k \cdot \mathbf{v}_j) \right],$$

which is the same as the right side of (3.6). Hence (3.5) is proved. □

Observe that $K_h + K_h^* = G + G^*$ and $M_h + M_h^* = (B^1\mu) + (B^1\mu)^*$. By setting $\mathbf{v} = \mathbf{u}$ in (3.5), we get immediately the discrete version of (2.5).

Lemma 3.2 (Second Discrete Identity). *For functions* $\mathbf{u}$ *defined at even grid points, there holds*

$$(K_h\mathbf{u}, \mathbf{u})_h + \langle M_h\mathbf{u}, \mathbf{u}\rangle_h = (G\mathbf{u}, \mathbf{u})_h + \langle (B^1\mu)\mathbf{u}, \mathbf{u}\rangle_h. \tag{3.7}$$

Consequently, we have:

Lemma 3.3 (Basic Inequality). *For* $\mathbf{u}$ *satisfying the approximate boundary condition* $M_h\mathbf{u} = 0$, *there is a (generic) constant* $C > 0$ *such that*

$$\|\mathbf{u}\|_h \le C\|K_h\mathbf{u}\|_h.$$

The existence and uniqueness of the solution of (3.1), (3.2) can be proved as follows. We see that the unknowns in (3.1) include those at the boundary and in the extension. For unknowns in the extension, we substitute the boundary condition (3.2) into (3.1). Note that

$$B^1\beta = S^x\alpha^1, \quad B^2\mathbf{u} = S^{2x}\mathbf{u} \quad \text{on } x_+,$$
$$B^1\beta = -S^{-x}\alpha^1, \quad B^2\mathbf{u} = S^{-2x}\mathbf{u} \quad \text{on } x_-,$$

etc. Hence, by substituting

$$(B^1\beta)B^2\mathbf{u} = (B^i\mu)B^0\mathbf{u}$$

into (3.1), we then have a square finite system of linear equations with unknowns at all grid points $x_j \in H$ (not including the points in the extension). The Basic Inequality insures uniqueness of the solution of the system of linear equations. But since these equations are a square finite system, uniqueness of the solution also insures its existence. We summarize:

Theorem 3.1. *The system of finite difference equations* (3.1) *and boundary condition* (3.2) *possesses a unique solution.*

We now study the error between the approximate solution $\mathbf{u}_h$ of (3.1), (3.2) and the solution $\mathbf{u}$ of (2.2), (2.3).

We may express K in a form slightly different from (2.2), by the use of (2.1). That is,

$$K\mathbf{u} = 2\nabla\cdot(\alpha\mathbf{u}) - (\nabla\cdot\alpha)\mathbf{u} + G\mathbf{u}. \tag{3.8}$$

In order to relate $L_h^2(\Omega)$ to the usual $L^2(\Omega)$ space, we introduce a projection $r_h\colon C^0(\overline{\Omega}) \to L_h^2(\Omega)$ defined by

$$(r_h\mathbf{u})_j = \mathbf{u}(x_j), \quad \forall x_j \in H,$$

and an injection $p_h\colon L_h^2(\Omega) \to L^2(\Omega)$ defined by

$$p_h\mathbf{u}_h(x) = (\mathbf{u}_h)_j, \quad \forall x \in P_j \cap \Omega.$$

We immediately have $\|p_h\mathbf{u}_h\|_0 \le \|\mathbf{u}_h\|_h$, $\forall \mathbf{u}_h \in L_h^2(\Omega)$.

We shall need the following lemma which is given in [7].

Lemma 3.4. *Let g be a function defined on a finite region $P \subset \mathbf{R}^2$, and suppose that g satisfies a Lipschitz condition, i.e., there is a constant $C > 0$ such that $|g(x) - g(y)| \le C|x - y|$ for all $x, y \in P$. Then, if A is the area of P and $|x - x_0| \le h$ in P, we have*

$$\left| g(x_0) - \frac{1}{A}\int_P g(x) \right| \le Ch.$$

We now state the convergence properties of the method.

Theorem 3.2. *Suppose that $\mathbf{u} \in C^2(\overline{\Omega})$ is the solution of* (2.2), (2.3). *Let $\mathbf{u}_h \in L_h^2(\Omega)$ be the unique solution of* (3.1), (3.2). *Then*

$$\|r_hK\mathbf{u} - K_hr_h\mathbf{u}\|_h = O(h), \tag{3.9}$$

$$|r_hM\mathbf{u} - M_hr_h\mathbf{u}|_h = O(h). \tag{3.10}$$

Moreover, the discrete error converges at the rate

$$\|\mathbf{u}_h - r_h\mathbf{u}\|_h = O(h^{1/2}), \tag{3.11}$$

$$\|p_h\mathbf{u}_h - \mathbf{u}\|_0 = O(h^{1/2}). \tag{3.12}$$

Proof. Using the Second Discrete Identity (3.7), the positive definiteness of G and positive semidefiniteness of μ, we have for some constant $C > 0$

$$\|\mathbf{u}_h - r_h\mathbf{u}\|_h^2 \le C[(\mathbf{u}_h - r_h\mathbf{u},\, K_h(\mathbf{u}_h - r_h\mathbf{u}))_h + \langle \mathbf{u}_h - r_h\mathbf{u},\, M_h(\mathbf{u}_h - r_h\mathbf{u})\rangle_h].$$

Writing $K_h(\mathbf{u}_h - r_h\mathbf{u}) = r_h\mathbf{f} - r_h\mathbf{f} + r_hK\mathbf{u} - K_hr_h\mathbf{u}$, using the boundary conditions $M_h\mathbf{u}_h = r_hM\mathbf{u} = 0$ on H_B and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|\mathbf{u}_h - r_h\mathbf{u}\|_h^2 &\le C[\|\mathbf{u}_h - r_h\mathbf{u}\|_h\|K_h(\mathbf{u}_h - r_h\mathbf{u})\|_h + |\mathbf{u}_h - r_h\mathbf{u}|_h|M_h(\mathbf{u}_h - r_h\mathbf{u})|_h] \\ &= C[\|\mathbf{u}_h - r_h\mathbf{u}\|_h\|r_hK\mathbf{u} - K_hr_h\mathbf{u}\|_h + |\mathbf{u}_h - r_h\mathbf{u}|_h|M_hr_h\mathbf{u}|_h].\end{aligned}$$

We shall show that $\|r_hK\mathbf{u} - K_hr_h\mathbf{u}\|_h = O(h^1)$ and $|M_hr_h\mathbf{u}|_h = O(h^1)$. Then, (3.4), (3.9), and (3.10) imply (3.11).

From the definition of $\|\cdot\|_h$, we have

$$\|r_hK\mathbf{u} - K_hr_h\mathbf{u}\|_h^2 = \sum_e (r_hK\mathbf{u} - K_hr_h\mathbf{u})^2 4h^2.$$

We now obtain a suitable bound for $|K\mathbf{u}(x_j) - (K_hr_h\mathbf{u})_j|$. By (3.8),

$$\begin{aligned}&|K\mathbf{u}(x_j) - (K_hr_h\mathbf{u})_j| \\ &\quad = \left|2\nabla\cdot(\alpha\mathbf{u})(x_j) - (\nabla\cdot\alpha)\mathbf{u}(x_j) - \frac{1}{A_j}\sum_k L_{j,k}\gamma_{j,k}\mathbf{u}_k\right| \\ &\quad \le \left|2\nabla\cdot(\alpha\mathbf{u})(x_j) - \frac{1}{A_j}\sum_k L_{j,k}\gamma_{j,k}(\mathbf{u}_k + \mathbf{u}_j)\right| \\ &\qquad + \left|(\nabla\cdot\alpha)\mathbf{u}(x_j) - \frac{1}{A_j}\sum_k L_{j,k}\gamma_{j,k}\mathbf{u}_j\right|.\end{aligned} \tag{3.13}$$

Consider the first term in the last expression above:

$$\begin{aligned}&\left|2\nabla\cdot(\alpha\mathbf{u})(x_j) - \frac{1}{A_j}\sum_k L_{j,k}\gamma_{j,k}(\mathbf{u}_k + \mathbf{u}_j)\right| \\ &\quad \le \left|2\nabla\cdot(\alpha\mathbf{u})(x_j) - \frac{2}{A_j}\int_{P_j}\nabla\cdot(\alpha\mathbf{u})\right| \\ &\qquad + \frac{1}{A_j}\left|\sum_k\int_{\Gamma_{j,k}} 2(\gamma\mathbf{u} - (\gamma\mathbf{u})_{j,k})\right| \\ &\qquad + \frac{1}{A_j}\left|\sum_k\int_{\Gamma_{j,k}} (2(\gamma\mathbf{u})_{j,k} - \gamma_{j,k}(\mathbf{u}_k + \mathbf{u}_j))\right|.\end{aligned} \tag{3.14}$$

Using Lemma 3.4, we have

$$\left|2\nabla\cdot(\alpha\mathbf{u})(x_j) - \frac{2}{A_j}\int_{P_j}\nabla\cdot(\alpha\mathbf{u})\right| = O(h). \tag{3.15}$$

We now examine a Taylor series expansion for $\gamma\mathbf{u}$ about the point $x_{j,k} = (x_j + x_k)/2$:

$$\gamma(x_{j,k} + tz)\mathbf{u}(x_{j,k} + tz) = (\gamma\mathbf{u})_{j,k} + t\left(\frac{d}{dt}(\gamma\mathbf{u})\right)_{j,k} + \frac{t^2}{2}g(\xi^1), \tag{3.16}$$

$$\gamma(x_{j,k} - tz)\mathbf{u}(x_{j,k} - tz) = (\gamma\mathbf{u})_{j,k} - t\left(\frac{d}{dt}(\gamma\mathbf{u})\right)_{j,k} + \frac{t^2}{2}g(\xi^2), \tag{3.17}$$

where z is a unit vector orthogonal to $x_j - x_k$, t is a scalar parameter, $g(\xi) = (g_1(\xi_1), g_2(\xi_2))$, g_i is the ith component of the vector $(d^2/dt^2)(\gamma \mathbf{u})$, 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 (3.16) and (3.17), we obtain the following bound:

$$\left| \int_{\Gamma_{j,k}} \gamma \mathbf{u} - (\gamma \mathbf{u})_{j,k} \right| = O(h^3). \tag{3.18}$$

Since $\mathbf{u} \in C^2$, we have

$$\mathbf{u}_j = \mathbf{u}_{j,k} - h\mathbf{u}'_{j,k} + \frac{h^2}{2}\mathbf{u}''(\xi_1), \qquad \mathbf{u}_k = \mathbf{u}_{j,k} + h\mathbf{u}'_{j,k} + \frac{h^2}{2}\mathbf{u}''(\xi_2),$$

where the derivatives are directional derivatives in the direction $x_k - x_j$. Hence, we have

$$|2\mathbf{u}_{j,k} - (\mathbf{u}_j + \mathbf{u}_k)| < Ch^2.$$

This means that

$$\begin{aligned} &\left| \int_{\Gamma_{j,k}} \gamma_{j,k}(2\mathbf{u}_{j,k} - (\mathbf{u}_j + \mathbf{u}_k)) \right| \\ &\qquad \le L_{j,k} \|\gamma_{j,k}\| \, |2\mathbf{u}_{j,k} - (\mathbf{u}_j + \mathbf{u}_k)| = O(h^3). \end{aligned} \tag{3.19}$$

Using (3.15), (3.18), and (3.19) in (3.14), we obtain

$$\left| 2\nabla \cdot (\alpha \mathbf{u})(x_j) - \frac{1}{A_j} \sum_k L_{j,k} \gamma_{j,k} (\mathbf{u}_k + \mathbf{u}_j) \right| = O(h). \tag{3.20}$$

We now consider the second term on the right of (3.13):

$$\begin{aligned} &\left| (\nabla \cdot \alpha)\mathbf{u}(x_j) - \frac{1}{A_j} \sum_k L_{j,k} \gamma_{j,k} \mathbf{u}_j \right| \\ &\qquad \le \left| (\nabla \cdot \alpha)\mathbf{u}(x_j) - \frac{1}{A_j} \int_{P_j} (\nabla \cdot \alpha)\mathbf{u} \right| \\ &\qquad\quad + \left| \frac{1}{A_j} \int_{P_j} (\nabla \cdot \alpha)(\mathbf{u} - \mathbf{u}_j) \right| + \left| \frac{1}{A_j} \sum_k \int_{\Gamma_{j,k}} (\gamma - \gamma_{j,k}) \mathbf{u}_j \right|. \end{aligned} \tag{3.21}$$

Again, by Lemma 3.4, we get

$$\left| (\nabla \cdot \alpha)\mathbf{u}(x_j) - \frac{1}{A_j} \int_{P_j} (\nabla \cdot \alpha)\mathbf{u} \right| = O(h). \tag{3.22}$$

Since $\mathbf{u}$ satisfies a Lipschitz condition, $|x - x_j| < h$ for all $x \in P_j$, and since $\|\nabla \cdot \alpha\|_0$ is uniformly bounded in Ω, we have

$$\left| \frac{1}{A_j} \int_{P_j} (\nabla \cdot \alpha)(\mathbf{u} - \mathbf{u}_j) \right| \le \|\nabla \cdot \alpha\|_0 |\mathbf{u} - \mathbf{u}_j| = O(h). \tag{3.23}$$

Since $\gamma_{j,k}$ is evaluated at the midpoint of $\Gamma_{j,k}$, we can use a Taylor series analysis, as in deriving equation (3.18), to get

$$\left| \frac{1}{A_j} \sum_k \int_{\Gamma_{j,k}} (\gamma - \gamma_{j,k}) \mathbf{u}_j \right| = O(h). \tag{3.24}$$

Combining (3.22), (3.23), and (3.24) in (3.21), we obtain

$$\left|(\nabla\cdot\alpha)\mathbf{u}(x_j)-\frac{1}{A_j}\sum_k L_{j,k}\gamma_{j,k}\mathbf{u}_j\right|=O(h). \tag{3.25}$$

Substituting (3.20) and (3.25) in (3.13), we obtain

$$|K\mathbf{u}(x_j)-(K_h r_h\mathbf{u})_j|=O(h),$$

from which (3.9) follows. Next, we prove (3.10). By (3.3), we have

$$\begin{aligned}M_h r_h\mathbf{u}&=B^1\mu B^0 r_h\mathbf{u}-B^1\beta B^2 r_h\mathbf{u}\\&=B^1\mu B^0 r_h\mathbf{u}-MB^0 r_h\mathbf{u}-B^1\beta B^2 r_h\mathbf{u}\\&=B^1\beta B^0 r_h\mathbf{u}-B^1\beta B^2 r_h\mathbf{u}.\end{aligned}$$

Hence,

$$\begin{aligned}|M_h r_h\mathbf{u}|_h&\le|B^1\beta|_h|B^0 r_h\mathbf{u}-B^2 r_h\mathbf{u}|_h\\&\le C\sum_j\sum_k\left|\int_{\Gamma_{j,k}}\mathbf{u}_j-\mathbf{u}_k\right|,\end{aligned}$$

for all $x_j\in H_B$, which implies that $|M_h r_h\mathbf{u}|_h=O(h)$, i.e., that (3.10) holds. Finally, we prove (3.12). We have

$$\begin{aligned}\|p_h\mathbf{u}_h-\mathbf{u}\|_0&=\|p_h r_h\mathbf{u}-p_h(r_h\mathbf{u}-\mathbf{u}_h)-\mathbf{u}\|_0\\&\le\|p_h r_h\mathbf{u}-\mathbf{u}\|_0+\|p_h(\mathbf{u}_h-r_h\mathbf{u})\|_0.\end{aligned}$$

Since

$$\|p_h(\mathbf{u}_h-r_h\mathbf{u})\|_0\le\|\mathbf{u}_h-r_h\mathbf{u}\|_h,$$

and

$$\|p_h r_h\mathbf{u}-\mathbf{u}\|_0^2=\sum_j\int_{Q_j}(\mathbf{u}_j-\mathbf{u})^2=O(h^2),$$

where $Q_j=P_j\cap\Omega$, we get (3.12). This concludes the proof. □

Remark. In the proof of this theorem, we did not require $\mu+\mu^*$ to be strictly positive definite; Katsanis [7] proved the rate of convergence using this condition. The domain considered in this paper is a rectangle, and the rate of convergence we have achieved is $O(h^{1/2})$. If a more complicated domain is encountered with piecewise smooth boundary, then Katsanis's finite difference scheme can be used. In this case the rate of convergence would still be $O(h^{1/2})$; however $\|r_h K\mathbf{u}-K_h r_h\mathbf{u}\|_h$ would be $O(h^{1/2})$ instead of $O(h)$. Note that (3.9) and (3.10) can be interpreted as the consistency of the operators K and K_h and the operators M and M_h, respectively.

4. A MODEL PROBLEM

Katsanis's method was motivated primarily by the numerical treatment of the Tricomi problem, which reduces to a first-order system with positive definite boundary matrix μ in (2.3). The following model problem shows that the above restriction needs to be removed and hence the method given in the previous section can be used. We consider the boundary value problem

$$\sigma(x,y)\phi_y(x,y)-\phi_{xx}(x,y)=f(x,y),\quad\forall(x,y)\in\Omega, \tag{4.1}$$

$$\begin{cases} \phi(\pm 1, y) = 0, & \forall y \in [0, 1], \\ \phi(x, 0) = 0, & \forall x \in [0, 1], \\ \phi(x, 1) = 0, & \forall x \in [-1, 0], \end{cases} \tag{4.2}$$

where $\Omega = (-1, 1) \times (0, 1)$ and the coefficient $\sigma(x, y)$ changes sign in Ω. There have been a number of papers addressing this kind of mixed-type heat equations (see e.g. [2, 3, 8, 11]). For a discussion of possible applications of such equations, we refer to [12].

We briefly show how the boundary value problem (4.1), (4.2) can be formulated as a Friedrichs's system. By a change of dependent variables,

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad u_1 = e^{-\lambda y}\phi, \quad u_2 = e^{-\lambda y}\phi_x,$$

equation (4.1) may be written as the symmetric first-order system

$$A_1\mathbf{u}_x + A_2\mathbf{u}_y + A_0\mathbf{u} = \mathbf{f}, \tag{4.3}$$

where

$$A_1 = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}, \quad A_0 = \begin{pmatrix} \lambda\sigma & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} e^{-\lambda y} f \\ 0 \end{pmatrix}.$$

Note that (4.3) in general is not positive. To make (4.3) symmetric positive, we multiply both sides by a 2×2 matrix

$$T = \begin{pmatrix} a & b\sigma \\ 0 & a \end{pmatrix},$$

where a and b are functions of x and y to be specified later. Then the forward-backward heat equation (4.1), (4.2) can be expressed in symmetric positive form, with properly chosen a, b, and λ, by

$$K\mathbf{u} = \mathbf{f} \quad \text{in } \Omega, \tag{4.4}$$

$$M\mathbf{u} = (\mu - \beta)\mathbf{u} = 0, \quad \forall (x, y) \in \partial\Omega, \tag{4.5}$$

where

$$K\mathbf{u} = \alpha^1\mathbf{u}_x + (\alpha^1\mathbf{u})_x + \alpha^2\mathbf{u}_y + (\alpha^2\mathbf{u})_y + G\mathbf{u},$$

$$\alpha^1 = \frac{1}{2}\begin{pmatrix} -b\sigma & -a \\ -a & 0 \end{pmatrix}, \qquad \alpha^2 = \frac{1}{2}\begin{pmatrix} a\sigma & 0 \\ 0 & 0 \end{pmatrix},$$

$$\alpha^0 = \begin{pmatrix} \lambda a\sigma & b\sigma \\ 0 & a \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} ae^{-\lambda y} f \\ 0 \end{pmatrix},$$

$$G = \alpha^0 - (\alpha^1)_x - (\alpha^2)_y, \qquad \beta = \frac{1}{2}\begin{pmatrix} (an_y - bn_x)\sigma & -an_x \\ -an_x & 0 \end{pmatrix},$$

$$\mu = \frac{1}{2}\begin{pmatrix} q^+ + q^- & -an_x \\ an_x & 0 \end{pmatrix}.$$

Here, $q^{\pm}$ are nonnegative functions defined by $(an_y - bn_x)\sigma = q^+ - q^-$. Then

$$\frac{G + G^*}{2} = \frac{1}{2}\begin{pmatrix} 2\lambda a\sigma - (a\sigma)_y + (b\sigma)_x & b\sigma + a_x \\ b\sigma + a_x & 2a \end{pmatrix}$$

will be positive definite and μ will be positive semidefinite, and $\ker(\mu - \beta) \oplus \ker(\mu + \beta) = \mathbf{R}^2$.

We summarize the conditions on the choice of a, b, and λ in order to have a positive system with admissible boundary condition:

(I) T is piecewise differentiable,
(II) $q^{\pm}$ are nonnegative functions on $\partial\Omega$,
(III) $G + G^*$ is positive definite.

The existence and uniqueness of the solution of (4.1), (4.2) follows immediately by using Friedrichs's results [5]. The reduction of the second-order problem to a first-order system not only simplifies the proof of existence and uniqueness but also is more convenient for the numerical treatment of the problem. Note also that, by Theorem 3.2, the solution of (4.1), (4.2) is required to have continuous derivatives of order 3 in x, compared with order 4 for the method in [12].

We give some examples on the choice of a, b, and λ.

Example 1. The case of $\sigma(x, y) = x^m$ with m an odd positive integer has been considered in [6]. For this we choose $\lambda = 0$, $a = 1$, and b such that $bx^m = x$. After a simple calculation, we have

$$G + G^* = \begin{pmatrix} 1 & x \\ x & 2 \end{pmatrix},$$

which is positive definite for all $x \in [-1, 1]$.

Example 2. We now show an example for which $\sigma(x, y) = x + \frac{1}{8}y$. Let $\lambda = 0.1$, $a = 2$, and $b = 1$. After a simple calculation, we have

$$G + G^* = \begin{pmatrix} \frac{3}{4} + 0.2(x + \frac{1}{8}y) & x + \frac{1}{8}y \\ x + \frac{1}{8}y & 4 \end{pmatrix},$$

which is again positive definite for all $x \in [-1, 1]$, $y \in [0, 1]$. As mentioned in the introduction, the iterative method proposed in [12] requires a more restrictive condition on the coefficient function σ, namely $\sigma_y \leq 0$ in Ω.

Example 3. Our numerical experiment was based on the example for which $\sigma(x, y) = x$. This kind of coefficient function appearing in the equation has been considered by many authors (see for example [11, 3]). In our computations f is taken to be

$$\begin{aligned} f(x, y) &= 2x(x^2 - 1)y[(y - 1)^2 - 4x^2 + y(y - 1)] \\ &\quad - 2y^2[(y - 1)^2 - 24x^2 + 4] \quad \forall x \geq 0,\ y \in [0, 1], \\ f(x, y) &= 2x(x^2 - 1)(y - 1)(2y^2 - y - 4x^2) \\ &\quad - 2(y - 1)^2(y^2 - 24x^2 + 4) \quad \forall x \leq 0,\ y \in [0, 1]. \end{aligned}$$

Denote the boundary $\partial\Omega$ by $\Gamma_1 \cup \cdots \cup \Gamma_6$,

$$\begin{aligned} \Gamma_1 &= \{(x, y): x \in [-1, 0],\ y = 0\}, \\ \Gamma_2 &= \{(x, y): x = -1,\ y \in [0, 1]\}, \\ \Gamma_3 &= \{(x, y): x \in [-1, 0],\ y = 1\}, \\ \Gamma_4 &= \{(x, y): x \in [0, 1],\ y = 1\}, \\ \Gamma_5 &= \{(x, y): x = 1,\ y \in [0, 1]\}, \\ \Gamma_6 &= \{(x, y): x \in [0, 1],\ y = 0\}. \end{aligned}$$

To formulate (4.1), (4.2) as a symmetric positive system, we choose $\lambda = 0.1$, $a = 1$, and $b = 1$. Then

$$G = \begin{pmatrix} 0.1x + \frac{1}{2} & x \\ 0 & 1 \end{pmatrix}.$$

We see that $(G+G^*)$ is positive definite in Ω. We need to evaluate the matrices μ and β along all boundaries. A straightforward calculation gives the values for μ, β, and M shown in Table 4.1.

TABLE 4.1

	2μ	2β	M
Γ_1	$\begin{pmatrix} -x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
Γ_2	$\begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}$	$\begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}$
Γ_3	$\begin{pmatrix} -x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -x & 0 \\ 0 & 0 \end{pmatrix}$
Γ_4	$\begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
Γ_5	$\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} -1 & -1 \\ -1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$
Γ_6	$\begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -x & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$

Of course, μ is positive semidefinite. Also, $\ker(\mu - \beta) \oplus \ker(\mu + \beta) = \mathbf{R}^2$, so that the boundary condition (4.2) is admissible. Theorem 3.2 assures us of essentially $O(h^{1/2})$ convergence in the L^2 norm. However, the results shown in Table 4.2 are better, i.e., $O(h)$.

TABLE 4.2

h	L^∞ error	L^2 error	L^2 rate
1/4	14.120	4.897	
1/8	6.077	1.706	1.52
1/12	4.033	1.048	1.20
1/16	3.464	0.764	1.10
1/20	3.145	0.602	1.06
1/24	2.956	0.496	1.05

ACKNOWLEDGMENT

The author would like to thank Professor A. K. Aziz for his guidance in this work. The author would also like to express his gratitude to the referee for valuable comments on the paper.

BIBLIOGRAPHY

1. A. K. Aziz and J.-L. Liu, *A weighted least squares method for the backward-forward heat equation*, SIAM J. Numer. Anal. **28** (1991), 156–167.
2. M. S Baouendi and P. Grisvard, *Sur une équation d'évolution changeant de type*, J. Funct. Anal. **2** (1968), 352–367.
3. R. Beals, *On an equation of mixed type from electron scattering theory*, J. Math. Anal. Appl. **58** (1977), 32–45.
4. C. K. Chu, *Type-insensitive finite difference schemes*, Ph.D. Thesis, New York University, 1958.
5. K. O. Friedrichs, *Symmetric positive differential equations*, Comm. Pure Appl. Math. **11** (1958), 333–418.
6. J. A. Goldstein and T. Mazumdar, *A heat equation in which the diffusion coefficient changes sign*, J. Math. Anal. Appl. **103** (1984), 533-564.
7. T. Katsanis, *Numerical solution of symmetric positive differential equations*, Math. Comp. **22** (1968), 763–783.
8. T. LaRosa, *The propagation of an electron beam through the solar corona*, Ph.D. Thesis, Dept. of Physics and Astronomy, University of Maryland, 1986.
9. P. Lesaint, *Finite element methods for symmetric hyperbolic equations*, Numer. Math. **21** (1973), 244–255.
10. P. Lesaint and P. A. Raviart, *Finite element collocation methods for first order systems*, Math. Comp. **33** (1979), 891–918.
11. J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limits non linéaires*, Gauthier-Villars, Paris, 1969, pp. 337–343.
12. V. Vanaja and R. B. Kellogg, *Iterative methods for a forward-backward heat equation*, SIAM J. Numer. Anal. **27** (1990), 622–635.

DEPARTMENT OF APPLIED MATHEMATICS, NATIONAL CHIAO TUNG UNIVERSITY, HSINCHU, TAIWAN

E-mail address: jinnliu@cc.nctu.edu.tw