MEMOIRS OF NUMERICAL MATHEMATICS

NUMBER 6

1979

The MEMOIRS OF NUMERICAL MATHEMATICS will publish original papers concerned with numerical mathematics, and henceforth will be published once a year.

All correspondence should be addressed to

Sin Hitotumatu
Research Institute for Mathematical Sciences
Kyoto University, Sakyo-ku, Kyoto
606, Japan

CONTENTS

On a Mixed Finite Element Scheme for Buckling Analysis of Plates

Fumio KIKUCHI*

1. Introduction

Our problem is expressed by the buckling equation of clamped plates

$$(1)\quad \left\| \begin{aligned} &\lambda\,\Delta^2 u + \sum_{i,j=1}^{2} \tau_{ij}\,\partial^2 u/\partial x_i \partial x_j = 0 \qquad \text{in } \Omega, \\ &\qquad\qquad\qquad u = \partial u/\partial n = 0 \qquad \text{on } \partial\Omega, \end{aligned} \right.$$

where $\Omega \subset R^2$ is a non-obtuse polygonal domain with boundary $\partial\Omega$, Δ the Laplace operator, $\partial/\partial n$ differentiation in the outward normal direction of $\partial\Omega$, $x = (x_1, x_2)$ the independent variable of R^2, and τ_{ij} $(1 \leqq i,j \leqq 2)$ are sufficiently smooth given functions of x such that

$$(2)\qquad \tau_{12} = \tau_{21}\ , \qquad \sum_{j=1}^{2} \partial\tau_{ij}/\partial x_j = 0 \qquad (i = 1,2)\ .$$

We are interested in a non-trivial function $u = u(x)$ and a (real) number λ that satisfy (1).

In our mixed method, we first decompose (1) into a system

$$(3)\quad \left\| \begin{aligned} &-\Delta u = v\ , \qquad -\lambda\,\Delta v + \sum_{i,j=1}^{2} \tau_{ij}\,\partial^2 u/\partial x_i \partial x_j = 0 \qquad \text{in } \Omega, \\ &\qquad\qquad\qquad u = \partial u/\partial n = 0 \qquad\qquad \text{on } \partial\Omega, \end{aligned} \right.$$

* Institute of Space and Aeronautical Science, University of Tokyo, Tokyo, 153, Japan.

and then use appropriate finite element spaces and weak forms for discretization. This type of mixed method has been proposed by many investigators, see for example Kikuchi [1], and Ciarlet and Glowinski [2].

As for the error analysis of the mixed method for the above eigenvalue problem, we can refer to Ishihara [3], where error estimates of approximate eigenpairs are derived for the piecewise linear finite element model considered over nearly uniform triangulations. The analysis is based on Miyoshi's results for the mixed approximation of the corresponding boundary value problems [4]. It also uses the Rayleigh and the min-max principles, and hence its validity is restricted to equations such as $\lambda\Delta^2 u + \Delta u = 0$.

The main objective of this note is to present error analysis of the above mixed method applied to more general cases, where the Rayleigh and the min-max principles are not available, or the employed meshes are not necessarily uniform. We will use the so-called spectral projection to our end, see for example Grigorieff [5]. We will also utilize the results obtained by Scholz [6] to obtain order estimates of errors of approximate eigenpairs. In our analysis, it is essential to prove a certain compactness property of a family of approximate operators. The techniques developed in this note will be available for higher order finite elements, and also for the different type of mixed models where u itself as well as its all second order derivatives are considered as independent functions to be approximated (se for example Miyoshi [4]).

In this note, C will be used as a generic positive constant, which may take different values when it appears in different places.

2. Preliminaries

We will use the Sobolev spaces $H^m(\Omega)$ and $H_0^m(\Omega)$ equipped with the same norm $\| \ \|_m$ $(m = 1,2,\ldots)$. The inner product and the norm of $L_2(\Omega)$ will be denoted by $(\ , \ )$ and $\| \ \|$, respectively.

The problem (1) expressed in a weak form is to find a pair $\{\lambda,u\} \in R^1 \times H_0^2(\Omega)$ such that $u \neq 0$ and

$$(4) \qquad \lambda(\Delta u,\Delta\bar{u}) = b(u,\bar{u}) \quad \text{for all} \quad \bar{u} \in H_0^2(\Omega) \ ,$$

where

$$(5) \qquad b(u,\bar{u}) = \sum_{i,j=1}^{2} (\tau_{ij}\partial u/\partial x_i,\partial\bar{u}/\partial x_j) \quad .$$

Here we have used the relation (2).

Another expression may be given to the same problem, if a weak form of (3) is employed: find a triplet $\{\lambda,u,v\} \in R^1 \times H_0^1(\Omega) \times H^1(\Omega)$ such that $u \neq 0$ and

$$(6) \quad (\nabla u,\nabla\bar{v}) = (v,\bar{v}) \quad (^\forall \bar{v} \in H^1(\Omega)\,) \ , \quad \lambda(\nabla v,\nabla\bar{u}) = b(u,\bar{u}) \quad (^\forall \bar{u} \in H_0^1(\Omega)\,),$$

where $(\nabla u,\nabla\bar{v})$ for example implies

$$(7) \qquad (\nabla u,\nabla\bar{v}) = \sum_{i=1}^{2} (\partial u/\partial x_i,\partial\bar{v}/\partial x_i) \quad .$$

In order to check the equivalence between (4) and (6), notice that the solution u of (4) is necessarily belongs to $H^4(\Omega)$ since Ω is a non-obtuse polygonal domain, see Mizutani [7]. Then the condition $v = -\Delta u \in H^1(\Omega)$ in (6) can be easily justified.

3. Finite Element Scheme

We triangulate Ω in the usual way. We assume that the considered triangulation T^h for $h > 0$ is κ-regular in the following sense, see Scholz [6]:

(i) $T \in T^h$ is a (closed) triangle, each side of which is either a portion of $\partial\Omega$ or a side of an adjacent triangle in T;

(ii) there is a fixed constant $\kappa > 0$ such that two circles K_1 and K_2 exist with the properties

$$\text{radius of } K_1 = h/\kappa \quad , \quad \text{radius of } K_2 = \kappa h \quad ,$$

$$\text{and} \quad K_1 \subset T \subset K_2 \quad \text{for all} \quad T \in T^h.$$

In the sequel, we will only consider a κ-regular family of triangulations $\{T^h\}$ with a fixed $\kappa > 0$, and the case when $h \downarrow 0$.

Let $X^h = X^h(T^h)$ be the space of continuous functions which are linear polynomials in each $T \in T^h$. Then X^h is a finite dimensional subspace of $H^1(\Omega)$. We also use the space X_0^h of all functions in X^h that vanish on $\partial\Omega$, that is, $X_0^h = X^h \cap H_0^1(\Omega)$.

Using the above two finite element spaces X^h and X_0^h, we can introduce a mixed finite element approximation to (6): find a triplet $\{\lambda_h, u_h, v_h\} \in R^1 \times X_0^h \times X^h$ such that $u_h \neq 0$ and

$$(8) \quad \left\| \begin{array}{lll} (\nabla u_h, \nabla v_h) = (v_h, \bar{v}_h) & (\forall \bar{v}_h \in X^h) & , \\ \\ \lambda_h(\nabla v_h, \nabla \bar{u}_h) = b(u_h, \bar{u}_h) & (\forall \bar{u}_h \in X_0^h) & . \end{array} \right.$$

For each $u_h \in X_0^h$, we can define $\Delta_h u_h \in X^h$ by the relation

$$(9) \qquad (\Delta_h u_h, v_h) = -(\nabla u_h, \nabla v_h) \quad \text{for all} \quad v_h \in X^h .$$

Then we can rewrite the system of equations (8) into a single one for $\{\lambda_h, u_h\} \in R^1 \times X_0^h$:

$$(10) \qquad \lambda_h (\Delta_h u_h, \Delta_h \bar{u}_h) = b(u_h, \bar{u}_h) \quad \text{for all} \quad \bar{u}_h \in X_0^h ,$$

where v_h has been eliminated by the relation $v_h = -\Delta_h u_h$.

4. Properties of Δ_h

In the space $H_0^2(\Omega)$, $\| \Delta u \|$ can be regarded as a norm equivalent to $\| u \|_2$, and hence we have the following properties: Let $\{u_i\}_{i=1}^{\infty}$ be a sequence in $H_0^2(\Omega)$ such that $\| \Delta u_i \| \leqq C$ for a certain positive constant C. Then we can choose a sub-sequence, again denoted by $\{u_i\}_{i=1}^{\infty}$ for simplicity, such that for i → ∞,

$$(11) \qquad u_i \to u \quad \text{weakly in } H_0^2(\Omega), \text{ and strongly in } H_0^1(\Omega),$$

where u is a certain element in $H_0^2(\Omega)$. We will show that the operator Δ_h, when considered over an appropriate family of X_0^h, possesses properties in a sense corresponding to the above.

Lemma 1 There exists a positive constant $C = C(\Omega)$ such that

$$(12) \qquad \| u_h \| \leqq C \| \nabla u_h \| \leqq C^2 \| \Delta_h u_h \| \quad \text{for any} \quad u_h \in X_0^h,$$

where $\| \nabla u_h \| = (\nabla u_h, \nabla u_h)^{1/2}$.

Proof The former part of (12) is nothing but the Poincare inequality, while the latter is obtained by substituting $v_h = u_h \in X_0^h \subset X^h$ and using the Schwarz inequality in (9). ////

Lemma 2 Let $\{T^{h(i)}\}_{i=1}^{\infty}$ be a (κ-regular) sequence of triangulations such that $\lim_{i\to\infty} h(i) = 0$. Consider a sequence of functions $\{u_{h(i)}\}_{i=1}^{\infty}$ such that

$$(13) \qquad u_{h(i)} \in X_0^{h(i)} \ , \qquad \| \Delta_{h(i)} u_{h(i)} \| \leqq C \ ,$$

where C is a positive constant, and $X_0^{h(i)}$ and $\Delta_{h(i)}$ are the finite element space X_0^h and the operator Δ_h associated with $T^{h(i)}$, respectively. Then we can choose a subsequence, again denoted by $\{u_{h(i)}\}_{i=1}^{\infty}$ for simplicity, such that for i → ∞

$$(14) \qquad \left[\begin{array}{ll} \Delta_{h(i)} u_{h(i)} \to \Delta u & \text{weakly in } L_2(\Omega) \ , \\ \\ u_{h(i)} \to u & \text{strongly in } H_0^1(\Omega) \ , \end{array} \right.$$

where u is a certain element of $H_0^2(\Omega)$.

Proof In this proof, we omit the index i, and h → 0 implies i → ∞.

1° Since $\Delta_h u_h$ is uniformly bounded in $L_2(\Omega)$, we can show the existence of a subsequence, again denoted by $\{u_h\}$, such that for h → 0 ,

$$\begin{array}{ll} \Delta_h u_h \to w & \text{weakly in } L_2(\Omega) \ , \\ u_h \to u & \text{weakly in } H_0^1(\Omega) \text{ and strongly in } L_2(\Omega) \ , \end{array}$$

where w and u are certain elements of $L_2(\Omega)$ and $H_0^1(\Omega)$, respectively, and we have used the results of Lemma 1. On the other hand, we can find for each $v \in H^1(\Omega)$ a sequence $\{v_h\}$ such that

$$v_h \in X^h, \text{ and } v_h \to v \text{ strongly in } H^1(\Omega) \text{ as } h \to 0.$$

The existence of such an approximate sequence is assured by the usual approximation theory, see for example Ciarlet [8]. Taking the limit of

$$(\nabla u_h, \nabla v_h) = -(\Delta_h u_h, v_h) \quad ,$$

we find

$$(*) \qquad (\nabla u, \nabla v) = -(w,v) \quad \text{for all} \quad v \in H^1(\Omega) \ .$$

Then we have for $h \to 0$

$$\| \nabla u_h \|^2 = -(\Delta_h u_h, u_h) \to -(w,u) = \| \nabla u \|^2 \quad ,$$

from which follows the strong convergence of u_h to u in $H_0^1(\Omega)$.

2° Choosing v from $H_0^1(\Omega)$ in (*), we have

$$(\nabla u, \nabla v) = -(w,v) \quad \text{for all} \quad v \in H_0^1(\Omega) \ .$$

Since Ω is a non-obtuse polygonal domain, we can assure that $u \in H_0^1(\Omega) \cap H^2(\Omega)$ and $\Delta u = w$, see Kondrat'ev [9]. Applying the Green's formula to the left-hand side of (*) yields

$$-(\Delta u, v) + \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\gamma = -(w,v) \quad \text{for all} \quad v \in H^1(\Omega),$$

where $d\gamma$ is the line element on $\partial\Omega$. We can now conclude that $\partial u/\partial n = 0$ on $\partial\Omega$, and hence $u \in H_0^2(\Omega)$. ////

5. Fundamental Properties of the Approximate Eigenvalue Problems

As before, we assume that the considered family of triangulations is κ-regular. Define σ, σ_h, and $\sigma(\varepsilon)$ by

$$\sigma = \text{set of all eigenvalues for the problem (4)},$$

$$\sigma_h = \text{set of all eigenvalues for the problem (10)},$$

$$\sigma(\varepsilon) = \bigcup_{\lambda \in \sigma}]\lambda - \varepsilon, \lambda + \varepsilon[\quad ,$$

where ε is an arbitrary positive number. Clearly σ is a bounded countable subset of R^1 without any accumulation points except zero. In particular, zero is either an eigenvalue of (4) or an accumulation point of σ. These properties directly follow from the theory of compact operators. We can also show that σ_h is a finite bounded subset of R^1, the boundedness being uniform with respect to h, as follows from Lemma 1. Notice that $\sigma(\varepsilon)$ is an open set.

In the sequel, we will only consider a ***non-zero*** eigenvalue $\lambda_0 \in \sigma$, which can be fixed arbitrarily. The constants to appear in the forthcoming lemmas and theorems may differ with λ_0, but we will make no special distinction among them. Define

$$E = \text{eigenspace for } \lambda_0 \ ,$$

$$m = \text{dimension of } E;\ 1 \leq m < \infty \ ,$$

$$\{\phi_i\}_{i=1}^m = \text{a basis of } E \text{ such that } (\Delta\phi_i, \Delta\phi_j) = \delta_{ij} \quad (1 \leq i,j \leq m) \ ,$$

where δ_{ij} is the Kronecker delta.

Let K be a positive constant such that

$$[\lambda_0 - K, \lambda_0 + K] \cap \sigma = \{\lambda_0\} \ .$$

Define

E^h = linear hull of eigenfunctions corresponding to all eigenvalues in $\sigma_h \cap [\lambda_0 - K, \lambda_0 + K]$,

m_h = dimension of E^h ; $0 \leqq m_h < \infty$,

$\{\phi_{hi}\}_{i=1}^{m_h}$ = a basis of E^h such that each ϕ_{hi} is an eigenfunction of (10) and satisfies $(\Delta_h\phi_{hi}, \Delta_h\phi_{hj}) = \delta_{ij}$ $(i,j \leqq m_h)$,

λ_{hi} = eigenvalue corresponding to ϕ_{hi} $(i \leqq m_h)$.

The following lemmas may be derived from the results established in the preceding section and by reduction to absurdity.

Lemma 3 For each $\varepsilon > 0$, we can find a positive number $h_0 = h_0(\varepsilon)$ such that

$$(15) \qquad \sigma_h \subset \sigma(\varepsilon) \qquad \text{for any} \quad h \leqq h_0 \ .$$

That is, any element of σ_h is close to a certain element of σ when h is sufficiently small.

Proof Fix $\varepsilon_0 > 0$, and consider the case $\varepsilon \leqq \varepsilon_0$. Then there exists a positive constant L such that

$$\sigma_h \subset [-L,L] \quad , \qquad \sigma(\varepsilon) \subset [-L,L]$$

Assume the contrary to (15). Then we can find a sequence $\{\{\lambda_{h(i)}, u_{h(i)}\}\}_{i=1}^{\infty}$ such that

$$\lim_{i\to\infty} h(i) = 0 \quad , \qquad \lambda_{h(i)} \in \sigma_{h(i)} \setminus \sigma(\varepsilon) \quad ,$$

$u_{h(i)} \in X_0^{h(i)}$ is an eigenfunction of (10) corresponding to $\lambda_{h(i)}$ and satisfies $\| \Delta_{h(i)} u_{h(i)} \| = 1$,

where $\{X_0^{h(i)}\}_{i=1}^{\infty}$ is a sequence of finite element spaces associated with an appropriate sequence of triangulations $\{T^{h(i)}\}_{i=1}^{\infty}$. Hereafter, we omit the index i, and h → 0 means i → ∞. Since $|\lambda_h| + \| \Delta_h u_h \|$ is uniformly bounded, we can apply Lemma 2 to show the existence of a subsequence, again denoted by $\{\{\lambda_h, u_h\}\}$, such that for $h \to 0$

$$\lambda_h \to \lambda \in R^1 \quad ,$$

$$u_h \to u \quad \text{strongly in } H_0^1(\Omega) \quad ,$$

$$\Delta_h u_h \to \Delta u \quad \text{weakly in } L_2(\Omega) \quad ,$$

where u is an element of $H_0^2(\Omega)$. Notice that λ belongs to a closed set $[-L,L] \setminus \sigma(\varepsilon)$, and in particular $\lambda \notin \sigma(\varepsilon)$. On the other hand, we can find for each $\bar{u} \in H_0^2(\Omega)$ an approximate sequence $\{\bar{u}_h\}$ such that $\bar{u}_h \in X_0^h$ and for $h \to 0$

$$\bar{u}_h \to \bar{u} \quad \text{strongly in } H_0^1(\Omega) \quad ,$$

$$\Delta_h \bar{u}_h \to \Delta \bar{u} \quad \text{strongly in } L_2(\Omega) \quad .$$

Here we have used the results of Scholz [6] to assure the existence of a sequence strongly convergent in the above sense. Taking the limit of

$$\lambda_h(\Delta_h u,\Delta_h\bar{u}_h) = b(u_h,\bar{u}_h) \quad ,$$

we have

$$\lambda(\Delta u,\Delta\bar{u}) = b(u,\bar{u}) \quad \text{for any} \quad \bar{u} \in H_0^2(\Omega) \ .$$

Since $\lambda \notin \sigma$, the above relation implies $u = 0$. Taking the limit of

$$\lambda_h \| \Delta_h u_h \|^2 = b(u_h,u_h) \quad ,$$

we have $\lambda = b(u,u) = 0$. Since $\lambda \notin \sigma$, this means $\lambda = 0$ is an accumulation point of σ. But λ also belongs to $[-L,L] \setminus \sigma(\varepsilon)$, and we have a contradiction. ////

Lemma 4 We can find a positive constant h_0 such that

(16) $$m_h \leq m \qquad \text{for any} \quad h \leq h_0 \ .$$

That is, the dimension of E^h never exceeds that of E when h is sufficiently small.

Proof We only sketch the proof. Assume the contrary to (16). Then we can find a sequence of triangulations $\{T^{h(i)}\}_{i=1}^{\infty}$ such that

$$\lim_{i\to\infty} h(i) = 0 \quad , \qquad m_{h(i)} \geq m + 1 \quad ,$$

where $m_{h(i)}$ is the dimension of $E^h = E^{h(i)}$ associated with $T^{h(i)}$. In the sequel, we omit i as in the proof of Lemma 3. We can find a sequence $\{u_h^*\}$ such that

$$u_h^* = \sum_{j=1}^{m+1} \alpha_j \phi_{hj} \in E^h \ , \qquad \| \Delta_h u_h^* \| = 1 \ ,$$

$$(\Delta_h u_h^*, \Delta_h \phi_{hj}) = 0 \qquad \text{for} \quad 1 \leqq j \leqq m \ ,$$

where $\{\phi_{hj}\}_{j=1}^{m_h}$ has been already defined in this section, and α_j for $1 \leqq j \leqq m+1$ are real coefficients. Notice that u_h^* is a linear combination of the first (m+1) functions of $\{\phi_{hj}\}_{j=1}^{m_h}$, and that we can always find u_h^* by an appropriate choice of α_j. As in the proof of the preceding lemma, we can choose a suitable subsequence of $\{u_h\}$ to show the existence of a *non-zero* function $u^* \in H_0^2(\Omega)$ such that

$$\lambda_0(\Delta u^*, \Delta \bar{u}) = b(u^*, \bar{u}) \qquad \text{for all} \quad \bar{u} \in H_0^2(\Omega) \ ,$$

$$(\Delta u^*, \Delta \phi_j) = b(u^*, \phi_j) \qquad \text{for} \quad 1 \leqq j \leqq m \ .$$

Here it is especially to be noted that the eigenvalues associated with E^h necessarily converge to λ_0 as h → 0 due to Lemma 3. The above relations clearly contradict the fact that E is the eigenspace for λ_0, and the proof is completed. ////

6. Existence and Error Estimation of the Approximate Eigenpairs

In the preceding section, we have established some fundamental properties of the approximate eigenpairs. This section is devoted to the proof that there actually exists a nice approximation to each eigenpair of (4), except that for the zero eigenvalue. Essentially, we will rely upon the standard technique, that is, the use of the spectral projection. In our description, we will be mainly concerned with the idea or the outline of analysis.

Let us consider, in the complex plane, a circumference Γ with its center and direction taken as λ_0 and anticlockwise, respectively. The radius of Γ is fixed so small that there is no element of σ, except λ_0, either on or inside Γ. Consider the following two problems : for each $\lambda \in \Gamma$ and $f \in H_0^1(\Omega)$, find $u \in H_0^2(\Omega)$ and $u_h \in X_0^h$ such that

$$(17) \qquad \lambda(\Delta u,\Delta\bar{u}) - b(u,\bar{u}) = \lambda_0^{-1}b(f,\bar{u}) \quad ; \ \forall \ \bar{u} \in H_0^2(\Omega) \ ,$$

and

$$(18) \qquad \lambda(\Delta_h u_h,\Delta_h\bar{u}_h) - b(u_h,\bar{u}_h) = \lambda_0^{-1}b(f,\bar{u}_h) \quad ; \ \forall \ \bar{u}_h \in X_0^h \ ,$$

respectively. For each $\lambda \in \Gamma$ and $f \in H_0^1(\Omega)$, we can show that u exists uniquely in $H_0^2(\Omega)$, and that u_h also does uniquely in X_0^h when h is small enough. These results are derived by the use of theory of compact operators, or the approximate compactness properties established in Lemma 2. Although we have not stated explicitly, we have considered real functions only. Since λ is a complex number, u and u_h are in general complex-valued functions even for real f, and hence we are obliged to concern ourselves in complex-valued world. However, the final results can be again expressed in real-valued world, as will become clear later.

Now we can define the following two operators for each $\lambda \in \Gamma$:

$$Q(\lambda) : H_0^1(\Omega) \to H_0^2(\Omega) \ ;$$

$$(19) \qquad Q(\lambda)f = u \in H_0^2(\Omega) \text{ of } (17) \text{ for each } f \in H_0^1(\Omega) \ ,$$

$$Q_h(\lambda) : H_0^1(\Omega) \to X_0^h ;$$

$$Q_h(\lambda)f = u_h \in X_0^h \text{ of (18) for each } f \in H_0^1(\Omega) . \tag{20}$$

It is to be noted that $Q(\lambda)f$ and $Q_h(\lambda)f$ (for h small enough) are uniformly continuous in $\lambda \in \Gamma$ for each f, the continuities being those with respect to the metrics of $H_0^2(\Omega)$ and X_0^h, respectively. Notice also Lemma 5 to be presented later. Define

$$P : H_0^1(\Omega) \to H_0^2(\Omega) ;$$

$$P f = \frac{1}{2\pi i} \int_\Gamma Q(\lambda) f \, d\Gamma \quad , \tag{21}$$

$$P_h : H_0^1(\Omega) \to X_0^h ;$$

$$P_h f = \frac{1}{2\pi i} \int_\Gamma Q_h(\lambda) f \, d\Gamma \quad , \tag{22}$$

where f is an arbitrary function in $H_0^1(\Omega)$, and i the imaginary unit. We can show that P is actually a mapping of $H_0^1(\Omega)$ into E such that

$$P u = u \qquad \text{for any} \quad u \in E . \tag{23}$$

On the other hand, P_h is a mapping of $H_0^1(\Omega)$ into E^h well defined for h small enough, but does not possess the property corresponding to (23) in the strict sense.

In the following, we will present three lemmas without proofs. The first one may be proved with the aid of Lemma 2, while the second and the last one are known results.

Lemma 5 Let λ be an arbitrary number on Γ. For each $f \in L_2(\Omega)$ and $g \in H^{-1}(\Omega)$ (= the dual of $H_0^1(\Omega)$), consider $u_h \in X_0^h$ such that

$$(24) \qquad \lambda(\Delta_h u_h, \Delta_h \bar{u}_h) - b(u_h, \bar{u}_h) = (f, \Delta_h \bar{u}_h) + (g, \bar{u}_h) \; ; \; \forall \; \bar{u}_h \in X_0^h \; ,$$

where for the expression $(g, \bar{u}_h)$ we have used (,) as the pairing on $H^{-1}(\Omega) \times H_0^1(\Omega)$. Then for h small enough, u_h exists uniquely in X_0^h, and satisfies

$$(25) \qquad \| \Delta_h u_h \| \leqq C(\| f \| + \| g \|_{H^{-1}(\Omega)}) \qquad ,$$

where the positive constant C can be independent of λ, h, f, and g.

Remark 1 The corresponding properties hold for the continuous version of (24).

Lemma 6 (Mizutani [7]) For each $\lambda \in \Gamma$ and $f \in H_0^2(\Omega)$, $Q(\lambda)f$ belongs to $H^4(\Omega) \cap H_0^2(\Omega)$ with the estimation

$$(26) \qquad \| Q(\lambda) f \|_4 \leqq C \| f \|_2 \qquad ,$$

where C can be independent of λ and f.

Remark 2 Notice that f belongs to $H_0^2(\Omega)$ and Ω is a non-obtuse polygonal domain.

Lemma 7 (Scholz [6]) For each $\lambda \in \Gamma$ and $f \in H_0^2(\Omega)$, consider $Q(\lambda)f$ and $Q_h(\lambda)f$ when h is small enough. Then

$$(27) \qquad \| Q_h(\lambda)f - Q(\lambda)f \| + h^{1/2} |\ln h| \, \| \Delta_h Q_h(\lambda)f - \Delta Q(\lambda) f \| \leqq C \, h \, |\ln h|^2 \| f \|_2 \, ,$$

$$(28) \qquad \| Q_h(\lambda)f - Q(\lambda)f \|_1 \leqq C h^{3/4} |\ln h|^{3/2} \| f \|_2 \quad ,$$

where C can be independent of λ, f, and h.

Now we can follow the standard procedure. Notice that $\phi_i = P\phi_i \in H^4(\Omega) \cap H_0^2(\Omega)$, and consider

$$(29) \qquad \phi^*_{hi} = P_h\phi_i \quad ; 1 \leqq i \leqq m \quad .$$

Then we can show that ϕ^*_{hi} and ϕ_i satisfy the estimations corresponding to (27) and (28). Therefore, for h small enough, $\phi^*_{h1}, \ldots, \phi^*_{hm}$ are linearly independent and belong to E^h. On the other hand, dim $E^h = m_h \leqq m$ from Lemma 4, and hence $\{\phi_{hi}\}_{i=1}^m$ can be regarded as a basis of E^h.

Notice that $m_h = m$ and consider the basis $\{\phi_{hi}\}_{i=1}^m$ introduced in section 5. Then each ϕ_{hi} can be uniquely expressed as

$$(30) \qquad \phi_{hi} = \sum_{j=1}^m \alpha_{ij}\phi^*_{hj} \qquad (1 \leqq i \leqq m)$$

by choosing the coefficients α_{ij} $(1 \leqq i,j \leqq m)$ appropriately. Define ψ_i by

$$(31) \qquad \psi_i = \sum_{j=1}^m \alpha_{ij}\phi_j \in E \quad (1 \leqq i \leqq m) \quad ,$$

where α_{ij} are the same as those in (30). Noting that α_{ij} are uniformly bounded with respect to i, j, and h, we have

$$(32) \quad \| \phi_{hi} - \psi_i \| + h^{1/2} |\ln h| \| \Delta_h\phi_{hi} - \Delta\psi_i \| \leqq C h |\ln h|^2 \quad ,$$

(33) $$\| \phi_{hi} - \psi_i \|_1 \leqq C h^{3/4} |\ln h|^{3/2} \quad ,$$

where C can be chosen independent of i and h.

The above estimations implicate that there exists a function $\psi_i \in E$ sufficiently close to each ϕ_{hi}. In general, $\{\psi_i\}_{i=1}^m$ is not orthonormal in the sense $(\Delta\psi_i, \Delta\psi_j) = \delta_{ij}$. However, the basis $\{\phi_{hi}\}_{i=1}^m$ has been chosen such that $(\Delta_h\phi_{hi}, \Delta_h\phi_{hj}) = \delta_{ij}$, and hence we can find an appropriate orthonormal basis of E, each basis function of which is close to one of ϕ_{hi} $(1 \leqq i \leqq m)$, cf. Chapter 6 of Strang and Fix [10]. On the other hand, the error of each λ_{hi} can be evaluated by the use of the relation $\lambda_{hi} = b(\phi_{hi}, \phi_{hi}) / \| \Delta\phi_{hi} \|^2$.

The final results can be summarized as follows.

Theorem 1 For h small enough, we can choose a basis $\{\phi_i^*\}_{i=1}^m$ of E such that

(34) $$(\Delta\phi_i^*, \Delta\phi_j^*) = \delta_{ij} \quad ,$$

(35) $$\| \Delta_h\phi_{hi} - \Delta\phi_i^* \| \leqq C h^{1/2} |\ln h| \quad ,$$

(36) $$\| \phi_{hi} - \phi_i^* \|_1 \leqq C h^{3/4} |\ln h|^{3/2} \quad ,$$

where $1 \leqq i,j \leqq m$, $\{\phi_{hi}\}_{i=1}^m$ is the basis of E^h defined in section 5, and C can be chosen independent of h and i. Notice that the choice of $\{\phi_i^*\}_{i=1}^m$ may differ with the triangulation T^h. As for the eigenvalue approximation, we have

(37) $$|\lambda_{hi} - \lambda_0| \leqq C h^{3/4} |\ln h|^{3/2} \quad \text{for } 1 \leqq i \leqq m \ .$$

Remark 3 At present, it is not certain whether we may expect the estimations

$$\| \phi_{hi} - \phi_i^* \| \leqq Ch|\ln h|^2 \quad \text{for } 1 \leqq i \leqq m .$$

7. Concluding Remarks

We have performed error analysis of a mixed finite element model applied to linear buckling analysis of thin elastic plates. The obtained order estimates of errors of the approximate eigenpairs have been summarized as a theorem. We have only sketched the proofs, and the detailed analysis will be reported elsewhere. The author believes that the principles and techniques established in this note will be available for the analysis of various finite element models.

This work is in part supported by the Grant-in Aid for the Scientific Research from the Ministry Education. The author would like to express his deepest appreciation to Mr. T. Hanada and Professor T. Ushijima of the University of Electro-Communications for their valuable advices and discussions.

References

[1] Kikuchi, F., A finite element method for plate bending analysis by decomposition of differential operator, J. Nucl. Sci. Technol., 8 (1971) 597-599.

[2] Ciarlet, P.G. and Glowinski, R., Dual iterative techniques for solving a finite element approximation of the biharmonic equation, Comput. Methods Appl. Mech. Engng, 5 (1975) 277-295.

[3] Ishihara, K., The buckling of plates by the mixed finite element

method, Memoirs of Numerical Mathematics, 5 (1978) 73-82.

[4] Miyoshi, T., A finite element method for the solution of fourth order partial differential equations, Kumamoto J. Sci. (Math.), 9 (1973) 87-116.

[5] Grigorieff, R.D., Discrete Approximation von Eigenwertproblemen, I. Qualitative Konvergenz, Numer. Math., 24 (1975) 355-374, II. Konvergenzordnung, Ibid, 415-433, III. Asymptotische Entwicklungen, Numer. Math., 25 (1975) 79-97.

[6] Scholz, R., A mixed method for 4th order problems using linear finite elements, R.A.I.R.O., Analyse numerique, 12 (1978) 85-90.

[7] Mizutani, A., On the regularities of solutions of biharmonic equations in domains with angular points — notes on a paper of Kondrat'ev, Kōkyū-roku of Reserach Institute of Mathematical Sciences, University of Kyoto, No. 329 (1978) 2-9.

[8] Ciarlet, P.G., The Finite Element Method for Elliptic Problems, North-Holland (1978)

[9] Kondrat'ev, V.A., Boundary problems for elliptic equations in domains with conical or angular points, Trans. Moscow Math. Soc., 16 (1967) 227-313.

[10] Strang, G. and Fix, G.J., An Analysis of the Finite Element Method, Prentice-Hall (1973).

[11] Dym. C.L. and Shames, I.H., Solid Mechanics, A Variational Approach, McGraw-Hill (1973).

[12] Canuto, C., Eigenvalue approximation by mixed methods, R.A.I.R.O., Analyse numerique, 12 (1978) 27-50.

[13] Kikuchi, F., A mixed finite element model for fourth order eigenvalue problems, to appear.

The Initial-Value Adjusting Method for Solving Problems of the Least Squares Type of Ordinary Differential Equations

By

Taketomo MITSUI*

Abstracts

An application of the initial-value adjusting method for the problem of the least squares type is considered. Convergence property of the method is discussed. An illustrative numerical example is given.

§1. The Initial-Value Adjusting Algorithm.

We are concerned with numerical procedures solving the problems of the least squares type for ordinary differential equations as following:

Find a solution of the differential equation

$$\frac{dx}{dt} = X(x,t), \quad a < t < b \tag{1.1}$$

which minimizes the value

$$J = \frac{1}{2}\sum_{j=1}^{N} {}^t[L_j x(t_j)-d_j][L_j x(t_j)-d_j], \tag{1.2}$$

where x and X are real n-dimensional vectors, t_j are given points on $I = [a,b]$, $a=t_1 < t_2 < \cdots < t_N=b$, and L_j

*Research Institute for Mathematical Sciences, Kyoto University, Kyoto, 606 Japan.

and d_j are given n×n matrices and n-dimensional vectors respectively. As far as the equation (1.1) is stable, the problem is equivalent to find an initial value η such that the solution $x(t)$ starting from η minimizes J.

Let us define a linear operator $\mathbb{L} : C(I) \to \mathbb{R}^{nN}$ such that for $x \in C(I)$

$$(1.3) \quad \mathbb{L}x = {}^t(L_1 x(t_1), \cdots, L_N x(t_N)).$$

Denote the nN-dimensional vector ${}^t(d_1, \cdots, d_N)$ by $\mathbb{d}$. Note that J is represented by

$$(1.4) \quad J = \frac{1}{2}\,{}^t\{\mathbb{L}x - \mathbb{d}\}\{\mathbb{L}x - \mathbb{d}\}.$$

After the initial-value adjusting method proposed by Ojika and Kasue [7], our *algorithm* can be expressed in the following steps.

Step 0. Choose a suitable perturbation parameter ε and an initial value $\eta_0 \in \mathbb{R}^n$, and set $k=0$.

Step 1. Compute the numerical solution $x_k(t)$ of (1.1) for the initial condition $x_k(a)=\eta_k$, and obtain the resulting value $\mathbb{L}_k=\mathbb{L}x_k$ and $J_k=\frac{1}{2}{}^t\{\mathbb{L}_k-\mathbb{d}\}\{\mathbb{L}_k-\mathbb{d}\}$.

Step 2. If the value J_k varies slightly in comparison with J_{k-1} (i.e. $J_{k-1}-J_k$ is sufficiently close to 0 by the criterion given in advance), terminate the iteration. Otherwise, go to the next step. (If $k=0$, skip this step.)

Step 3. Set $j=1$.

Step 4. Compute the numerical solution $y_k^{(j)}(t)$ of (1.1) for the initial condition $y_k^{(j)}(a) = \eta_k+\varepsilon e_j$. Here e_j means the j-th uint vector of $\mathbb{R}^n$.

Step 5. Replace j by $j+1$, and return to Step 4 until $j=n$.

Step 6. Determine the $nN\times n$ matrix $S(\varepsilon;x_k)$ (the *adjusting matrix*) such that

$$(1.5)\qquad S(\varepsilon;x_k) = \Big(\frac{1}{\varepsilon}\{\mathbb{L}y_k^{(1)}-\mathbb{L}_k\},\cdots,\frac{1}{\varepsilon}\{\mathbb{L}y_k^{(j)}-\mathbb{L}_k\},\cdots,$$

$$\frac{1}{\varepsilon}\{\mathbb{L}y_k^{(n)}-\mathbb{L}_k\}\Big).$$

Step 7. Determine the initial value η_{k+1} for the next iteration by

$$(1.6)\qquad \eta_{k+1} = \eta_k-\{{}^tS(\varepsilon\,;x_k)S(\varepsilon;x_k)\}^{-1}\,{}^tS(\varepsilon;x_k)\{\mathbb{L}_k-\mathbb{a}\}.$$

Then replace k by $k+1$, and return to Step 1.

Computational Remarks. In the above process, the numerical integration of the differential equation (1.1) are carried out by a suitable step-by-step method, for example, the Runge-Kutta method. Since the matrix ${}^tS(\varepsilon;x_k)S(\varepsilon;x_k)$ is an $n\times n$ symmetric, positive definite matrix, the square-root-free Cholesky's method is preferable for the solution of the linear equation in (1.6).

We shall denote the Euclidean norm of an n-dimensional vector x by $\|x\|$. $C(I)$ stands for the Banach space of vector-valued continuous functions on I, equipped with the norm

$$\|x\|_C = \max_{t\in I}\|x(t)\|.$$

$C^1(I)$ means a subset of $C(I)$ of continuously differentiable functions on I. The norms for matrices $\mathbb{R}^m \to \mathbb{R}^{m'}$ and other linear operators should be taken as the induced norms by the corresponding vector norms.

Let $\mathfrak{D}$ be the domain of the tx-space bounded on x, intercepted by two hyperplanes $t=a$ and $t=b$. The boundary points of $\mathfrak{D}$ on the hyperplanes $t=a$ and $t=b$ are supposed to be included in $\mathfrak{D}$ and to make an open set on each hyperplane. Put

$D = \{x \in C^1(I);\ (t,x(t)) \quad \text{for } t \in I\},$

$D' = \{\ x \in C(I);\ (t,x(t)) \quad \text{for } t \in I\}.$

We shall analyse the initial-value adjusting algorithm within the above framework. Because of space limitations, we shall omit all the proofs of the statements, which will be shown in other papers [5],[6].

§2. Some Preliminaries.

We shall adopt the following assumptions to the problem (1.1) and (1.2).

Assumption 1. *$X(x,t)$ is defined and sufficiently continuously differentiable with respect to x on $\mathfrak{D}$. X and its derivatives are continuous with respect to t on I.*

Let us consider an operator $\mathfrak{F}$ mapping $\mathbb{R}^n$ to a function of $C^1(I)$ along the flow generated by the differential equation (1.1):

$$\frac{d}{dt}\mathfrak{F}\eta = X[\mathfrak{F}\eta(t),t], \qquad a < t < b,$$

$$\mathfrak{F}\eta(a) = \eta.$$

Define a function $J(\eta)$ by

(2.1) $$J(\eta) = \frac{1}{2}\,{}^t\{\mathbb{L}\mathfrak{F}\eta - \mathbb{a}\}\{\mathbb{L}\mathfrak{F}\eta-\mathbb{a}\}.$$

Let $\Phi(t;x)$ stand for the matrizant of the linear homogeneous matrix differential equation

$$\frac{d\Phi}{dt} = X_x[x(t),t]\Phi\ , \qquad a< t < b.$$

We shall call the matrix $\mathbb{L}\Phi(\cdot;\mathfrak{F}\eta)$ the *G-matrix* and denote it by $G(\eta)$.

Assumption 2. *There exists a vector* η^* *such that* $(a,\eta^*)\in\mathcal{D}$, $\mathfrak{F}\eta^*\in D$,

(2.2) $${}^tG(\eta^*)\{\mathbb{L}\mathfrak{F}\eta^*- \mathbb{a}\}=0$$

and

(2.3) $$\det {}^tG(\eta^*)G(\eta^*)\neq 0$$

hold. Furthermore, there exists a positive constant $\bar{\varepsilon}$ *such that for* $0 < |\varepsilon| \leq \bar{\varepsilon}$ *the initial value problem*

$$\frac{dy}{dt} = X(y,t), \quad a < t < b,$$

$$y(a) = \eta^*+\varepsilon e_j$$

has a unique solution in D *for every* j.

We shall call η^* the *exact (local) isolated minimal point* of $J(\eta)$. Let L_0 be the bound of the operator-norm of $\mathbb{L}$ on D.

For the matrizant $\Phi(t;x)$ the estimation

(2.4) $$\|\Phi(\cdot;x)\| \leq M_0 \qquad \text{for} \quad x\in D'$$

is evident. Since D is open in $C^1(I)$, for a positive constant M_1 we can take a positive number Δ such that in the Δ-neighbourhood B_Δ of η^*

$$B_\Delta = \{ \eta \in \mathbb{R}^n;\quad \|\eta - \eta^*\| \leqq \Delta \},$$

the inverse of *compound G-matrix* defined by $\mathcal{G}(\eta) = {}^tG(\eta)G(\eta)$, exists and have the estimation

$$(2.5) \qquad \|\mathcal{G}^{-1}(\eta)\| \leqq M_1.$$

Replace Δ by a suitable value less than Δ if necessary, then it is possible that

$$\|\mathcal{F}\eta - \mathcal{F}\eta^*\|_C \leqq \delta_0$$

holds for $\eta \in B_\Delta$ and small positive δ_0. We shall fix the number Δ.

Lemma 1. *For* $\eta \in B_\Delta$, *put* $x(t) = \mathcal{F}\eta(t)$. $y^{(j)}(t)$ *in Step 4 of the algorithm has the following expression.*

$$(2.6) \qquad y^{(j)}(t) = x(t) + \varepsilon\{\varphi^{(j)}(t) + v^{(j)}(t)\} \quad on \quad I,$$

where $\varphi^{(j)}(t)$ *is the j-th column vector of* $\Phi(t;x)$ *and* $v^{(j)}(t)$ *satisfies the differential equation*

$$(2.7) \qquad \frac{dv^{(j)}(t)}{dt} = \frac{1}{\varepsilon}\{ X[x(t) + \varepsilon\{\varphi^{(j)}(t) + v^{(j)}(t)\}, t] - X[x(t), t]\} - $$
$$- X_x[x(t), t]\varphi^{(j)}(t), \qquad a < t < b$$

subject to the initial condition

$$(2.8) \qquad v^{(j)}(a) = 0.$$

Moreover, for arbitrary small $|\varepsilon|$ *there exists a positive constant* C^* *such that* $v^{(j)}(t)$ *satisfies*

(2.9) $\|v^{(j)}\|_C \leq C^*|\varepsilon|$.

Through the equation

$$S(\varepsilon;\mathfrak{F}\eta) - G(\eta) = \mathbb{L}(v^{(1)},\cdots,v^{(j)},\cdots,v^{(n)}),$$

Lemma 1 immediately implies that there exists a small positive number ε_0 such that for $0 \leqq |\varepsilon| \leqq \varepsilon_0$ the estimation

(2.10) $\|S(\varepsilon;\mathfrak{F}\eta) - G(\eta)\| \leqq C^{**}\varepsilon \equiv \delta_1$

holds in B_Δ. Furthermore, for $0 \leqq |\varepsilon| \leqq \varepsilon_0$ and $\eta\in B_\Delta$ the inverse of the *compound adjusting matrix* ${}^tS(\varepsilon;\mathfrak{F}\eta)S(\varepsilon;\mathfrak{F}\eta)$ exists and

(2.11) $\|\{{}^tS(\varepsilon;\mathfrak{F}\eta)S(\varepsilon;\mathfrak{F}\eta)\}^{-1}\| \leqq \dfrac{M_1}{1-\delta_1M_1(2L_0M_0+\delta_1)}$

holds.

Corresponding to our iterative process, an operator $\mathcal{A}$ mapping $\mathbb{R}^n$ into itself is defined by the following:

The domain of $\mathcal{A}$ is identical to B_Δ.

(2.12) $\mathcal{A}\eta = \eta - \{{}^tS(\varepsilon;\mathfrak{F}\eta)S(\varepsilon;\mathfrak{F}\eta)\}^{-1}\,{}^tS(\varepsilon;\mathfrak{F}\eta)\{\mathbb{L}\mathfrak{F}\eta - \mathbb{a}\}$

for $\eta\in B_\Delta$.

The iterative process is simply represented by

(2.13) $\eta_{k+1} = \mathcal{A}\eta_k$, $k=0,1,2,\cdots$.

Thus our aim is concentrated on the analysis of the operator $\mathcal{A}$.

§3. Fixed Point of the Operator $\mathcal{A}$.

It is noteworthy that the exact minimal point η^* is *not* a fixed point of $\mathcal{A}$, because a fixed point of $\mathcal{A}$ must satisfy the equation

$$(3.1) \qquad \{{}^tS(\varepsilon;\mathcal{F}\eta)S(\varepsilon;\mathcal{F}\eta)\}^{-1}\, {}^tS(\varepsilon;\mathcal{F}\eta)\{\mathbb{L}\mathcal{F}\eta-\mathbb{d}\}=0,$$

i.e.

$$(3.1)' \qquad {}^tS(\varepsilon;\mathcal{F}\eta)\{\mathbb{L}\mathcal{F}\eta-\mathbb{d}\}=0.$$

On the other hand, from (2.2) η^* satisfies

$${}^tG(\eta^*)\{\mathbb{L}\mathcal{F}\eta^*-\mathbb{d}\}=0.$$

$S(\varepsilon;\mathcal{F}\eta)$ is surely an approximation of $G(\eta)$, but is not identical to that as far as $\varepsilon\neq 0$.

Therefore, we should answer the question whether a fixed point of $\mathcal{A}$ exists. We shall show that it actually exists in some neighbourhood of η^* by the implicit function theorem.

Define a function $\mathcal{T}(\eta,\varepsilon)$ by

$$(3.2) \qquad \mathcal{T}(\eta,\varepsilon) = {}^tS(\varepsilon;\mathcal{F}\eta)\{\mathbb{L}\mathcal{F}\eta-\mathbb{d}\}$$

for (η,ε) such that $\eta\in B_\Delta$ and $0 < |\varepsilon| \leq \varepsilon_0$.

Lemma 2. *Define* $\mathcal{T}(\eta,0)$ *by*

$$(3.3) \qquad \mathcal{T}(\eta,0) = \lim_{\varepsilon\to 0}\mathcal{T}(\eta,\varepsilon),$$

then

$$(3.4) \qquad \mathcal{T}(\eta,0) = {}^tG(\eta)\{\mathbb{L}\mathcal{F}\eta-\mathbb{d}\}$$

holds.

The following two lemmas say how the adjusting matrix depends on ε or η when they make a small variation.

Lemma 3. *For* $\eta \in B_\Delta$ *and* $|\varepsilon_1|, |\varepsilon_2| \leq \varepsilon_0$, *the estimation*

$$(3.5) \qquad \| S(\varepsilon_1; \mathcal{F}\eta) - S(\varepsilon_2; \mathcal{F}\eta) \| \leq \mathit{const.}\, |\varepsilon_1 - \varepsilon_2|$$

holds. Here the constant does not depend on η, ε_1 *and* ε_2.

Lemma 4. *For sufficiently small numbers* $\tilde{\Delta}$ *and* $\tilde{\varepsilon}$, *the equation*

$$(3.6) \qquad S(\varepsilon; \mathcal{F}(\eta+\xi)) - S(\varepsilon; \mathcal{F}\eta) = \mathbb{L}P(\cdot; \mathcal{F}\eta, \varepsilon)\xi + o(\|\xi\|)$$

holds if η, $\eta+\xi \in B_{\tilde{\Delta}}$ *and* $|\varepsilon| \leq \tilde{\varepsilon}$. *Here* $P(t; \mathcal{F}\eta, \varepsilon)$ *is a linear mapping which maps* $\xi \in \mathbb{R}^n$ *to an* $n \times n$ *matrix* $P(t; \mathcal{F}\eta, \varepsilon)\xi$ *with components of continuous functions on* I.

Lemmas 2 and 3 and the implicit function theorem imply the desired result.

Theorem 1. *There exist positive numbers* Δ_1 *and* ε_1 *such that, for any* ε *satisfying* $|\varepsilon| \leq \varepsilon_1$, *the equation* $\mathcal{J}(\eta, \varepsilon) = 0$ *has a unique solution* $\hat{\eta} = \hat{\eta}(\varepsilon)$ *in the ball* $\|\eta - \eta^*\| \leq \Delta_1$, *and* $\hat{\eta} \to \eta^*$ *as* $\varepsilon \to 0$.

We shall denote the fixed point of $\mathcal{A}$ by $\hat{\eta} = \hat{\eta}(\varepsilon)$ whose existence has been guaranteed by Theorem 1 and call it the *approximate minimal point* of $J(\eta)$ because it is an approximation of η^* and tends to it as $\varepsilon \to 0$.

§4. Convergence of the Iterative Process.

Since $\hat{\eta}$ is defined by $\mathcal{J}(\hat{\eta},\varepsilon)=0$, the equation

$$(4.1)\quad \{{}^{t}S(\varepsilon;\mathcal{F}\hat{\eta})S(\varepsilon;\mathcal{F}\hat{\eta})\}^{-1}\ {}^{t}S(\varepsilon;\mathcal{F}\hat{\eta})\{\mathbb{L}\mathcal{F}\hat{\eta}-\mathbb{d}\}=0$$

holds if $\|\hat{\eta}-\eta^{*}\| \leqq \Delta$. Then we have

$$\begin{aligned}\mathcal{A}\eta - \hat{\eta} &= \eta - \hat{\eta} - \{{}^{t}S(\varepsilon;\mathcal{F}\eta)S(\varepsilon;\mathcal{F}\eta)\}^{-1}\mathcal{J}(\eta,\varepsilon)\\ &= \eta - \hat{\eta} - \{{}^{t}S(\varepsilon;\mathcal{F}\eta)S(\varepsilon;\mathcal{F}\eta)\}^{-1}\{\mathcal{J}(\eta,\varepsilon)-\mathcal{J}(\hat{\eta},\varepsilon)\}\\ &= \{{}^{t}S(\varepsilon;\mathcal{F}\eta)S(\varepsilon;\mathcal{F}\eta)\}^{-1}\{\ {}^{t}S(\varepsilon;\mathcal{F}\eta)S(\varepsilon;\mathcal{F}\eta)(\eta-\hat{\eta})-\\ &\quad - \mathcal{J}_{\eta}(\hat{\eta},\varepsilon)(\eta-\hat{\eta})\}+o(\|\eta-\hat{\eta}\|).\end{aligned}$$

Here $\mathcal{J}_{\eta}(\eta,\varepsilon)$ stands for the Jacobian of $\mathcal{J}$ with respect to η. Utilizing the results of the previous section, we can investigate the estimation for the term

$${}^{t}S(\varepsilon;\mathcal{F}\eta)S(\varepsilon;\mathcal{F}\eta)(\eta-\hat{\eta})-\mathcal{J}_{\eta}(\hat{\eta},\varepsilon)(\eta-\hat{\eta}).$$

Then, with the estimation (2.11), the contraction mapping principle brings the following

Theorem 2. *There exists a positive number* Δ_2 *such that the iterative process (2.13) starting from any* η *in the* Δ_2*-neighbourhood of* $\hat{\eta}$ *converges to* $\hat{\eta}$.

Remark. The results in the previous section also gives the explicit formula for the Fréchet derivative of $\mathcal{A}$ at $\hat{\eta}$. By the consideration of $\mathcal{A}'(\hat{\eta})$, we can conclude that the convergence of the iteration would not be expected to be quadratic in the neighbourhood of $\hat{\eta}$ while a fixed ε is chosen to be apart from zero.

Next, we shall investigate the approximation order of $\hat{\eta}$ for η^*. The following improves the statement of Lemma 3.

Lemma 5. *For* $\eta \in B_\Delta$ *and* $|\varepsilon|, |\varepsilon+\nu| \leqq \varepsilon_0$, *the equation*

$$(4.2) \qquad S(\varepsilon+\nu;\mathfrak{F}\eta) - S(\varepsilon;\mathfrak{F}\eta) = \mathbb{L}p(\cdot;\mathfrak{F}\eta,\varepsilon)\nu + o(|\nu|)$$

holds, where $p(t;\mathfrak{F}\eta,\varepsilon)$ *is an* $n\times n$ *matrix with components of continuous functions on* I.

By virtue of Lemma 5 we obtain informations about the derivatives of $\mathcal{T}(\eta,\varepsilon)$ with respect to ε in some neighbourhood of $(\eta^*,0)$, which implies the following

Theorem 3. *In the neighbourhood*

$\{(\eta,\varepsilon);\ \|\eta-\eta^*\| \leqq \Delta_3$ *and* $|\varepsilon| \leqq \varepsilon_3\}$,

$\hat{\eta}(\varepsilon)$ *whose existence is guaranteed in Theorem 1 has the estimation such as*

$$(4.3) \qquad \|\hat{\eta}(\varepsilon)-\eta^*\| \leqq const.|\varepsilon| \quad as \quad \varepsilon\to 0.$$

§5. An Illustrative Example.

Let us consider the following problem originally mentioned in [4], which occurs relating to the tubular flow chemical reactor with axial mixing.

The differential equation is

$$(5.1) \qquad \frac{d^2x}{d\tau^2} - 6\frac{dx}{d\tau} - 12x^2 = 0, \qquad 0 < \tau < 1.$$

By the transformation

$$(5.2)\qquad t = 1 - 2\tau,$$

the equation (5.1) can be reduced to the following differential equation:

$$(5.3)\qquad \frac{d^2x}{dt^2} + 3\frac{dx}{dt} - 3x^2 = 0, \quad -1 < t < 1.$$

The constraining condition of least squares type is as follows:

$$(5.4)\qquad \sum_{j=0}^{10} \{x(t_j) - y_j\}^2 = \text{minimum},$$

where

$$(5.5)\qquad t_j = 0.2j - 1.0 \qquad (j = 0,1,\cdots,10)$$

and y_j $(j = 0,1,\cdots,10)$ are given in Table 1.

Table 1.

j	0	1	2	3	4	5
y_j	0.38727	0.39476	0.41305	0.43862	0.47017	0.50764
	6	7	8	9	10	
	0.55172	0.60372	0.66559	0.74012	0.83129	

This problem is equivalent to the example in [2].

We shall rewrite the problem into the vector form as (1.1) and (1.2). The equation (5.3) is transformed into

$$(5.6)\qquad \begin{cases} \dfrac{dx_1}{dt} = x_2, \\ \dfrac{dx_2}{dt} = 3x_1^2 - 3x_2, \end{cases} \qquad -1 < t < 1.$$

Let x be the vector $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, then (5.6) is equivalent to

$$(5.6)' \qquad \frac{dx}{dt} = \begin{bmatrix} x_2 \\ 3x_1^2 - 3x_2 \end{bmatrix}, \qquad -1 < t < 1$$

and the functional value J to be minimized is

$$(5.7) \qquad J = \frac{1}{2} \sum_{j=0}^{10} {}^t[L_j x(t_j) - d_j][L_j x(t_j) - d_j].$$

Here t_j $(j=0,1,\cdots,10)$ are the same as in (5.5), the matrices L_j $(j=0,1,\cdots,10)$ are

$$L_0 = L_1 = \cdots = L_{10} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

and the vectors d_j $(j=0,1,\cdots,10)$ are given by

$$d_j = \begin{bmatrix} y_j \\ 0.0 \end{bmatrix}.$$

The results of the numerical computation carried out by FACOM M-190 in the Data Processing Center, Kyoto University are shown in Tables 2∿4. For the numerical integration of ordinary differential equations the Runge-Kutta-Gill method programed by T. Ojika was used. All the calculations were carried out in the double precision arithmetic.

Table 2. Results obtained by the iteration.

iteration times	0		1		2		3		4	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
t_0=-1.0	0.5	0.0	0.37806769956	0.10806510612	0.38702086627	0.00261251861	0.38727177735	-0.00004288586	0.38727191330	-0.00004431630
t_1=-0.8	0.51252090985	0.11507113359	0.40171204036	0.12821189883	0.39490532424	0.07015010707	0.39476040385	0.06869739433	0.39476032659	0.06869661205
t_2=-0.6	0.54334736769	0.18909213769	0.42937707807	0.14869156970	0.41341909766	0.11229585764	0.41304637216	0.11139645273	0.41304617227	0.11139596872
t_3=-0.4	0.58721750119	0.24863472740	0.46135283076	0.17162147360	0.43913778607	0.14393243601	0.43861177501	0.14326111462	0.43861149266	0.14326075367
t_4=-0.2	0.64281390289	0.30821629103	0.49830337268	0.19871943583	0.47081948324	0.17287829974	0.47016701758	0.17226427328	0.47016666726	0.17226394342
t_5=0.0	0.71115159268	0.37748477983	0.54123626991	0.23176626442	0.50841638291	0.20369262002	0.50763820223	0.20303615170	0.50763778440	0.20303579926
t_6=0.2	0.79503126629	0.46516837677	0.59155143668	0.27296261250	0.55263785706	0.23962085391	0.55171808024	0.23885031164	0.55171758641	0.23884989812
t_7=0.4	0.89914650517	0.58192807415	0.65116185255	0.32531012072	0.60480645049	0.28369352839	0.60371555428	0.28274087195	0.60371496863	0.28274036083
t_8=0.6	1.0307690460	0.74350525298	0.72270288155	0.39313865121	0.66689448238	0.33949736231	0.66558820885	0.33828049978	0.66558750766	0.33827984706
t_9=0.8	1.2012163950	0.97571861039	0.80987497209	0.48294851916	0.74170948834	0.41193878825	0.74012428084	0.41034340656	0.74012343005	0.41034255099
t_{10}=1.0	1.4287209566	1.3241632086	0.91800946931	0.60486691312	0.83325492973	0.50828488609	0.83129911340	0.50613875295	0.83129806389	0.50613760232
J_k	4.93×10^{-1}		1.16×10^{-2}		5.81×10^{-6}		1.12×10^{-10}		1.10×10^{-10}	

convergence criterion 1.0×10^{-7}

step size for the Runge-Kutta method 0.003125

$\varepsilon = 1.0\times10^{-8}$

Table 3. The results for different starting values (I) ($\varepsilon=1.0\times10^{-8}$)

		(a)	(b)	(c)
Starting Values	$x_1(-1.0)$	0.5	0.4	0.1
	$x_2(-1.0)$	0.0	0.0	0.0
Iteration times		4	3	5
J_0		4.93×10^{-1}	4.60×10^{-3}	1.16
J_1		1.16×10^{-2}	1.54×10^{-6}	3.21×10^{-1}
J_2		5.81×10^{-6}	1.11×10^{-10}	3.45×10^{-3}
J_3		1.12×10^{-10}	1.10×10^{-10}	5.70×10^{-7}
J_4		1.10×10^{-10}		1.10×10^{-10}
J_5				1.10×10^{-10}

For the above three cases all the converged values at each t_j coincide to eleven figures. The values are shown in Table 2.

convergence criterion 1.0×10^{-7}

step-size for the Runge-Kutta method 0.003125

Table 4.

The results for different starting values (II) ($\varepsilon=1.0\times10^{-6}$)

		(a)	(b)	(c)
Starting values	$x_1(-1.0)$	0.5	0.4	0.1
	$x_2(-1.0)$	0.0	0.0	0.0
Iteration times		4	3	5
	J_0	4.93×10^{-1}	4.60×10^{-3}	1.15
	J_1	1.16×10^{-2}	1.54×10^{-6}	3.21×10^{-1}
	J_2	5.81×10^{-6}	1.11×10^{-10}	3.45×10^{-3}
	J_3	1.12×10^{-10}	1.10×10^{-10}	5.70×10^{-7}
	J_4	1.10×10^{-10}		1.10×10^{-10}
	J_5			1.10×10^{-10}

For the above three cases all the converged values at each t_j coincide to eleven figures. The values are following:

t_j	x_1	x_2
-1.0	0.38727191327	-0.00004431619
-0.8	0.39476032658	0.06869661211
-0.6	0.41304617226	0.11139596875
-0.4	0.43861149265	0.14326075369
-0.2	0.47016666726	0.17226394343
0.0	0.50763778440	0.20303579926
0.2	0.55171758641	0.23884989812
0.4	0.60371496863	0.28274036083
0.6	0.66558750766	0.33827984706
0.8	0.74012343005	0.41034255099
1.0	0.83129806389	0.50613760232

convergence criterion 1.0×10^{-7}

step-size for the Runge-Kutta method 0.003125

References

[1] Banks, H. T. & Groome, Jr., G. M., Convergence theorems for parameter estimation by quasilinearization, J. Math. Anal. Appl., 42(1973), 91-109.

[2] Fujii, M., Numerical solutions to problems of the least squares type for ordinary differential equations by the use of Chebyshev series, Bull. Fukuoka Univ. Education, 27, Prat III(1978), 15-26.

[3] Красносельский, М.А., Вайнинко, Г.М. и др., Приблженное решение операторных уравнений, "Наука", Москва, 1969.

[4] Lee, E. S., Quasilinearization and invariant imbedding, Academic P., New York, 1968.

[5] Mitsui, T., On the convergence of the initial-value adjusting method for nonlinear boundary value problems, submitted to Publ. RIMS, Kyoto Univ.

[6] Mitsui, T., The initial-value adjusting method for problems of the least squares type of ordinary differential equations, in preparation.

[7] Ojika, T. & Kasue, Y., Initial-value adjusting method for the solution of nonlinear multipoint boundary-value problems, to appear in J. Math. Anal. Appl.

[8] Urabe, M., The Newton method and its application to boundary value problems with nonlinear boundary conditions, U. S. -Japan Sem. Different. Funct. Eqs., Benjamin, New York, 1967, pp383-410.

[9] Urabe, M., On the Newton method to solve problems of the least squares type for ordinary differential equations, Mem. Fac. Sci., Kyushu Univ., Ser. A, 29(1975), 173-183.

A Direct Method for Solving Two-dimensional One Phase Stefan Problems

By

Vitoriano RUAS B. SANTOS*

Summary

A new algorithm is proposed for solving the Stefan problem for the case when the initial region occupied by the medium in a given phase is a starshaped two-dimensional domain.

The evolution of this region as time increases is determined by plotting the free boundary directly. At each time step this is approximated by a polygon in the interior of which the heat equation is solved with piecewise linear finite elements.

Numerical experiments indicate that the method is stable and that convergence is to be expected, under suitable assumptions on the regularity of both the boundary and initial data of the problem.

1. Introduction

We shall be concerned here with an algorithm for solving one-phase Stefan problems in two-dimension space. The basic tool for the algorithm is a method of automatic generation of triangular finite element meshes that the author initially proposed for solving boundary value problems defined on stationary star-

* Computer Science Department, Pontifícia Universidade Católica do Rio de Janeiro, Brazil
and Department of Mathematics, University of Tokyo.

shaped domains [8]. In that work many technical details of its implementation are given, while regularity properties of the so generated mesh have been analysed in [9].

This automatic triangulation process is used here both for plotting the free boundary and for adjusting the mesh, so that the heat equation can be conveniently solved just in the domain occupied by one of the phases of a certain medium undergoing a change of phase. This is what we call a direct method of solution of a Stefan problem.

Actually, our algorithm generalizes the one proposed by Mori [6] for the one-dimensional case. In his work the free boundary is determined after each increase in time of a fixed value Δt, while the mesh containing a fixed number of intervals moves according to the new shape of the domain. As it was well remarked by Prof. Fujita [3], our triangulation process could be the appropriate means of doing the same thing in the two-dimensional case, provided that the outer boundary of the initial domain is starshaped and that it remains starshaped as it evolves in time. Of course, since the domain is in itself one of the unknowns of the problem, the latter condition cannot be satisfied a priori. However, one might expect that, in most practical cases, the domain becomes more and more regular as time increases, in the sense that it actually remains starshaped, although rigorous proofs are not available yet.

Let us also say that Mori applied with success a slight generalization of his one phase algorithm to a two-dimensional strip domain [7]. Later Bonnerot and Jamet proposed the direct solution of one phase two-dimensional Stefan problems by time-space finite elements [5][1]. Though their method works well in many cases, as shown by their examples, they did not propose any solution to the problem of dealing with the spatial mesh as time increases. We believe this may cause either numerical inconvenience or difficulties of implementation for practical cases, unless a suitable solution of this problem is a part of the algorithm itself (see comments in Section 5).

Finally we note that a number of reasonable algorithms for indirect solution of the two-dimensional two phase problem are available at present. As a significant example of those we mention the work of Ciavaldini [1], that is based on a suitable transformation of the heat equation into a semilinear problem. This is solved on the fixed bounded domain occupied by the medium in both phases, and the position of the free boundary is then determined with the help of a function of the solution of the transformed equation. By introducing appropriate modifications, this approach can also be used for solving the pure one phase problem. However, in this case, although one can linearize the discrete problem, it is necessary to perform the calculations for a domain much larger than the one occupied by the only phase undergoing changes of temperature. This can be impractical in the case where the domain evolves to very large regions compared to the initial one.

2. The Continuous Problem

As we will comment in Section 5, with the eventual introduction of simple modifications, our algorithm can be applied to

[1] Actually, their algorithm gives approximate problems very close to Mori's in the cases he considers.

the solution of a very wide class of Stefan problems. However, we shall confine ourselves here to the solution of the special case of continuous problem defined on domains whose growth is not at all limited a priori, as described below. Our algorithm seems to attain the best of its efficiency compared to others particularly in the case where the domain may become much larger than the initial one.

In the sequel we introduce the formulation of the one phase Stefan problem we shall consider here, as a special case of the general multidimensional one treated by Friedman [2].

Let a certain medium exist in two different phases 1 and 2. Phase 1 occupies initially a certain bounded region Ω^0 of the x-y plane, whose boundary consists of two disjoint curves Γ^0 and Γ^*. Γ^0 is the interface separating both phases at time $t = 0$ and Γ^* is a fixed curve where heat sources are to be applied. We assume that Γ^* lies in the interior of Γ^0 and we denote its own interior by Ω^*.

We assume also that both $\Omega^0 \cup \bar{\Omega}^*$ and Ω^* are starshaped domains such that there exists one point $0 \in \Omega^*$ for which the segments joining it to any other point either of Ω^* or $\Omega^0 \cup \bar{\Omega}^*$ lie completely in those sets, respectively. An example of sets Ω^* and Ω^0 satisfying this assumption is illustrated in Figure 1.

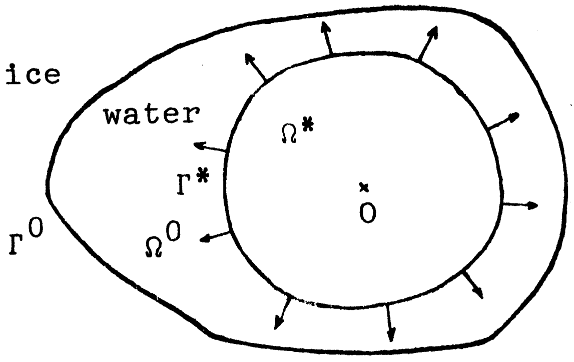

Figure 1

As time increases a change of phase 2 into phase 1 will occur in the medium lying in the region beyond Γ^0. So, the corresponding problem will be defined on the region $\mathbb{R}^2 - \Omega^*$.

As a practical example of this situation we have the problem of melting ice by heating some already melted region Ω^0 with a pipeline represented by Γ^* (Figure 1).

We shall denote by $\Omega(t)$ and $\Gamma(t)$ the domain occupied by the medium in phase 1 and its outer boundary respectively, at time t, $t \geq 0$. Clearly we have $\Omega(0) = \Omega^0$, $\Gamma(0) = \Gamma^0$ and that $\Gamma(t) \cup \Gamma^*$ is the boundary of $\Omega(t)$.

Let

$$(1) \qquad \phi(x, y, t) = 0$$

be the equation of the free boundary $\Gamma(t)$. We assume that ϕ is such that $\phi(x, y, t) < 0$ for $(x, y) \in \Omega(t) \cup \bar{\Omega}^*$ and $\phi(x, y, t) > 0$ for $(x, y) \notin \overline{\Omega(t) \cup \Omega^*}$. Now, supposing that the heating process takes place up to time T, for the temperature u of the medium in phase 1 we have the heat equation:

$$(2) \quad \frac{\partial u}{\partial t} = \Delta u \quad \text{in} \quad \Omega(t) \qquad \forall t \in [0, T]$$

with the initial condition:

$$(3) \quad u(x, y, 0) = u_0(x, y) \qquad \text{in} \quad \Omega^0$$

and the boundary conditions:

$$(4) \qquad u(\vec{x}, t) = g(\vec{x}, t)$$

or

$$(4)' \quad \frac{\partial u}{\partial \nu}(\vec{x}, t) = g'(\vec{x}, t)$$ (ν denotes the unit outer normal vector of Γ^*),

for $\vec{x} \in \Gamma^*$, $\forall t \in [0, T]$, and

$$(5) \qquad u(\vec{x}, t) = 0 \quad \text{for} \quad \vec{x} \in \Gamma(t), \quad \forall t \in [0, T].$$

Also for each time t, the so-called Stefan condition holds on $\Gamma(t)$

$$(6) \qquad (\nabla\phi, \nabla u)\Big|_{\vec{x}\in\Gamma(t)} = \kappa\frac{\partial\phi}{\partial t} .$$

where κ is a positive constant and $(\cdot, \cdot)$ denotes the scalar product in $\mathbb{R}^2$.

Thus the problem we want to solve is finding ϕ and u to satisfy (2) ~ (6).

As we have already mentioned, the algorithm we shall employ for solving this problem, is constructed upon a process of automatic triangulation. This in turn is based on the representation of the free boundary by an equation in polar coordinates (ρ, θ). So it will be useful to take ϕ to be an expression of the form:

$$(7) \qquad \phi = \rho - s(\theta, t).$$

In order to do so we must choose first of all a suitable origin of coordinates lying in Ω^*. According to the assumptions that we have made it is possible to choose such an origin so that both Γ^* and $\Gamma(t)$ can be represented in polar coordinates[2]. Thus it makes sense to write ϕ in form (7), and we have:

$$\nabla\phi = (\cos\theta + \frac{1}{s}\frac{\partial s}{\partial\theta}\sin\theta, \ \sin\theta - \frac{1}{s}\frac{\partial s}{\partial\theta}\cos\theta)$$

and

[2] Strictly speaking $\Omega^0 \cup \bar{\Omega}^*$ should be "non-singular" star-shaped. For the corresponding definition and other details see [9].

and

$$\frac{\partial \phi}{\partial t} = - \frac{\partial s}{\partial t}.$$

Thus (6) becomes:

$$(8) \qquad \left[\frac{\partial u}{\partial \rho} - \frac{1}{s}\frac{\partial s}{\partial \theta}\frac{\partial u}{\partial \sigma}\right]\Bigg|_{(\rho,\theta)\epsilon\Gamma(t)} = - \kappa \frac{\partial s}{\partial t}$$

where $\frac{\partial u}{\partial \rho}$ and $\frac{\partial u}{\partial \sigma}$ are, respectively, derivatives of u in two perpendicular directions (see Figure 2). The meaning of the first derivative is clear whereas the second one is simply $\frac{1}{\rho}\frac{\partial u}{\partial \theta}$.

On the other hand, according to (5), the tangential derivative $\frac{\partial u}{\partial \tau}$ on $\Gamma(t)$ must vanish.

So denoting by α the angle between the polar radius and by ν the outer normal on $\Gamma(t)$, we have:

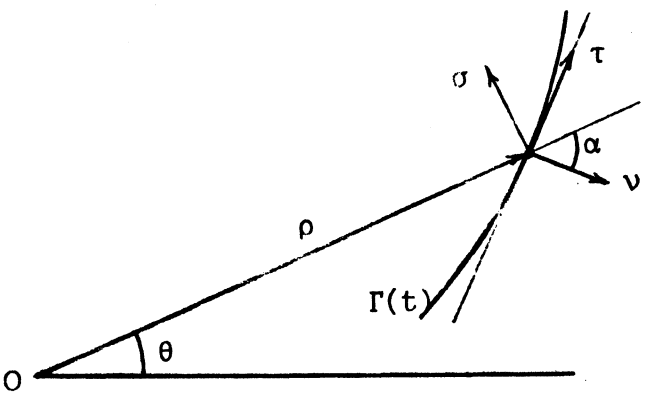

Figure 2

$$(9) \qquad \frac{\partial u}{\partial \rho}\sin\alpha + \frac{\partial u}{\partial \sigma}\cos\alpha = 0.$$

Using (8) and (9) and taking into account that

$$\tan\alpha = \frac{1}{s}\frac{\partial s}{\partial \theta},$$

we get:

$$(10) \qquad \left[1 + \left(\frac{1}{s}\frac{\partial s}{\partial \theta}\right)^2\right]\frac{\partial u}{\partial \rho}\bigg|_{\rho=s(\theta,t)} = - \kappa \frac{\partial s}{\partial t}.$$

(10) is the expression of the Stefan condition when u is given in polar coordinates.

3. The Algorithm

We shall now describe the algorithm that we propose for solving (2) ∼ (6) numerically.

First of all the discretization of (2) ∼ (5) is performed by standard methods applied to the corresponding variational formulation, namely:

$$(11) \qquad \int_{\Omega(t)} \frac{\partial u}{\partial t} v\, dx = - \int_{\Omega(t)} \nabla u \nabla v\, dx \quad \forall v \in V, \quad u \in \tilde{V}$$

if the Dirichlet-type boundary condition (4) holds, or

$$(11)' \quad \int_{\Omega(t)} \frac{\partial u}{\partial t} v\, dx = - \int_{\Omega(t)} \nabla u \nabla v\, dx + \int_{\Gamma^*} g' v\, ds \quad \forall v \in V', \ u \in V'$$

in case the Neumann-type boundary condition (4)' holds. Here V and V' are the subspaces of $H^1[\Omega(t)]$ of functions that vanish everywhere on $\Gamma(t) \cup \Gamma^*$ and on $\Gamma(t)$, respectively, and $\tilde{V}$ is the linear variety of V consisting of functions w such that $w = g$ on Γ^*.

For the discretization in space of (11) and (11)' we use a triangulation of $\Omega(t)$ on which we construct finite element spaces V_h or V_h' of piecewise linear continuous functions. Functions g and g' will be replaced by their piecewise linear interpolate g_h and g_h' which coincide with them at the nodes lying on Γ^*, respectively. Thus in case (4) holds, the approximate solution u_h will belong to $\tilde{V}_h$, discrete analogue of $\tilde{V}$, i.e., the linear variety of V_h such that $w_h \in \tilde{V}_h$

implies $w_h = g_h$ on the inner polygonal boundary approximating Γ^*. If (4)' holds, integration will be performed along this polygon instead of Γ^*.

For the discretization in time we employ standard schemes. This means that once obtained a linear system of ordinary differential equations after discretization in space:

$$(12) \qquad M_h(t)\frac{\partial u_h(t)}{\partial t} + A_h(t)u_h(t) = 0 \ ,$$

we start with $u_h^0 = u_{0_h}$, u_{0_h} being the V_h'-interpolate of u_0, and calculate $u_h^1, u_h^2, \ldots, u_h^n, \ldots$, approximations of $u_h(t)$ at increasing times values $t_1, t_2, \ldots, t_n, \ldots$ by

$$M_h^n(\tau_n)\frac{u_h^n - u_h^{n-1}}{t_n - t_{n-1}} + A_h^n(\tau_n)[wu_h^n + (1-w)u_h^{n-1}] = 0$$

where $\tau_n = wt_n + (1-w)t_{n-1}$, and $w \in [0,1]$ is the scheme parameter.

As a matter of fact, we take a constant increment of time Δt, so that $t_n = n\Delta t$, and approximate $\Omega(t_n)$, say by Ω_h^n.

Integrations are performed on a weighted domain given by $w\Omega_h^n + (1-w)\Omega_h^{n-1}$. Matrices M_h^n and A_h^n do not really depend on t, for the approximate domain Ω_h^n changes discretely. Thus the argument τ_n above only accounts for this interpolation.

Before describing how we determine Ω_h^n, we should give a short account of the triangulation method that we use.

Let $\rho = s^0(\theta)$ and $\rho = s^*(\theta)$ be the equations in polar coordinates of Γ^0 and Γ^*, respectively.

Now, given $\varepsilon > 0$ (possibly small), we choose integers m and p such that:

$$(13) \qquad \max_{\theta\in[0,2\pi]} \left|\frac{s^*(\theta)}{p} - \frac{s^0(\theta)-s^*(\theta)}{m}\right| \le \varepsilon$$

and we define

$$M = 8(p+m).$$

Let the boundary of Ω^0 be approximated by polygons Γ_h^0 and Γ_h^* whose vertices are the intersections of Γ^0 and Γ^* with the lines $\theta = \theta_j$, $j = 0,1,\ldots,M-1$, and $\theta = \theta_i^*$, $i = 0,1,\ldots,8p-1$, respectively, where $\theta_j = j\beta$ and $\theta_i^* = i\beta^*$, β and β^* being given by

$$(14) \qquad \beta = \pi/4(p+m)$$

and

$$(14)' \qquad \beta^* = \pi/4p\ .$$

Let Ω_h^0 be the domain bounded by Γ_h^0 and Γ_h^*, and $\rho = s_h^0(\theta)$ be the equation of Γ_h^0 in polar coordinates.

Now the vertices $P_{k\ell}$ of the triangulation of Ω_h^0 are defined as follows:

$$(15) \qquad \begin{cases} P_{k\ell} = (\rho_{k\ell}, \theta_{k\ell}) \\ \theta_{k\ell} = \dfrac{2\pi\ell}{k} \\ \rho_{k\ell} = (k-p)\dfrac{s_h^0(\theta_{k\ell}) - s^*(\theta_{k\ell})}{m} + s^*(\theta_{k\ell}) \\ k = p,\ p+1,\ \ldots,\ p+m; \quad \ell = 1,\ 2,\ \ldots,\ 8k\ . \end{cases}$$

An illustration of the so obtained triangulation is given in Figure 6 (see Section 4) for $p = 4$ and $m = 2$.[3] For other details see [9].

The meaning of ε is establishing a relation between p

[3] This triangulation is slightly different of the one we consider in [9]. The essential difference is due to the exclusion of the triangles lying in Ω^*.

and m so that the length of all the edges of the mesh remain of the same order as we refine the mesh. In the ideal case ε should be chosen to be at least $O(m^{-2})$. In this way we can immediately conclude that the spacial step size h is defined as $O(m^{-1})$ or equivalently as $O(p^{-1})$. But in most practical cases ε is just $O(h) = O(m^{-1})$ and still the above equivalence could be verified.

Notice that by this process, Γ_h^0 is a polygon of M vertices and that the domain is divided into 8 basic sectors of amplitude $\pi/4$, where the same kind of partition is performed. Also, in each sector we have $p+m$ boundary triangles, i.e., $p+m$ triangles adjacent to Γ_h^0 along an edge; between neighboring pairs of those boundary triangles we have $p+m-1$ interposed triangles, intersecting Γ_h^0 by one vertex.

Let us now introduce the process for determining the position of the free boundary Γ_h^n at the n-th time step, $n = 1, 2, \ldots,$ and consequently Ω_h^n, the interior of $\Gamma_h^n \cup \Gamma_h^*$.

First we denote by s_j^0 the polar radii of Γ_h^0 in the directions θ_j, $j = 0, 1, \ldots, M-1$.

Now we assume that during the fixed time increment Δt, Γ_h^{n-1} moves to Γ_h^n in the way shown in Figure 3, which represents a part of the domain.

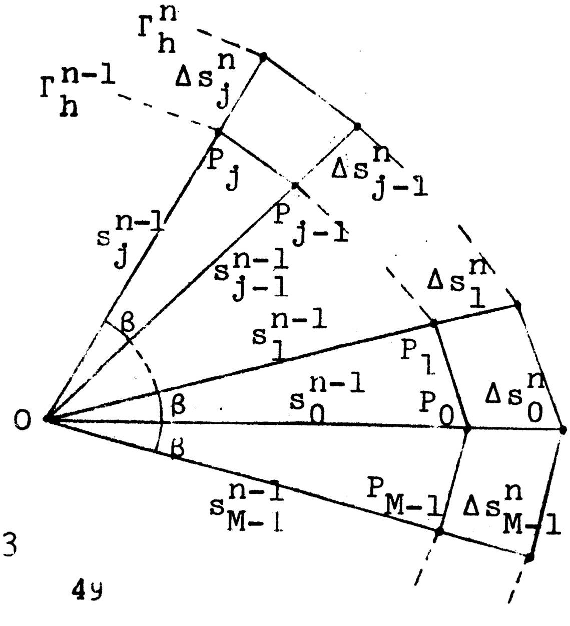

Figure 3

Δs_j^n being the displacement of the boundary in the direction θ_j, $j = 0,1,\ldots,M-1$, the polar radii s_j^n defining the new boundary Γ_h^n are given by

$$s_j^n = s_j^{n-1} + \Delta s_j^n , \quad n = 1,2,\ldots .$$

The increments Δs_j^n are calculated by discretizing the Stefan condition (10) as follows:

$$(16) \qquad -\kappa\frac{\Delta s_j^n}{\Delta t} = \frac{\partial u_h^{n-1}}{\partial\rho}(P_j)\left[1+\left(\frac{s_{j+1}^{n-1}-s_{j-1}^{n-1}}{2\beta s_j^{n-1}}\right)^2\right]$$

where $s_M^{n-1} = s_0^{n-1}$, $s_{-1}^{n-1} = s_{M-1}^{n-1}$, β is given by (14) and P_j are the vertices of Γ_h^n, as shown in Figure 3. $\frac{\partial u_h^{n-1}}{\partial\rho}(P_j)$ is calculated in two different ways according to the position of P_j:

1. If $j = k(p+m)$, $k = 0,1,\ldots,7$ then it is simply calculated along the edge $\overline{Q_jP_j}$ lying on the line $\theta = \theta_j$ (see Fig. 4a).

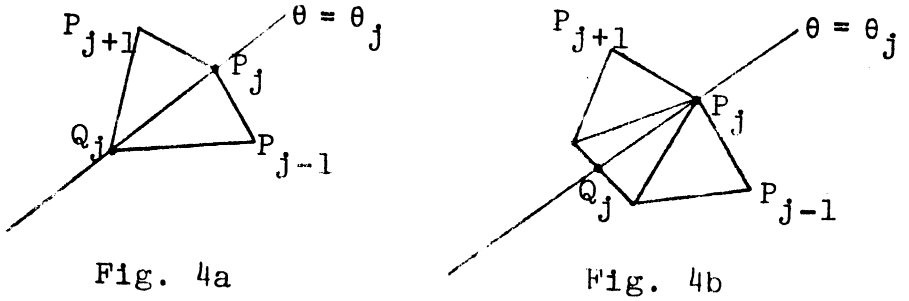

Fig. 4a

Fig. 4b

Figure 4

2. If $j \neq k(p+m)$, $k = 0,1,\ldots,7$, then $\frac{\partial u_h^{n-1}}{\partial\rho}(P_j)$ is calculated along the segment $\overline{Q_jP_j}$ (see Figure 4b) lying in the interposed triangle between the boundary triangles having P_j as a common vertex.

We note that due to the definition of our triangulation process, the latter operation always makes sense.

Finally, keeping the number of triangles constant, we simply define the new mesh by the automatic triangulation method (15) by replacing s_h^0 by s_h^n, for $n > 0$, $\rho = s_h^n(\theta)$ being the equation of the polygon Γ_h^n.

We should note that in our method each basis function φ_j depends on time, for the corresponding node P_i moves at each time step. Denoting by $N(h)$ the number of nodes and by $u_i(t)$ the coefficient of $u_h(t)$ associated with φ_i, that is, with node P_i, we have

$$\frac{\partial u_h}{\partial t} = \sum_{i=1}^{N(h)} \left(\frac{du_j(t)}{dt} \varphi_j + \frac{\partial \varphi_j}{\partial t} u_j \right).$$

Now recalling (12) we have:

$$A_h(t) = \{\int_{\Omega(t)} [\nabla\varphi_i \nabla\varphi_j + \varphi_i \frac{\partial \varphi_j}{\partial t}] dx\}$$

and

$$M_h(t) = \{\int_{\Omega(t)} \varphi_i \varphi_j dx\} .$$

The mass matrix $M_h(t)$ is obviously symmetric but $A_h(t)$ is non-symmetric due to the second term in the integrand. Actually the non-symmetric component of $A_h(t)$ is matrix $V_h(t)$,

$$V_h(t) = \{\int_{\Omega(t)} \varphi_i \frac{\partial \varphi_j}{\partial t} dx\}$$

that is called by Mori the velocity matrix, since it accounts for the effect of displacement of the nodes with respect to time.

Another point that is worth a remark is the following. Since for this algorithm we do not increase the number of triangles, and since the edges of Γ_h^* have fixed length at every time step, triangles closer to Γ_h^* will tend to elongate as n increases. Although from the accuracy viewpoint this may not be a desirable situation, as far as the rate of convergence is concerned, this violation of the classical angle condition for unbounded n does not yield any disadvantage. Indeed, as it has been proved by Jamet [4], the essential condition for maintaining the optimal rate of convergence of triangular finite element approximate solutions is not the lower boundedness of the angles, but the fact that no angle approaches 180° as the mesh is refined. Such a situation is not occurring here.

4. Numerical Results

We have tested the numerical viability of the algorithm proposed here which generally presented a good performance. Some singificant examples of the obtained results are given below. All the calculations were done in double precision on the HITAC/8800-8700 of the University of Tokyo Computer Center. For the solution of the systems of linear equations we used the Gaussian method for band matrices.

Example 1: For the given initial domain

$$\Omega_0 = \{(x, y)/1 < \rho < 2\}$$

we take as an exact solution the function

$$u = \frac{1}{2}(t + 2 - \frac{\rho^2}{t+2})$$

that satisfies a non-homogeneous equation of the form:

$$\frac{\partial u}{\partial t} - \Delta u = f .$$

The equation of the free boundary is $\rho = t + 2$. We preferred to choose (4)' as a boundary condition, so that we can compare computed values of u at the stationary nodal points of the grid, i.e., those lying on Γ^*.

We calculate up to time $T = 1$ so that the final domain is

$$\Omega(T) = \{(x, y) / 1 < \rho < 3\}.$$

Due to the symmetry of the problem we have performed the calculations only for the sector $0 \leq \theta \leq \pi/4$.

We define $h = 1/m$ and we take $p = m$ (so, in (13) $\varepsilon = 0$). We also take $w = 1$ as the scheme parameter so that we obtain a usual implicit scheme. In this way we choose $\Delta t = h/5$.

Table 1: Computed values of u for $\rho = 1$.

θ	h	t = 0.1	t = 0.5	t = 1.0
0	1/2	0.8194	1.0528	1.3175
π/8		0.8111	1.0424	1.3068
0	1/4	0.8154	1.0487	1.3206
π/8		0.8114	1.0442	1.3161
0	1/8	0.8132	1.0483	1.3255
π/8		0.8117	1.0467	1.3237
0	1/16	0.8123	1.0488	1.3289
π/8		0.8118	1.0483	1.3283
Exact Value		0.8119	1.0500	1.3333

Table 2: Maximal absolute errors of the computed solution.

h	t = 0.1	t = 0.5	t = 1.0
1/2	0.0147	0.0661	0.1310
1/4	0.0070	0.0319	0.0649
1/8	0.0033	0.0156	0.0322
1/16	0.0016	0.0077	0.0160

Remarks: 1. Those maximal absolute errors occur on the computed free boundary where the computed solution vanishes.

2. From Table 2 one can see that the observed rate of convergence in the maximum-norm is one for each t.

Table 3: Position of the free boundary for $\theta = 0$ and computer time.

h	t = 0.1	t = 0.5	t = 1.0	Comp. Time
1/2	2.0875	2.4423	2.8807	4.019sec.
1/4	2.0940	2.4713	2.9396	5.963sec.
1/8	2.0971	2.4857	2.9696	25.284sec.
1/16	2.0986	2.4929	2.9847	254.324sec.
Exact Value	2.1000	2.5000	3.0000	

Remarks: 1. For other values of θ the computed values are nearly the same which means that the free boundary is stably plotted.

2. One can observe linear convergence for the position of the free boundary.

Example 2: We take an example similar to the preceding one. Only this time Γ^* is reduced to a point, namely the centre of the circular domain:

$$\Omega^0 = \{(x, y) / 0 < \rho < 1\} .$$

The exact solution in this case is chosen to be:

$$u = \frac{1}{2}\left(t + 1 - \frac{\rho^2}{t+1}\right)$$

so that the equation of the free boundary is $\rho = t + 1$.

In this case it only makes sense choosing boundary condition (4).

Again we take $h = 1/m$, $w = 1$ and $h = \Delta t/5$. Obviously this time $p = 0$. We perform the calculations only in the sector $0 \leq \theta \leq \pi/4$, up to time $T = 1$, i.e., the final domain is given by

$$\Omega(T) = \{(x, y) / 0 < \rho < 2\}.$$

Table 4: Maximal absolute errors of the computed solution

h	t = 0.1	t = 0.5	t = 1.0
1/2	0.0299	0.1317	0.2516
1/4	0.0141	0.0643	0.1271
1/8	0.0067	0.0315	0.0635
1/16	0.0032	0.0156	0.0316

Remarks: 1. The basic convergence properties shown in the preceding example can be observed here also, although convergence itself seems to become slower.

2. For the position of the free boundary the preceding remark also applies.

Example 3. We take data symmetric with respect to the 4 directions $\theta = i\pi/4$, $i = 0, 1, 2, 3$ and we solve the following problem:

$$\begin{cases} \frac{\partial u}{\partial t} = \Delta u \quad \text{in} \quad \Omega(t) \times [0, 1], \quad \text{plus Stefan condition on} \quad \Gamma(t), \\ u(0, \vec{x}) = u_0 \quad \text{in} \quad \Omega^0, \quad u_0 = 2 - \rho - \frac{\cos 4\theta}{10} \\ \Gamma^* \quad \text{is given by} \quad s^*(\theta) = 1 \quad \forall \theta \\ \Gamma^0 \quad \text{is given by} \quad s^0(\theta) = 2 - \frac{\cos 4\theta}{10} \\ \frac{\partial u}{\partial \nu}(t, \vec{x}) = 1 \quad \text{for} \quad \vec{x} \in \Gamma^* \quad \text{and} \quad t \in [0, 1] \\ u(t, \vec{x}) = 0 \quad \text{for} \quad \vec{x} \in \Gamma(t) \quad \text{and} \quad t \in [0, 1]. \end{cases}$$

We define $h = 1/m$ and we take $p = m$ (so that $\varepsilon < 0.1m^{-1}$). We use again $w = 1$ but this time we take $\Delta t = h/2$.

For any quantity ϑ, we denote by ϑ_h its approximation for a certain value of h, and we define $\delta\vartheta_h$ by:

$$\delta\vartheta_h = 10^4 |\vartheta_h - \vartheta_{2h}|.$$

Table 5: Position of the free boundary for $\theta = 0$.

t →	0.25		0.50		0.75		1.00	
h	$s_h(0,t)$	δs_h	$s_h(0,t)$	δs_h	$s_h(0,t)$	δs_h	$s_h(0,t)$	δs_h
1/2	2.1500		2.3110		2.4407		2.5533	
1/4	2.1153	347	2.2585	525	2.3750	657	2.4768	765
1/8	2.0960	193	2.2303	282	2.3400	350	2.4366	402
1/16	2.0859	101	2.2155	148	2.3219	181	2.4158	208
1/32	2.0807	52	2.2080	75	2.3127	92	2.4053	105

Remark: Again the observed rate of convergence is one for the position of the free boundary.

Table 6: Computed values of u for $\rho = 1$ and $\theta = 0$.

t →	0.25		0.50		0.75		1.00	
h	$u_h(1,0,t)$	δu_h	$u_h(1,0,t)$	δu_h	$u_h(1,0,t)$	δu_h	$u_h(1,0,t)$	δu_h
1/2	0.8745		0.8545		0.8522		0.8605	
1/4	0.8650	95	0.8424	121	0.8426	96	0.8545	60
1/8	0.8541	109	0.8279	145	0.8291	135	0.8427	118
1/16	0.8469	72	0.8183	56	0.8201	90	0.8346	81
1/32	0.8428	41	0.8130	53	0.8151	50	0.8301	45

Remarks: 1. It seems that the optimal rate $\Delta t/h$ for the purpose of economy is far less than 1/2. Indeed we observed that for the same value of h the

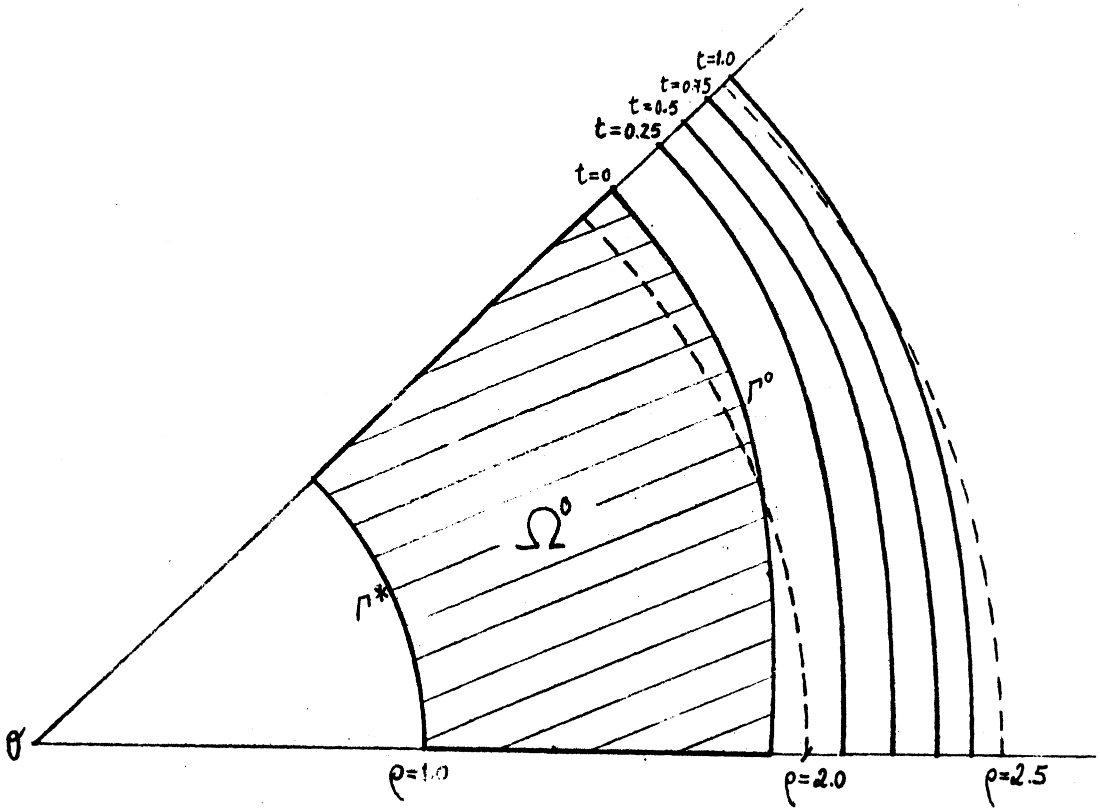

Figure 5

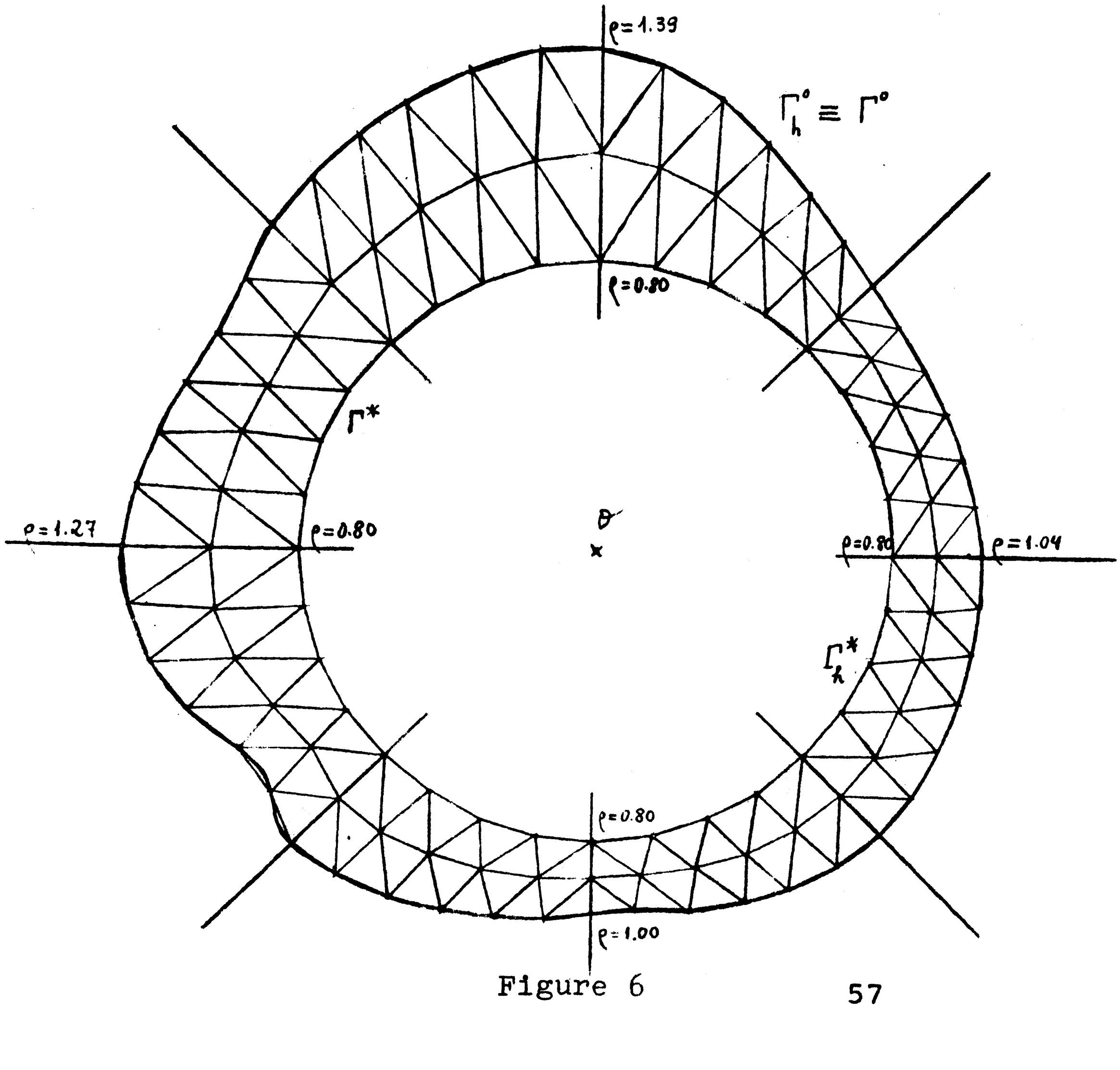

Figure 6

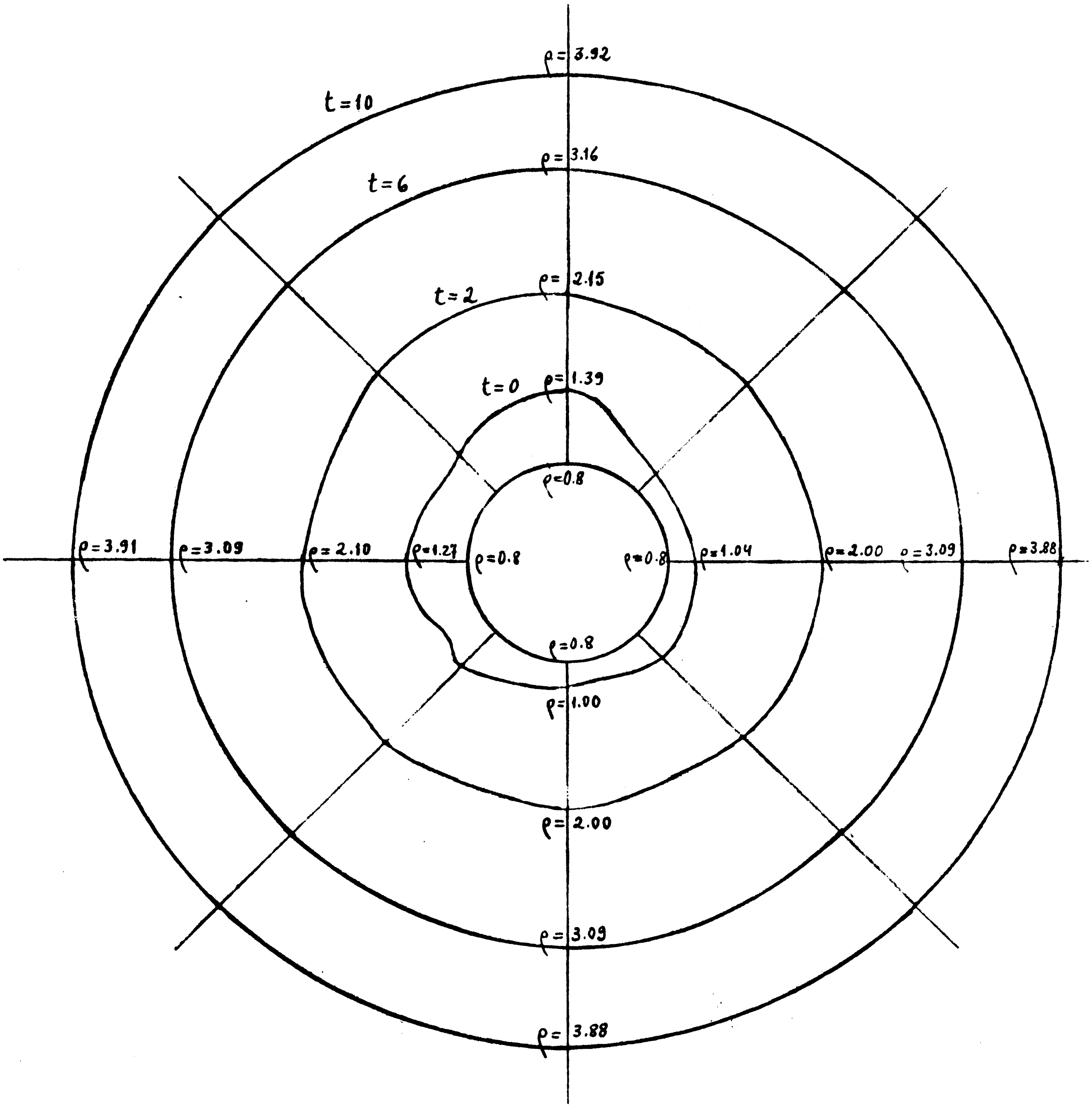
ρ= 3.92
t=10
ρ= 3.16
t=6
ρ= 2.15
t=2
t=0
ρ= 1.39
ρ=0.8
ρ=3.91
ρ=3.09
ρ= 2.10
ρ=1.27
ρ=0.8
ρ=0.8
ρ=1.04
ρ=2.00
ρ=3.09
ρ=3.88
ρ=0.8
ρ= 1.00
ρ= 2.00
ρ= 3.09
ρ= 3.88

Figure 7

approximate values obtained with $\Delta t = h/5$ are much closer to the estimated solution (by extrapolation to the limit) than those in Table 5. So, in practice Δt should be small even for $w = 1$.

2. The number of operations necessary for solving the system of linear equations is $O(m^4)$ whereas the composition of matrices at each time step is only $O(m^2)$. So we suggest also the following procedure:

   Choose $w = 0$, i.e., an explicit scheme, with lumped mass system [6]. In this case there is no need to solve a system of linear equations at each time step. Although for explicit schemes we must take $\Delta t = O(h^2)$ this could be advantageous in some cases (for instance, when m is sufficiently large). This statement is based on the fact that the number of operations with this modified scheme becomes $O(m^4)$ instead of $O(m^5)$, and is also supported by the preceding remark.

In Figure 5 we show the evolution of the free boundary for increasing values of t and $h = 1/32$. The dotted lines show circles with center 0.

Example 4: We have tested our algorithm to a whole domain by solving a non-symmetric problem similar to the one of the preceding example. We have observed basically the same properties, although in this case we could not use too small mesh sizes because the computer time increases sharply. So we prefer showing the results obtained for the following problem in which the domain evolves up to large values of T:

- Γ^0 is given by a set of 96 points equally spaced in θ and with polar radii ranging from 1.0 to 1.4, as shown in Figure 6. So we consider Γ^0 to be a polygon.
- Γ^* is defined by $\rho = 0.8$.
- u_0 is the function such that $\frac{\partial u_0}{\partial \rho} = 1 \quad \forall \vec{x} \in \Omega^0, \quad u_0|_{\Gamma^0} = 0$.
- $\frac{\partial u}{\partial \nu}(\vec{x}, t) = 1 \quad \forall t \in [0, T]$ and $\vec{x} \in \Gamma^*$.

The aspect of the initial triangulation is shown in Figure 6 for $p = 4$, $m = 2$. We actually calculate with $p = 8$ and $m = 4$ (so that $\varepsilon \leq 0.05$). Defining $h = 0.8/p$ we choose $\Delta t = h$, i.e., $\Delta t = 0.1$. We calculate up to time $T = 10.0$, and we show in Figure 7 the evolution of the free boundary as time increases. In this way the computation lasts about 3 minutes.

5. Concluding Remarks

Although we did not consider explicitly other cases, it should be clear that every one-phase Stefan problem in which the domain evolves as a starshaped one can be treated in a similar way to that described in Sections 2 and 3. In particular we mention the case of bounded domains such as the region Ω shown in Figure 8. By introducing appropriate modifications, we can take into account the gradual transformation of edges of the approximate free boundary into edges of the fixed external boundary $\tilde{\Gamma}$, and then solve the problem similarly (see [5]).

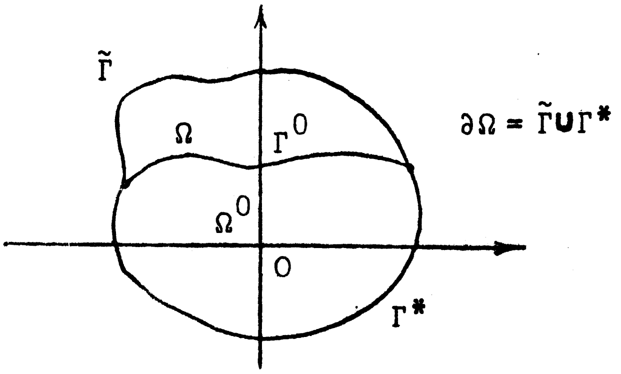

Figure 8

Unfortunately generalization of our algorithm to some cases such as that of the two phase problem, seems not to be so straightforward. However, as we have mentioned other algorithms based on the direct determination of the free boundary could be efficiently used in such cases, provided that good methods for adjusting the spacial mesh step by step are available. Indeed, one of the main features of our triangulation method is generating triangles whose angles remain reasonably bounded away from zero, or, in other words, approximately equal. In so doing the number of nodes of the mesh necessary to attain a given precision can be minimized and computer time saved.

So, generally speaking, we think that our work could be a starting point for research on the application of automatic discretization processes to the direct solution of two or three-dimensional free boundary problems. This is because we also believe that with this work we have helped to contradict the long-prevailing opinion, that this approach is inadequate to the numerical solution of such problems.

Acknowledgements

The author wishes to express his gratitude to Prof. H. Fujita of the Department of Mathematics of the University of Tokyo for extending a warm welcome to him and for giving valuable suggestions which were essential for the development of these results.

Thanks should be extended to Prof. T. Ushijima of the Department of Information Mathematics of the University of Electro-Communications, for his kind assistance in various academic matters during the author's study in Japan.

This work was supported by the Japanese Ministry of Education (Mombusho).

References

[1] Ciavaldini, J. F., Analyse Numérique d'un Problème de Stefan à Deux Phases par une Méthode d'Eléments Finis, SIAM Journal on Numerical Analysis, Vol.12, 3 (1975), 464-487.

[2] Friedman, A., The Stefan Problem in Several Space Variables, Trans. Amer. Math. Soc., 132 (1968), 51-87.

[3] Fujita, H., Personal Communication, Department of Mathematics, University of Tokyo.

[4] Jamet, P., Estimations d'Erreur pour des Eléments Finis Droits Presque Dégénérés, R.A.I.R.O., Vol.10, 3 (1976), 43-61.

[5] Jamet, P. and Bonnerot, R., Numerical Computation of the Free Boundary for the Two-dimensional Stefan Problem by Time-Space Finite Elements, J. of Comp. Phys., Vol.25, 2 (1977), 163-181.

[6] Mori, M., Stability and Convergence of a Finite Element Method for Solving the Stefan Problem, Publ. RIMS, Kyoto Univ., 12 (1976), 539-563.

[7] Mori, M., A Finite Element Method for Solving Moving Boundary Problems, Preprints of IFIP Working Conference on Modelling of Environmental Systems, Tokyo, April 1976, 167-171.

[8] Ruas B. Santos, V., Sur l'Application de Quelques Eléments Finis non Conformes à la Résolution de Problèmes Biharmoniques, Thèse, Université Paris VI, April 1976.

[9] Ruas B. Santos, V., Automatic Generation of Triangular Finite Element Meshes (to appear in Computer and Mathematics with Applications).

Error Estimates for the Lumped Mass Approximation of the Heat Equation

By Teruo USHIJIMA

January 1979

Introduction.

In this paper error estimates for the lumped mass approximation of the inhomogeneous heat equation with zero Dirichlet boundary condition are considered. Generally the error is dominated by the term caused by the lumping effect. It is, however, bounded by the term having the same order as in the case of consistent mass approximation if the triangulation poseses the local symmetric property, which is a kind of regularity of triangulation, being defined in §1 of this paper. Here we restrict our consideration to a semi-discrete approximation scheme. Namely we adopt a system of ordinary differential equations as an approximate equation.

It may be well-known that the error is $O(h)$ in the lumped mass approximation for the heat equation. And classical analysis teaches us that the error is $O(h^2)$ in the usual difference scheme using five point differnce approximation of Δ, whereas this difference scheme can be regarded as a special case of the lumped mass approximation scheme, as was shown by Courant [4]. Original motivation of this work is to give a persuasible explanation of this discrepancy.

An idea of division of error term into two terms is essential. The first term represents the error of Ritz projection, which already appeared in the consistent mass approximation. The second term is proper to the lumped mass approximation, which is dominant, in general, being O(h) with respect to mesh size h. In the case of local symmetric triangulation, the estimate for the second term is improved up to $O(h^2)$, so that the first term is principal.

In §1, after giving formulation of the problem, we state main result concerning L^2-estimate in Theorem 2. An abstract routine to treat the error estimation is constructed in §2. §3 is devoted to prove Theorem 2. In §4, L^∞-estimates are driven as applications of our abstract routine.

As for L^∞-estimates, similar results to us have been obtained by Tabata [8] (see also [9]). He regards the lumped mass scheme as a finite difference scheme defined on an irregular mesh. In comparison with his method, this paper may be considered to describe an operator theoretical approach of the lumping method.

The author expresses his sincere thanks to Professor Kikuchi of Institute of Space and Aeronautical Science, University of Tokyo for his valuable advices and discussions during the preparation of this paper.

§1. Statement of the problem and L^2-estimates.

We consider the following continuous problem $(\mathcal{E})$.

$$(\mathcal{E})\begin{cases}\dfrac{\partial u(t,x)}{\partial t}-\Delta u(t,x) = f(t,x), & (t,x)\in(0,T]\times\Omega,\\ u(t,x) = 0, \quad (t,x)\in(0,T]\times\Gamma, & \\ u(0,x) = a(x), \quad x\in\Omega, & \end{cases}$$

where Ω is a bounded convex polygonal domain in n-dimensional Euclidean space $\mathbb{R}^n$ with its boundary Γ, and $f(t,x)$ and $a(x)$ are given known functions.

Let us denote the inner product of $L^2(\Omega)$ by $(\ ,\ )$. Namely

$$(u,v) = \int_\Omega u(x)v(x)dx \quad \text{for} \quad u,v\in L^2(\Omega).$$

Define the inner product $a(\ ,\ )$ of $V = H_0^1(\Omega)$ by the formula:

$$a(u,v) = \sum_{j=1}^{n} \left(\frac{\partial u}{\partial x_j},\frac{\partial v}{\partial x_j}\right) \quad \text{for } u,v\in V.$$

Then there is a unique selfadjoint operator A in $L^2(\Omega)$ with its domain $D(A)$ satisfying the following properties from (1.1) to (1.4):

(1.1) $D(A) = \{v\in V: \Delta v\in L^2(\Omega)\}$,

(1.2) $Av = -\Delta v$ for $v\in D(A)$,

(1.3) $D(A^{1/2}) = V$,

(1.4) $a(u,v) = (Au,v)$, for $u\in D(A)$, $v\in V$.

As is well known,

$$D(A) = H_0^1(\Omega)\cap H^2(\Omega)$$

holds since Ω is a convex polygonal domain (see Kadlec [6]).

As usual, we seek the solution $u(t,x)$ of $(\mathcal{E})$ as the solution of the following variational formulation (Π) of $(\mathcal{E})$, or operational formulation (E) of $(\mathcal{E})$.

$$(\Pi)\begin{cases}\frac{d}{dt}(u(t),v)+a(u(t),v) = (f(t),v), & 0<t\leqq T,\ v\in V,\\ u(t)\in V, & 0<t\leqq T,\\ u(0) = a.\end{cases}$$

$$(E)\begin{cases}\frac{du(t)}{dt} + Au(t) = f(t), & t>0,\\ u(0) = a.\end{cases}$$

Following Fujii [5], let us formulate the lumped mass approximation method. Let positive numbers h be indices. Assume that there is a triangulation $\mathcal{T}_h = \{T: \text{simplex},\ \text{diam}(T) \leqq h\}$ of Ω, and that the family of triangulations $\{\mathcal{T}_h : h>0\}$ is regular in the sense of Ciarlet-Raviart [2]. Namely there is a positive constant σ satisfying that

$$h(T)/\rho(T) \leqq \sigma \quad \text{for } T\in\mathcal{T}_h,\ h>0,$$

where $h(T)$ is a diameter of T and $\rho(T)$ is a diameter of the inscribed sphere of T.

Now we set our approximation space V_h of V as follows.

$$V_h = \{v_h \in C(\bar{\Omega}_h) : v_h|_\Gamma = 0,$$
$$v_h|_T \in P_1 \quad \text{for any } T\in\mathcal{T}_h\},$$

where $C(\bar{\Omega})$ denotes the space of continuous functions on $\bar{\Omega}$, and $v|_S$ denotes the restriction of the function v to the set S, and P_k denotes the totality of polynomials with degree at most k. A vertex of $T\in\mathcal{T}_h$ is said to be a nodal point. Let us count the interior and the boundary nodal points of Ω as $b_1, b_2, \cdots, b_N$, and $b_{N+1}, b_{N+2}, \cdots, b_{N+M}$, respectively. There exists uniquely a set of functions $\{w_j : 1\leqq j\leqq N+M\}$ with the properties:

$$w_j \in C(\bar{\Omega}_h),$$
$$w_j|_T \in P_1 \quad \text{for any } T\in\mathcal{T}_h,$$
$$w_j(b_k) = \delta_{jk}.$$

Clearly $\{w_j : 1\leqq j\leqq N\}$ forms a basis of V_h. Let S_j be the support of w_j, and let B_j be the lumped mass region corresponding to the nodal point b_j. (For the definition of lumped mass region, see, for example, p479 of Ushijima [10].) The characteristic function of B_j is denoted by $\bar{w}_j$. Let $\bar{V}_h$ be the linear space spanned by the set of functions $\{\bar{w}_j : 1\leqq j\leqq N\}$. The lumping operator L_h from V_h onto $\bar{V}_h$ is defined as the linear mapping naturally generated by the correspondence: $w_j \longrightarrow \bar{w}_j$. Namely we have

$$L_h v_h = \sum_{j=1}^{N} V_j \bar{w}_j \quad \text{for } v_h = \sum_{j=1}^{N} V_j w_j \epsilon V_h.$$

Let K_h be the inverse of L_h. The spaces V_h and $\bar{V}_h$ being closed subspaces of $L^2(\Omega)$, we can define adjoint operators L_h^*, and K_h^*, from $\bar{V}_h$ onto V_h, and from V_h onto $\bar{V}_h$, respectively, inner products of these spaces being induced from $L^2(\Omega)$. It holds that

$$(L_h^* \bar{v}_h, u_h) = (\bar{v}_h, L_h u_h), \quad \bar{v}_h \epsilon \bar{V}_h,\ u_h \epsilon V_h,$$

and that

$$(K_h^* v_h, \bar{u}_h) = (v_h, K_h \bar{u}_h) \quad v_h \epsilon V_h,\ \bar{u}_h \epsilon \bar{V}_h.$$

Let A_h be a linear operator acting on V_h defined by the formula:

$$(A_h u_h, v_h) = a(u_h, v_h), \quad u_h, v_h \epsilon V_h.$$

We say that A_h is the Galerkin approximation of A. Clearly A_h is a bounded positive self adjoint operator in the space V_h. Define a bounded linear operator $\tilde{A}_h$ acting on V_h by the formula:

$$\tilde{A}_h = K_h K_h^* A_h.$$

Let us denote by Π_h the linear interpolation operator from $C(\bar{\Omega})$ onto V_h. Let P_{1h} be the orthogonal projection from V onto V_h with respect to the inner product a(,), which is called Ritz projection, sometimes, and let P_{0h}, and $\bar{P}_h$, be the orthogonal projection from $L^2(\Omega)$ onto V_h, and $\bar{V}_h$, respectively, with respect to the inner product (,).

Our approximate problem $(\bar{\Pi}_h)$ can be written as follows.

$$(\bar{\Pi}_h) \begin{cases} \frac{d}{dt}(L_h u_h(t), L_h v_h) + a(u_h(t), v_h) = (L_h f_h(t), L_h v_h), \\ \qquad\qquad\qquad\qquad 0<t\leqq T,\ v_h \epsilon V_h, \\ u_h(t) \in V_h, \quad 0<t\leqq T, \\ u_h(0) = a_h, \end{cases}$$

where $f_h(t)$ is a V_h-valued function, and a_h is an element of V_h. This problem $(\bar{\Pi}_h)$ is equivalent to the following V_h-valued evolution equation $(\tilde{E}_h)$

$$(\tilde{E}_h) \begin{cases} \frac{du_h(t)}{dt} + \tilde{A}_h u_h(t) = f_h(t), & t>0, \\ u_h(0) = a_h. \end{cases}$$

Definition 1. A triangulation $\mathcal{T}_h$ is said to be locally symmetric if S_j is symmetric with respect to b_j for any interior nodal point b_j $(1 \leq j \leq N)$, namely, if it holds that

$-(S_j - b_j) = S_j - b_j, \quad 1 \leq j \leq N.$

It is noted that $w_j(x+b_j)$ is an even function of x if the triangulation is locally symmetric.

Two typical examples of locally symmetric triangulation in 2-dimensional case are illustraled in Fig. 1 and 2.

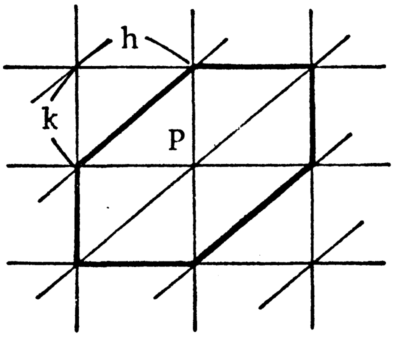

Fig. 1

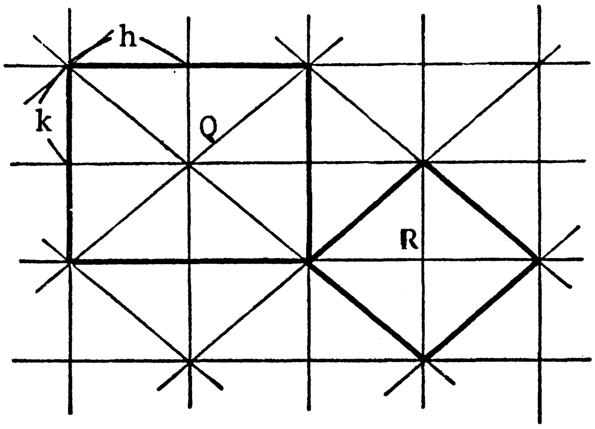

Fig. 2

In both examples, we obtain the stiffness matrix corresponding to usual five points difference formula for $-\Delta$. As for mass matrices, the matrix element of the point P is hk, where as that of Q, and R, are 4/3hk, and 2/3hk, respectively.

Theorem 2. Let Ω be a bounded convex polygonal domain in $\mathbf{R}^n$ where $n \leq 3$. Assume that $\{\mathcal{T}_h : h>0\}$ is regular. Then there is a constant C, which may depend on T, with the property that the estimate

$$\max_{0\leqq t\leqq T} \| u_h(t)-u(t) \|_{L^2(\Omega)}$$
$$\leqq C\{h^m(\max_{0\leqq t\leqq T} \| u(t) \|_{H^m(\Omega)}+\max_{0\leqq t\leqq T} \| \frac{\partial u}{\partial t} \|_{H^m(\Omega)}+\max_{0\leqq t\leqq T} \| f(t) \|_{H^m(\Omega)})$$
$$+\| a_h-P_{1h}a \|_{L^2(\Omega)}+\max_{0\leqq t\leqq T} \| f_h(t)-f(t) \|_{L^2(\Omega)}\}$$

holds for the solution u(t) of (E) and the solution $u_h(t)$ of (E_h) provided that $u(t)\epsilon D(A^{1+m/2})$, $\frac{\partial u}{\partial u}(t)\epsilon D(A^{m/2})$, and that f(t) is so smooth that the quantity of the right hand side may be meaningful. Here we can put m=2 if $\mathcal{T}_h$ is locally symmetric for any h, whereas m=1 in general case.

§2. An abstract theory for error estimation.

Now we consider the following five conditions.

Condition I. There is a Banach space X satisfying that

$$L^2(\Omega)\supset X\supset V_h.$$

Condition II. There is a Banach space Y contained in X and V, and a scalar function $\varepsilon(h)$ such that

$$\| P_{1h}v-v \|_X \leqq \varepsilon(h) \| v \|_Y, \quad v\epsilon Y.$$

Condition III. There is a Banach space Z contained in X, and a scalar function $\delta(h)$ such that

$$\| K_hK_h{}^*P_{0h}v-v \|_X \leqq \delta(h) \| v \|_Z, \quad v\epsilon Z.$$

Condition IV. There is a constant M_0 satisfying

$$\| e^{-t\tilde{A}_h}v_h \|_X \leqq M_0 \| V_h \|_X, \quad 0\leqq t\leqq T, \ v_h\epsilon V_h.$$

Condition V. The solution u(t) of (E) satisfies

$$u(t)\ \epsilon\ Y\cap D(A),\ Au(t)\ \epsilon\ Z, \text{ and } \frac{\partial u(t)}{\partial t}\ \epsilon\ Y$$

for any $t\epsilon[0,T]$.

Theorem 2. Under conditions from I to V, we have the following estimate with a suitable constant C.

$$\max_{0\leqq t\leqq T}\| u_h(t)-u(t) \|_X$$
$$\leqq C\{\varepsilon(h)(\max_{0\leqq t\leqq T}\| u(t) \|_Y+\max_{0\leqq t\leqq T}\| \frac{\partial u}{\partial t}\|_Y)$$
$$+\delta(h)\max_{0\leqq t\leqq T}\| Au(t) \|_Z+\| a_h-P_{1h}a\|_X+\max_{0\leqq t\leqq T}\| f_h-f\|_X\}.$$

Proof. Let $r_h(t)=P_{1h}u(t)$. We have

(2.1) $\tilde{A}_h r_h = K_hK_h{}^*P_{0h}Au,$

since $r_h=A_h{}^{-1}P_{0h}Au$. Let $g_h=\frac{d}{dt}(u_h-r_h)+\tilde{A}_h(u_h-r_h)$, then we have

(2.2) $g_h = (f_h-f)-(P_{1h}\frac{\partial u}{\partial t} - \frac{\partial u}{\partial t}) - (K_hK_h{}^*P_{0h}Au-Au)$.

In fact, by $(\tilde{E}_h)$ and (2.1),

$$g_h = f_h-\frac{d}{dt}r_h-\tilde{A}_hr_h$$
$$= f_h-P_{1h}\frac{\partial u}{\partial t} - K_hK_h{}^*P_{0h}Au$$
$$= f_h-P_{1h}\frac{\partial u}{\partial t} + \frac{\partial u}{\partial t} + Au - f - K_hK_h{}^*P_{0h}Au.$$

Therefore the function $e_h=u_h-r_h$ is the solution of the following evolution problem (2.3).

(2.3) $$\begin{cases} \frac{d}{dt}e_h + \tilde{A}_he_h = g_h & 0\leqq t\leqq T, \\ e_h(0) = a_h - P_{1h}a. \end{cases}$$

By Duhamel's principle we have

(2.4) $e_h(t) = e^{-t\tilde{A}_h}e_h(0) + \int_0^t e^{-(t-s)\tilde{A}_h}g_h(s)\, ds.$

Substituting condition II , III, V into (2.2), we obtain

$$\| g_h(t) \|_X \leqq \| f_h-f\|_X+\varepsilon(h)\| \frac{\partial u}{\partial t}\|_Y+\delta(h)\| Au\|_Z.$$

Therefore (2.4) and Condition IV imply the following estimate

$$\| e_h(t) \|_X \leqq M_0\| a_h-P_{1h}a\|_X$$
$$+M_0T(\| f_h-f\|_X+\varepsilon(h)\| \frac{\partial u}{\partial t}\|_Y+\delta(h)\| Au\|_Z).$$

On the other hand, Condition II implies

$$\| r_h(t)-u(t) \|_X = \| P_{1h}u(t)-u(t) \|_X$$
$$\leqq \varepsilon(h) \| u(t) \|_Y.$$

Thus the conclusion of Theorem 3 follows from the triangular inequality:

$$\| u_h(t)-u(t) \|_X \leqq \| u_h(t)-r_h(t) \|_X + \| r_h(t)-u(t) \|_X.$$

Corollary 4. Adding conditions from I to V, assume further that Y is a closed subspace of Z. Then we have the following estimate with $\tilde{\varepsilon}=\max(\varepsilon,\delta)$.

$$\max_{0\leqq t\leqq T} \| u_h(t)-u(t) \|_X$$
$$\leqq C\{\tilde{\varepsilon}(h)(\max_{0\leqq t\leqq T} \| u(t) \|_Z + \max_{0\leqq t\leqq T} \| \frac{\partial u}{\partial t} \|_Z + \max_{0\leqq t\leqq T} \| f \|_Z)$$
$$+ \| a_h - P_{1h}a \|_X + \max_{0\leqq t\leqq T} \| f_h - f \|_X\}.$$

Proof. Substitute $Au=f-\frac{\partial u}{\partial t}$ into the estimate in Theorem 3.

§3 Proof of Theorem 2.

First we prepare some Propositions.

Proposition 5. For any $p\in[1,\infty]$, operators $\overline{P}_h$, $K_h\overline{P}_h$ and $K_h{}^*P_{0h}$ can be considered as operators from $L^p(\Omega)$ to $L^p(\Omega)$ with their norms not greater than 1.

Proof. See Proposition 2.1 of Ushijima [10].

Proposition 6. For any $p\in[1,\infty]$, there is a constant $L=L_{p,n}$ independent of h such that

$$\| L_h v_h \|_{L^p(\Omega)} \leqq L \| v_h \|_{L^p(\Omega)}, \quad v_h \in V_h.$$

Proof. See Proposition 2.2 of Ushijima [10].

Proposition 7. For $a\in C(\overline{\Omega})$, we have

$$K_h K_h{}^* P_{0h} a - \mathbb{L}_h a$$

$$= \sum_{j=1}^{N}\{\frac{1}{m(B_j)}\int_{S_j}(a(x)-a(b_j))w_j(x)dxw_j\},$$

where $m(B_j)=\int_{B_j}dx$.

Proof. Noticing the orthogonality relation:

$$(\overline{w}_j,\overline{w}_k) = \delta_{jk}m(B_j),$$

we have

$$K_h{}^*P_{0h}a = \sum_{j=1}^{N}\frac{1}{m(B_j)}\int_{S_j}a(x)w_j(x)dx\overline{w}_j.$$

On the otherhand,

$$\Pi_h a = \sum_{j=1}^{N}a(b_j)w_j$$

$$= \sum_{j=1}^{N}\frac{1}{m(B_j)}\int_{S_j}a(b_j)w_jdxw_j,$$

where we use the equality:

$$m(B_j) = \int_{S_j}w_j(x)dx.$$

Therefore we have the conclusion.

The proof of Theorem 2 for m=1 is summarized in the following Proposition.

Proposition 8. Let X be $L^2(\Omega)$, and let Y and Z be V. Set $M_0=L_{2,n}$. Then the conditions from I to IV hold with suitable scalar functions $\varepsilon(h)$ and $\delta(h)$ which behave O(h) as h tends to 0.

Proof. Condition I is trivially valid. By a standard error estimation method for elliptic problem, we have

(3.1) $\quad \| P_{1h}v-v\|_{L^2(\Omega)} \leqq Ch\| v\|_{H^1(\Omega)}, \quad v\in V,$

which assures condition II (see Ciarlet [1]). Set $v_h=P_{1h}v$ for $v\in V$. By definition of P_{1h},

(3.2) $\quad \| v_h\|_V \leqq \| v\|_V.$

Since $\| v\|_V=\| \nabla v\|_{L^2(\Omega)}$ is equivalent to $\| v\|_{H^1(\Omega)}$, (3.1) implies

(3.3) $\quad \| v_h-v\|_{L^2(\Omega)} \leqq Ch\| v\|_V.$

Noticing $\Pi_h v_h = v_h$, we have

$$K_h{}^*P_{0h}v_h - L_h v_h$$

$$= \text{(by Proposition 7)}$$

$$= \sum_{j=1}^{N} \{ \frac{1}{m(B_j)} \int_{S_j} (v_h(x) - v_h(b_j)) w_j(x) dx \bar{w}_j \}$$

$$= \sum_{j=1}^{N} \{ \frac{1}{m(B_j)} \int_{S_j} (x - b_j, \nabla v_h(x))_n w_j(x) dx \bar{w}_j \},$$

for $v_h(x)$ is piece wise linear on S_j, where $(c,b)_n$ means the Euchidean inner product of $a,b \in \mathbb{R}^n$. Therefore

$$\| K_h{}^*P_{0h}v_h - L_h v_h \|^2$$

$$= \sum_{j=1}^{N} \frac{1}{m(B_j)} | \int_{S_j} (x - b_j, \nabla v_h(x))_n w_j(x) dx |^2$$

$$\leqq h^2 \sum_{j=1}^{N} \frac{1}{m(B_j)} \int_{S_j} | \nabla v_h |^2 dx m(S_j)$$

$$\leqq h^2 (n+1)^2 \| \nabla v_h \|_{L^2(\Omega)}{}^2 .$$

Hence (3.2) implies

(3.4) $\quad \| K_h{}^*P_{0h}v_h - L_h v_h \|_{L^2(\Omega)} \leqq (n+1)h \| v \|_V$.

Using Proposition 5, we have

$$\| K_h K_h{}^*P_{0h}v - v \|_{L^2(\Omega)}$$

$$\leqq \| (K_h K_h{}^*P_{0h} - 1)(v - v_h) \|_{L^2(\Omega)} + \| K_h K_h{}^*P_{0h}v_h - v_h \|_{L^2(\Omega)}$$

$$\leqq 2 \| v - v_h \|_{L^2(\Omega)} + \| K_h{}^*P_{0h}v_h - L_h v_h \|_{L^2(\Omega)}$$

$$\leqq \text{(by (3.2) and (3.4))}$$

$$\leqq Ch \| v \|_V .$$

Thus Condition III is established. Now let $\bar{A}_h = K_h{}^*A_h K_h$. Then $\bar{A}_h$ is a positive self adjoint operator acting in $\bar{V}_h$. So we have

$$\| e^{-t\bar{A}_h} \bar{v}_h \|_{L^2(\Omega)} \leqq \| \bar{v}_h \|_{L^2(\Omega)}, \quad t \geqq 0, \ \bar{v}_h \in \bar{V}_h .$$

Since $e^{-t\tilde{A}_h} = K_h e^{-t\bar{A}_h} L_h$,

$$\| e^{-t\tilde{A}_h} v_h \|_{L^2(\Omega)}$$

$\leqq$ (Proposition 5)

$\leqq \| L_h v_h \|_{L^2(\Omega)}$

$\leqq L_{2,n} \| v_h \|_{L^2(\Omega)}$.

Therfore Condition IV holds with $M_0 = L_{2,n}$.

The following Proposition completes the proof of Theorem 2 for m=2.

<u>Proposition 9.</u> Assume $\mathcal{T}_h$ be locally symmetric. Let X be $L_2(\Omega)$, and let Y and Z be $H_0^1(\Omega) \cap H^2(\Omega)$. Then Conditions II and III hold with some scalar functions $\varepsilon(h)$ and $\delta(h)$ which behave $O(h^2)$ as h tends to 0.

Proof. By a standard argument, we have

(3.5) $\| P_{1h} v - v \|_{L^2(\Omega)} \leqq Ch^2 \| v \|_{H^2(\Omega)}$, $v \in D(A) = H_0^1(\Omega) \cap H^2(\Omega)$,

(see Ciarlet[1]). This gives Condition II . To prove Condition III, first we note that $\Pi_h v$ is well defined for $v \in H^2(\Omega)$, for $H^2(\Omega) \subset C(\bar{\Omega})$ holds by Sobolev imbedding Theorem since we assumed $n \leqq 3$.

And we have

(3.6) $\| \Pi_h v - v \|_{L^2(\Omega)} \leqq Ch^2 \| v \|_{H^2(\Omega)}$ for $v \in H_0^1(\Omega) \cap H^2(\Omega)$,

(see Ciarlet [1]). Now we admit the following estimate (3.7) for a while.

(3.7) $\| K_h^* P_{0h} v - L_h \Pi_h v \|_{L^2(\Omega)} \leqq Ch^2 \| v \|_{H^2(\Omega)}$, $v \in H^2(\Omega)$.

Then we have for $v \in H_0^1(\Omega) \cap H^2(\Omega)$,

$\| K_h K_h^* P_{0h} v - v \|_{L^2(\Omega)}$

$\leqq \| K_h \|_{L(L^2(\Omega))} \| K_h^* P_{0h} v - L_h \Pi_h v \|_{L^2(\Omega)} + \| \Pi_h v - v \|_{L^2(\Omega)}$

$\leqq$ (by Proposition 5, (3.7), and (3.6))

$\leqq Ch^2 \| v \|_{H^2(\Omega)}$.

This gives Condition III.

Now we proceed to establish (3.7). Proposition 7 implies

$$(3.8) \quad \| K_h * P_{0h} v - L_h \Pi_h v \|^2_{L^2(\Omega)}$$
$$= \sum_{j=1}^{N} \frac{1}{m(B_j)} | \int_{S_j} (u(x)-u(b_j)) w_j(x)\,dx |^2 .$$

Let $\xi = x - b_j$, $u_j(\xi) = u(\xi + b_j)$, $w(\xi) = w_j(\xi + b_j)$ and $S = S_j - b_j$. Then we have

$$u(x) - u(b_j) = u_j(\xi) - u_j(0)$$
$$= (\nabla u_j(0), \xi)_n + \int_0^1 t (D^2 u_j(t\xi)\xi, \xi)_n dt,$$

where $D^2 u_j(\xi)$ the following n×n matrix:

$$D^2 u_j(\xi) = \left(\frac{\partial^2 u}{\partial x_j \partial x_k} (\xi + b_j)\right)_{1 \leqq j, k \leqq n} \cdot$$

Since S is symmetric with respect to the origin, we have

$$\int_{S_j} (u(x) - u(b_j)) w_j(x)\,dx$$
$$= \int_S (u_j(\xi) - u_j(0)) w(\xi)\,d\xi$$
$$= \int_S (\int_0^1 t((D^2 u_j)(t\xi)\xi, \xi)dt) w(\xi)\,d\xi$$
$$\leqq \int_0^1 t (\int_S \| D^2 u_j(t\xi) \|_{\mathbb{R}^n} \| \xi \|^2_{\mathbb{R}^n} d\xi) dt$$
$$\leqq \text{(by Schwartz inequality and } \| \xi \| \leqq \text{diam}(S)/2)$$
$$\leqq (\text{diam}(S)/2)^2 \int_0^1 t \| D^2 u_j(t\xi) \|_{L^2(S)} dt \ m(S)^{1/2}$$
$$= h^2 m(S)^{1/2} \int_0^1 t^{1-n/2} \| D^2 u_j(\xi) \|_{L^2(tS)} dt$$
$$\leqq h^2 m(S)^{1/2} \int_0^1 t^{1-n/2} \| D^2 u_j(\xi) \|_{L^2(S)} dt$$
$$= h^2 m(S)^{1/2} \frac{2}{4-n} \| D^2 u \|_{L^2(S_j)},$$

Substituting this estimate into (3.8), we have

$$\| K_h * P_{0h} v - L_h \Pi_h v \|^2_{L^2(\Omega)}$$
$$\leqq \left(\frac{2}{4-n}\right)^2 h^4 \sum_{j=1}^{N} \frac{m(S_j)}{m(B_j)} \| D^2 u \|^2_{L^2(S_j)}$$
$$\leqq (n+1)^2 \left(\frac{2}{4-n}\right)^2 h^4 \| D^2 u \|^2_{L^2(\Omega)}$$

This assures the validity of (3.7).

§4. L^∞-estimates for nonnegative triangulation.

Since we have prepared our abstract theory aiming a general purpose routine, we can automatically obtain some L^∞-estimates gathering already established results if the triangulation is restricted to be nonnegative. Following Ciarlet-Raviart [3], the triangulation $\mathcal{T}_h$ is said to be nonnegative if and only if it holds

(4.1) $a(w_i, w_j) \leqq 0$, for $i \neq j$, $1 \leqq i \leqq N$, $1 \leqq j \leqq N+M$.

In 2-dimensional problem, (4.1) is equivalet to the requirement that all the angles of triangles T of $\mathcal{T}_h$ are not greater than $\pi/2$. Following Fujii [5], this triangulation is said to be of acute type. Throughout this §, triangulations are assumed to be nonnegative.

Let $X=L^\infty(\Omega)$, $Y=W^{2,\infty}(\Omega)\cap V$ and $Z=W^{1,\infty}(\Omega)\cap V$. Then Condition I is trivially satisfied. Due to Ciarlet-Raviart [3], the following estimate (4.2) holds.

(4.2) $\| P_{1h}v-v\|_{L^\infty(\Omega)} \leqq Ch\| v\|_{W^{2,\infty}(\Omega)}$, for $v\in W^{2,\infty}(\Omega)\cap V$.

This implies the condition II with $\varepsilon(h)=O(h)$. Proposition 7 assures

(4.3) $\| K_h K_h{}^* P_{0h}v-\Pi_h v\|_{L^\infty(\Omega)} \lessapprox h\| v\|_{W^{1,\infty}(\Omega)}$.

Moreover mean value theorem implies

(4.4) $\| \Pi_h v-v\|_{L^\infty(\Omega)} \lessapprox h\| v\|_{W^{1,\infty}(\Omega)}$, $v\in W^{1,\infty}(\Omega)\cap V$.

Condition III with $\delta(h)=O(h)$ follows from (4.3) and (4.4). Due to Fujii [5], we have

(4.5) $\| e^{-t\bar{A}_h} \bar{v}_h\|_{L^\infty(\Omega)} \leqq \| \bar{v}_h\|_{L^\infty(\Omega)}$, $t>0$, $\bar{v}_h\in\bar{V}_h$.

(See p487 of Ushijima [10]. It is noted that the nonnegativity of triangulation is essential to establish (4.2) and (4.5).)

Since $e^{-t\tilde{A}_h}=K_h e^{-t\bar{A}_h}L_h$,

$$\| e^{-t\tilde{A}_h}v_h\|_{L^\infty(\Omega)} \leqq \text{(Proposition 5 and (4.5))}$$
$$\leqq \| L_h v_h\|_{L^\infty(\Omega)}$$
$$= \| v_h\|_{L^\infty(\Omega)}$$

holds for $v_h \epsilon V_h$. Thus Conditions IV with $M_0=1$ holds good. Hence Theorem 3 implies the following result.

<u>Theorem 10.</u> Let Ω be a bounded convex polygonal domain in $\mathbb{R}^n$ with arbitrary n. Assume that the triangulation $\mathcal{T}_h$ is nonnegative for any h. Then there is a constant C with the property that the estimate

$$\max_{0\leqq t\leqq T}\| u_h(t)-u(t)\|_{L^2(\Omega)}$$
$$\leqq C\{h(\max_{0\leqq t\leqq T}\| u(t)\|_{W^{2,\infty}(\Omega)}+\max_{0\leqq t\leqq T}\| \frac{\partial u}{\partial t}\|_{W^{2,\infty}(\Omega)}$$
$$+\max_{0\leqq t\leqq T}\| f(t)\|_{W^{1,\infty}(\Omega)})+\| a_h-P_{1h}a\|_{L^\infty(\Omega)}$$
$$+\max_{0\leqq t\leqq T}\| f_h-f\|_{L^\infty(\Omega)}\}$$

holds for the solution u(t) of (E) and the solution $u_h(t)$ of (E_h) provided that $u(t)\epsilon H_0^1(\Omega)\cap W^{2,\infty}(\Omega)$, $Au(t)\epsilon H_0^1(\Omega)\cap W^{1,\infty}(\Omega)$ and $\frac{\partial u}{\partial t}\ \epsilon H_0^1(\Omega)\cap W^{2,\infty}(\Omega)$, $0\leqq t\leqq T$.

A result of O(h)-convergence of this problem is already obtained by Tabata [8].

It is also possible to utilize the recent result of L^∞-estimate for the stationary problem due to Nitsche [7]. We restrict our problem to the 2-dimensional case, and assume further that our family of triangulation $\{\mathcal{T}_h: h>0\}$ satisfies the inverse assumption. Namely there is a positive constant ν satisfying that

(4.6) $\quad h \leqq \nu h(T) \quad$ for $T \in \mathcal{T}_h, h>0$.

Then Nitsche proves the following estimate (4.7):

(4.7) $\quad \| P_{1h}v - v \|_{L^\infty(\Omega)} \leqq Ch^2 |\log h| \; \|v\|_{W^{2,\infty}(\Omega)}$

provided that $v \in V \cap W^{2,\infty}(\Omega)$.

Theorem 11. Let Ω be a bounded convex polygonal domain in $\mathbb{R}^2$. Assume that the family of regular nonnegative triangulation $\{\mathcal{T}_h: h>0\}$ of Ω satisfies the inverse assumption (4.6), and that $\mathcal{T}_h$ is locally symmetric for any $h>0$. Then there is a constant C with the property that the estimate

$$\max_{0\leqq t\leqq T} \| u_h(t) - u(t) \|_{L^2(\Omega)}$$

$$\leqq C\{h^2 |\log h| (\max_{0\leqq t\leqq T} \|u(t)\|_{W^{2,\infty}(\Omega)} + \max_{0\leqq t\leqq T} \|\frac{\partial u}{\partial t}\|_{W^{2,\infty}(\Omega)}$$

$$+ \max_{0\leqq t\leqq T} \|f(t)\|_{W^{2,\infty}(\Omega)}) + \|a_h - P_{1h}a\|_{L^\infty(\Omega)}$$

$$+ \max_{0\leqq t\leqq T} \|f_h - f\|_{L^\infty(\Omega)}\},$$

holds for the solution u(t) of (E) and the solution $u_h(t)$ of (E_h) provided that u(t), Au(t), $\frac{\partial u}{\partial t}(t) \in H_0^1(\Omega) \cap W^{2,\infty}(\Omega)$, $0 \leqq t \leqq T$.

Proof. Let $X = L^\infty(\Omega)$, and let Y and Z be $D(A) \cap W^{2,\infty}(\Omega) = H_0^1(\Omega) \cap W^{2,\infty}(\Omega)$. Then Condition I is trivial. Condition II follows from (4.7) and Condition IV with $M_0 = 1$ is already shown. Condition IV is also proven by a step by step modification of the proof given in Proposition 9. Namely instead of (3.6) and (3.7), we have

(4.8) $\quad \| \Pi_h v - v \|_{L^\infty(\Omega)} \leqq Ch^2 \|v\|_{W^{2,\infty}(\Omega)} \quad$ for $v \in H_0^1(\Omega) \cap W^{2,\infty}(\Omega)$.

(4.9) $\| K_h * P_{0h} v - L_h \Pi_h v \|_{L^\infty(\Omega)} \leqq Ch^2 \| v \|_{W^{2,\infty}(\Omega)}$ for $v \in W^{2,\infty}(\Omega)$.

Corollary 5 implies the conclusion of Theorem.

References

[1] Ciarlet, P.G., The finite element method for elliptic problems, (North Holland, Amsterdam, 1977).

[2] Ciarlet, P.G., Raviart, P.A., General Lagrange and Hermite interpolation in R^n with applications to finite element methods, Arch. Rational Mech. Anal., 46, 177-199(1972).

[3] Ciarlet, P.G., Raviart, P.A., Maximum principle and uniform convergence for the finite element method, Comput. Meth. Appl. Mech. Engrg., 2, 17-31(1973).

[4] Courant, R., Variational methods for the solution of problems of equilibrium and vibrations, Bull. Amer. Math. Soc., 49, 1-23(1943).

[5] Fujii, H., Some remarks on finite element analysis of time-dependent field problems, IN Theory and practice in finite element structural analysis, 91-106 (Proceedings of 1973 Tokyo seminar of finite element analysis, Univ. Tokyo Press, Tokyo, 1973).

[6] Kadlec, J., On the regularity of the solution of the Poisson problem on a domain with boundary locally similar to the boundary of a covex open set (in Russian), Czech. Mat. J., 14, 386-393(1964).

[7] Nitsch, J.A., L^∞-convergence of finite element approximation, (Second conference on finite elements, Rennes, 1975).

[8] Tabata, M., Uniform convergence of the upwind finite element approximation for semilinear parabolic problems, J. Math. Kyoto Univ., 18, 327-351(1978).

[9] Tabata, M., L^∞-analysis of the finite element method, In Numerical analysis of evolution equations, 25-62 (Lecture notes in numerical and applied analysis 1, Kinokuniya, Tokyo, 1979).

[10] Ushijima, T., On the uniform convergence for the lumped mass approximation of the heat equation, Jour. Fac. Sci. Univ. Tokyo, Sec IA, 24, 477-490(1977).

Department of Information Mathematics
The University of Electro-Communications
1-5-1, Chofugaoka, Chofu-shi
Tokyo, 184, Japan.