Int. J. Solids Structures. 1967, Vol. 3, pp. 865 to 879. Pergamon Press Ltd. Printed in Great Britain

THE CONVERGENCE OF FINITE ELEMENT METHOD IN SOLVING LINEAR ELASTIC PROBLEMS

PIN TONG and T. H. H. PIAN

Massachusetts Institute of Technology

Abstract—This paper presents a theoretical development to show the sufficient conditions that will insure a finite element displacement analysis to converge to the exact displacement solutions when the size of the elements are progressively reduced. The order of such convergence is also estimated. The development is in connection with the three dimensional elasticity problem and the plate bending problem. A study is made to determine the possible means of evaluating the merits of different stiffness matrices to be used in the finite element analysis.

1. INTRODUCTION

THE various finite element displacement methods for static analysis of solid continuum have been identified by many authors [1–3] as the application of variational principles in mechanics in one form or another. A method which is based on assumed continuous and piecewise differentiable displacement functions is generally considered as an application of the minimum potential energy principle and the Ritz method. It has been assumed that the method will yield solutions which are converging to the exact solutions when the sizes of the finite elements are progressively reduced. However, such a statement is by no means trivial and, indeed, does not apply if some proper conditions have not been satisfied.

In general, an upper and a lower bound of the solution of the mathematical problem with positive definite variational functional can be obtained by the hypercircle method [4]. The present study will use a different approach and will give a theoretical account of the sufficient conditions that will insure the convergence of the numerical solution.* The theoretical development will also provide a means for estimating the order of convergence. The problem of three dimensional elasticity is considered first and is followed by a treatment of plate and shell problems. Finally, a study is made to determine the possible criteria that can be used to evaluate the merits of different stiffness matrices.

2. THREE DIMENSIONAL ELASTICITY

Let us consider an elastic body of volume V in equilibrium under external loads as shown in Fig. 1, bounded by a surface $\partial V = S_\sigma + S_u$. S_σ is the portion of the boundary surface where surface traction T_0 is prescribed and S_u is the surface where displacement $\underline{u}_0$ (continuous) is prescribed. The minimum potential energy theorem [5, 6] says that all of the admissible displacements (satisfying the prescribed displacement over S_u, being continuous in V and ∂V and having piecewise continuous second derivatives in V), the one

* After submission of the paper, the work of S. W. Key, A convergence investigation of the direct stiffness method, (Ph.D. Thesis, Univ. of Washington 1966) has been brought to the authors' attention. In this work, a similar idea in the proof of convergence has been used.

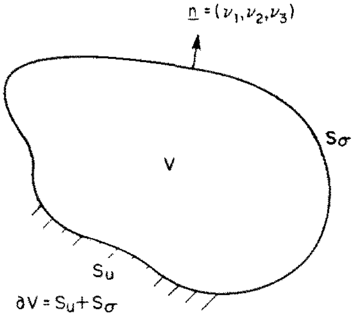
$\underline{n} = (\nu_1, \nu_2, \nu_3)$
S_σ
V
S_u
$\partial V = S_u + S_\sigma$

FIG. 1. Geometry and notation.

and only one which satisfies the equations of equilibrium

$$\frac{\partial}{\partial x_j}[C_{ijkl}e_{kl}] = F_i \qquad \text{in } V \tag{2.1}$$

and the boundary conditions

$$C_{ijkl}e_{kl}\nu_j = (T_i)_0 \qquad \text{on } S_\sigma \tag{2.2}$$

$(i = 1, 2, 3)$ is distinguished by a stationary value (minimum) of potential energy

$$\Pi(\underline{u}) = \tfrac{1}{2}D(\underline{u}) + \int_P F_i u_i \, \mathrm{d}V - \int_{S_\sigma} (T_i)_0 u_i \, \mathrm{d}S \tag{2.3}$$

$\tfrac{1}{2}D(\underline{u})$ is known as the strain energy and is defined by

$$D(\underline{u}) = D(\underline{u}, \underline{u}) \tag{2.4}$$

where

$$D(\underline{u}, \underline{u}) = \int_V C_{ijkl}e_{ij}e_{kl} \, \mathrm{d}V \tag{2.5}$$

$$e_{ij} = \tfrac{1}{2}(u_{i,j} + u_{j,i})$$

$\underline{u}$ is the admissible displacement vector with components (u_1, u_2, u_3). ν_j is the direction cosine of the normal on ∂V. The summation convention and rectangular cartesian coordinates (x_1, x_2, x_3) are used. C_{ijkl} are the elastic constants with

$$C_{ijkl} = C_{jikl} = C_{klij} \tag{2.6}$$

C_{ijkl} and F_i are assumed to be independent of u_i and $C_{ijkl}e_{ij}e_{kl}$ is assumed to be positive definite.

Let $\underline{u}$ be the solution of equation (2.1) and satisfy the prescribed boundary conditions, it is easy to see that one has

$$\Pi(\underline{u}) = -\tfrac{1}{2}D(\underline{u}) + \int_{S_u} C_{ijkl}e_{kl}\nu_j(u_i)_0 \, \mathrm{d}S \tag{2.7}$$

at the minimum of Π, and

$$D(\underline{u}, \underline{v}) = \int_{S_\sigma} (T_i)_0 v_i \, \mathrm{d}S + \int_{S_u} C_{ijkl} e_{kl} v_j v_i \, \mathrm{d}S - \int_V F_i v_i \, \mathrm{d}V \tag{2.8}$$

for any $\tilde{v}$ continuous first partial derivatives in V.

Let us divide the region V into small tetrahedrons, or in the case that V is axially symmetric, into circular rings of triangular cross-section. Each tetrahedron has a volume of the order ε^3 and a length of the edges of the order ε. For these tetrahedrons which are partly in V and partly outside V, we shall make their vertices either on ∂V or outside V. A typical tetrahedron is shown in Fig. 2. The position vector of the vertex n is denoted by $\underline{X}_n$. To each vertex n, subsequently called a nodal point, we assign the region D_n which is the domain comprised of all the tetrahedrons with nodal point n as their vertex.

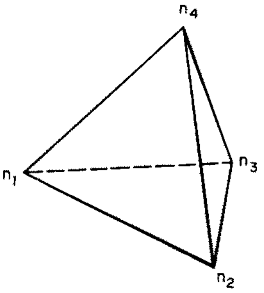

FIG. 2. Typical tetrahedral element.

Let $f_n(\underline{x})$ be given by

$$f_n(\underline{x}) = a_{np} + b_{np} x_1 + c_{np} x_2 + \mathrm{d}_{np} x_3 \tag{2.9}$$

in each tetrahedron p with point n as one of its vertices and let it be equal to zero outside D_n. The coefficients $a_{np} \ldots d_{np}$ in tetrahedron p are determined by requiring that

$$\begin{aligned} f_n(\underline{x}) &= 1 \quad \text{at } \underline{x} = \underline{X}_n \\ &= 0 \quad \text{at other three vertices} \end{aligned} \tag{2.10}$$

$f_n(\underline{x})$ is evidently continuous in V and subsequently will be called interpolation function. If the value of $\underline{u}$, at point $\underline{X}_n$ is denoted by $\underline{U}_n = (U_{1n}, U_{2n}, U_{3n})$, $\underline{u}$ can be written as

$$\underline{u} = \underline{u}^* + O(\varepsilon^2) \tag{2.11}$$

(since $\underline{u}$ has piecewise continuous second derivative), where

$$\underline{u}^* = \sum_n \underline{U}_n f_n(\underline{x}). \tag{2.12}$$

Substituting into equation (2.3), one can show that

$$\Pi(\underline{u}) = \Pi(\underline{u}^*) + O(\varepsilon) \tag{2.13}$$

Evidently

$$\Pi(u^*) \to \Pi(u) \tag{2.14}$$

as ε tends to zero.

Let us replace U_{in} of equation (2.12) by q_{in}, the so-called generalized coordinates, for those nodal points n which are not on S_u and denoted the new displacement function by $\underline{Q}$. Evidently, on S_u,

$$\underline{u}-\underline{Q} = \varepsilon \underline{h}(\underline{x}) \tag{2.15}$$

where $\underline{h}(\underline{x})$ is a bounded function. In fact, if $\underline{u}$ has piecewise continuous first partial derivative on S_u, $\underline{h}(\underline{x})$ is of the order ε on S_u and if $\underline{u}$ is zero on S_u, $\underline{h}(\underline{x})$ is also zero on S_u.

Since

$$\Pi(\underline{Q}) = \Pi(\underline{u})+\Pi(\underline{Q}-\underline{u})+D(\underline{u}, \underline{Q}-\underline{u}) \tag{2.16}$$

by equations (2.3) and (2.8)

$$\Pi(\underline{Q}) = \Pi(\underline{u})+\tfrac{1}{2}D(\underline{Q}-\underline{u})-\varepsilon E \tag{2.17}$$

where

$$E = \int_{S_u} C_{ijkl}e_{kl}v_j h_i(\underline{x})\,\mathrm{d}S \tag{2.18}$$

E is bounded and is independent of q_{in}. Thus one has

$$\Pi(\underline{Q})+\varepsilon E \geq \Pi(\underline{u}) \tag{2.19}$$

for all functions $\underline{Q}$ having piecewise continuous first partial derivatives and being equal to $\underline{u}$ at the nodal points on S_u. Write $\Pi(\underline{Q})$ in matrix form, i.e.

$$\Pi(\underline{Q}) = \tfrac{1}{2}D(\underline{Q})-\mathbf{q}^T\mathbf{T}+A \tag{2.20}$$

where

$$D(\underline{Q}) = \mathbf{q}^T\mathbf{K}\mathbf{q}$$

A is a constant which depends on $\underline{u}_0$ and is zero if $\underline{u}_0 = 0$ on S_u. $\mathbf{q}$ is the column vector with its components being the generalized coordinates q_{in}. $\mathbf{K}$ is symmetric and is called the stiffness matrix. If $\int_{S_u} \mathrm{d}S \neq 0$, $\mathbf{K}$ is positive definite, otherwise $\mathbf{K}$ is positive semi-definite.

The minimum of $\Pi(\underline{Q})$ is attained when $\mathbf{q}$ satisfies the equation

$$\mathbf{K}\mathbf{q} = \mathbf{T} \tag{2.21}$$

At the minimum, one has

$$\Pi(\mathbf{Q}) = -\tfrac{1}{2}\mathbf{q}^T\mathbf{K}\mathbf{q}+A = -\tfrac{1}{2}\mathbf{q}^T\mathbf{T}+A \tag{2.22}$$

and evidently, since $\underline{u}^* = \underline{Q}$ when $U_{in} = q_{in}$,

$$\Pi(\underline{u}^*) \geq \Pi(\underline{Q}) \tag{2.23}$$

By equation (2.19),

$$\Pi(\underline{u}^*)+\varepsilon E \geq \Pi(\underline{Q})+\varepsilon E \geq \Pi(\underline{u}) \tag{2.24}$$

Thus as ε tends to zero, the minimum of $\Pi(\underline{Q})$ tends to the exact minimum of $\Pi(\underline{u})$.

In order to show that the convergence of the approximate functional $\Pi(\underline{Q})$ to the exact functional $\Pi(\underline{u})$ actually implies the mean square convergence of the approximate solution $\underline{Q}$ itself to the exact solution u, i.e.

$$\int_V (\underline{u}-\underline{Q})^2 \, \mathrm{d}V \rightarrow 0 \tag{2.25}$$

as ε tends to zero, we would first consider the case that $\int_{S_u} \mathrm{d}S \neq 0$. In this case the solution $\underline{u}$ of equation (2.1) is unique and the stiffness matrix is positive definite. Let us consider the free vibration of the elastic body V with the density per unit volume to be unity, $F_i = 0$ in V, $\underline{u}_0 = 0$ on S_u and $(T_i)_0 = 0$ on S_σ. We see that the lowest eigenvalue, say λ, is positive and any continuous vector function $\underline{v} \neq 0$ in V with $\underline{v} = 0$ on S_u will satisfy

$$D(\underline{v}) \geq \lambda \int_V (\underline{v})^2 \, \mathrm{d}V. \tag{2.26}$$

Let $\underline{Q}$ be the function defined in equation (2.12) with its associated $\mathbf{q}$ satisfies equation (2.21). Let $\underline{f}$ be a continuous function in $V$and equal to $\underline{h}$ on S_u. Then, by equation (2.26)

$$D(\underline{u}-\underline{Q}-\varepsilon\underline{f}) \geq \lambda \int_V (\underline{u}-\underline{Q}-\varepsilon\underline{f})^2 \, \mathrm{d}V. \tag{2.27}$$

If $\underline{f}$ is so chosen that it vanishes everywhere except in the small neighborhood of S_u with volume of the order of ε, it can easily be seen that

$$\begin{aligned} D(\underline{u}-\underline{Q}-\varepsilon\underline{f}) &= D(\underline{u}-\underline{Q})+O(\varepsilon) \\ \int_V (\underline{u}-\underline{Q}-\varepsilon\underline{f})^2 \mathrm{d}V &= \int_V (\underline{u}-\underline{Q})^2 \, \mathrm{d}V + O(\varepsilon^2). \end{aligned} \tag{2.28}$$

Since

$$\begin{aligned} D(\underline{u}-\underline{Q}) &= D(\underline{u})+D(\underline{Q})-2D(\underline{u},\underline{Q}) \\ &= D(\underline{u})+D(\underline{Q})-2\int_{S_\sigma} (T_i)_0 Q_i \, \mathrm{d}S \\ &\quad +2\int_V F_i Q_i \, \mathrm{d}V - 2\int_{S_u} C_{ijkl}e_{kl}v_j(u_i)_0 \, \mathrm{d}S - 2\varepsilon \int_{S_u} C_{ijkl}e_{kl}v_i h_j \, \mathrm{d}S \\ &= -2\Pi(\underline{u})+2\Pi(\underline{Q})+2\varepsilon E \end{aligned}$$

Therefore, by equation (2.24)

$$D(\underline{u}-Q) \rightarrow 0 \quad \text{as} \quad \varepsilon \rightarrow 0$$

From equations (2.27) and (2.28), one concludes the mean square convergence of the approximate solution $\underline{u}$, i.e.

$$\int_V (\underline{u}-\underline{Q})^2 \, \mathrm{d}V \rightarrow 0 \quad \text{as} \quad \varepsilon \rightarrow 0$$

In the case of $\int_{S_u} \mathrm{d}S = 0$, i.e. stress is prescribed all over the boundary, the stiffness matrix is positive semidefinite. The solution for $\underline{u}$ or $\underline{Q}$ is unique only within a rigid body motion. But if one imposes additional conditions of removing the rigid body motion, then the proof of convergence is identical with the previous case.

In both cases, provided ∂V is sufficiently smooth, the rate of convergency of

$$\int_V (\underline{u}-\underline{Q})^2\,\mathrm{d}V \quad \text{and} \quad D(\underline{u}-\underline{Q})$$

are at least of the order ε.

We have shown that if the interpolation functions f_n, (1) are continuous in $V+\partial V$and have piecewise continuous first partial derivatives and (2) are able to approximate any sufficiently smooth function up to order ε^2 (see equation 2.11), the solution $\underline{Q}$ by the finite element analysis will tend to the exact solution in the sense of equation (2.25). The first condition is obviously also a necessary condition. If the approximate function $\underline{Q}$ is only piecewise continuous, say, $\underline{Q}$ has a finite jump across the boundary of a tetrahedron in V, then its first partial derivatives have an infinite jump over there; the square of such an infinite jump is not integrable, i.e. the strain energy $\frac{1}{2}D(\underline{Q})$ corresponding to such displacement $\underline{Q}$ is not defined at all. The total strain energy of the body is then not equal to the sum of the strain energy of each individual element. Using the latter to represent $D(\underline{Q})$ for minimization to obtain the approximate solution is meaningless. The second condition is evidently not necessary. In fact, what we have used is equation (2.24). If the interpolation functions can approximate closely the given u_0 on S_u and the approximate function $\underline{Q}$ satisfies equation (2.24) as ε tends to zero, then all the arguments used in this proof will follow.

3. PLATES AND SHELLS

In the linearized theory of plates, the inplane displacements and outplane displacements are decoupled. They can be treated separately. For the inplane displacement, the proof of convergence is exactly the same as that in the last section, except that here the region being considered is a plane. Therefore we only have to use triangular elements instead of tetrahedrons. We shall not consider it again. For the outplane displacement, the situation is a little different, because we have to deal with bending energy in the construction of the finite element equation. The bending energy of the plate involves second partial derivatives of the outplane displacement w. We may expect the continuity alone for the approximate function will not be sufficient. If the first partial derivative has a discontinuity at a certain point, then the second partial derivative has an infinite jump over there, and the bending energy of the plate is not defined at all. Therefore as was pointed out by Pian [7] the continuity of the first partial derivatives is a necessary condition for the assumed outplane displacement.

For the outplane displacement of a plate, one has to find w, which satisfies the equilibrium equation

$$\frac{\partial^2 m_{\alpha\beta}}{\partial x_\alpha \partial x_\beta} - \frac{\partial}{\partial x_\alpha} n_{\alpha\beta} \frac{\partial w}{\partial x_\beta} + q(x)w = p(x_\alpha) \tag{3.1}$$

in S and the appropriate boundary conditions on ∂S. In the expression,

$$m_{\alpha\beta} = C_{\alpha\beta\lambda\theta}\frac{\partial^2 w}{\partial x_\lambda \partial x_\theta}$$

and $n_{\alpha\beta}$ are stress couples and initial stress resultants respectively. Here cartesian coordinates are used and repeated indices indicate summation from one to two. For simplicity, we shall consider only simply supported boundary conditions, i.e.

$$w = w_0 \tag{3.2}$$

$$m_{\alpha\beta}\nu_\alpha\nu_\beta = M_0 \tag{3.3}$$

where ν_α is the direction cosine of the normal of ∂S.

Let the functional $D(w, \psi)$ and $D(w)$ be defined by

$$D(w, \psi) = \int_S \left[C_{\alpha\beta\lambda\theta}\frac{\partial^2 w}{\partial x_\alpha \partial x_\beta}\,\frac{\partial^2 \psi}{\partial x_\lambda \partial x_\theta} + n_{\alpha\beta}\frac{\partial w}{\partial x_\alpha}\,\frac{\partial \psi}{\partial x_\beta} + q(x)w\psi\right] \mathrm{d}S \tag{3.4}$$

$$D(w) = D(w, \psi)$$

Then the solution w of equations (3.1) to (3.3) minimizes the potential energy

$$\Pi(\phi) = \tfrac{1}{2}D(\phi) - \int_S p\phi \,\mathrm{d}S - \int_{\partial S} M_0 \frac{\partial \phi}{\partial n}\,\mathrm{d}l \tag{3.5}$$

over all admissible functions ϕ which are continuous, satisfy the rigid boundary condition (3.2) and have piecewise continuous second partial derivatives in S. If w is the solution of equations (3.1) to (3.3)

$$D(w, \psi) = \int_S p\psi \,\mathrm{d}S + \int_{\partial S} M_0 \frac{\partial \psi}{\partial n}\,\mathrm{d}l + \int_{\partial S} B(w)\psi \,\mathrm{d}l \tag{3.6}$$

where

$$B(w) = \frac{\partial}{\partial l}\left[m_{\alpha\beta}\nu_\beta \frac{\partial l}{\partial x_\alpha}\right] + \left(\frac{\partial m_{\alpha\beta}}{\partial x_\beta} - N_{\alpha\beta}\frac{\partial w}{\partial x_\beta}\right)\nu_\alpha$$

for all admissible ψ. In particular, setting $\psi = w$ and substituting into equation (3.5), one gets

$$\Pi(w) = -\tfrac{1}{2}D(w) - \int_{\partial S} B(w)w_0 \,\mathrm{d}l \tag{3.7}$$

which is the minimum of $\Pi(\phi)$ for all admissible ϕ.

In the proof of the convergence of the finite element method we first subdivide the region S into small nonoverlapping triangles or quadrilaterals with areas of order ε^2. To each vertex of the polygons, say at point P, we assign a subregion S_P which comprises all the adjacent polygons containing point P. Let $w(P)$ and $w_\alpha(P)$ denote the values of the function w and its first partial derivative with respect to x_α, respectively, at point P. Let the functions $f_P(x_1, x_2)$, $g_{P\alpha}(x_1, x_2)$, $(\alpha = 1, 2)$ be so defined that

$$f_P = \frac{\partial f_P}{\partial x_\alpha} = g_{P\alpha} = \frac{\partial g_{P\alpha}}{\partial x_\alpha} = 0 \qquad (\alpha = 1, 2) \qquad \text{on } \partial S_P \tag{3.8}$$

$$f_P = \frac{\partial g_{P\alpha}}{\partial x_\alpha} = 1 \qquad \text{at } P \tag{3.9}$$

$$\frac{\partial f_P}{\partial x_\alpha} = \frac{\partial g_{P\alpha}}{\partial x_\beta} = 0 \qquad (\alpha \neq \beta) \qquad \text{at } P \tag{3.10}$$

both f_P and $g_{P\alpha}(\alpha = 1, 2)$ vanish outside S_P. The functions f_P and $g_{P\alpha}$ so defined are called the interpolation functions.

For the present problem, if

$$w = w^*(\underline{x}) + O(\varepsilon^3) \qquad \text{in S} \tag{3.11}$$

where

$$w^* = \sum_P \left[w(P) f_P + \sum_{\alpha=1}^{2} w_\alpha(P) g_{P\alpha} \right] \tag{3.12}$$

for all sufficiently smooth functions w, then the proof of the convergence of the finite element solution to the exact solution are similar to that in the last section. We shall just sketch the step and indicate the results. Evidently

$$\Pi(w^*) \to \Pi(w)$$

as $\varepsilon \to 0$. The solution of the finite element method is obtained by minimizing

$$\Pi(F)$$

where

$$F = \sum_P \left[q(P) f_P + \sum_{\alpha=1}^{2} q_\alpha(P) g_{P\alpha} \right] \tag{3.13}$$

with respect to all $q(P)$ and $q_\alpha(P)$ subject to the condition that $q(P) = w(P)$ for $P \in \partial S$. The end result is

$$\tfrac{1}{2}[\Pi(F) - \Pi(w)] \geq D(w - F) + O(\varepsilon^2) \geq \lambda \int_S (w - F)^2 \, dS + O(\varepsilon^2) > 0 \tag{3.14}$$

while

$$\Pi(F) - \Pi(w) \to 0$$

as $\varepsilon \to 0$.

In the case of shells, the inplane and outplane displacements are in general coupled. Thus we have to consider them simultaneously in the potential energy of the shell. The argument used in the convergence proof is identically the same. We shall not repeat the proof.

Clough and Tocher [8] have suggested a procedure for constructing a displacement function to be used in the derivation of the stiffness matrix of a triangular plate element in bending. Let us show that their scheme does lead to an interpolation function that satisfies equations (3.8)–(3.12). Consider a typical triangle ΔPQR which comprises three subtriangles, ΔA, ΔB, and ΔC as shown in Fig. 3. Let

$$W_A = a_1 + a_2 x_1 + a_3 x_2 + \ldots + a_9 x_1 x_2^2 + a_{10} x_2^3 \tag{3.15}$$

be a polynomial of degree three defined in ΔA. To determine the ten a's, W_A and its first

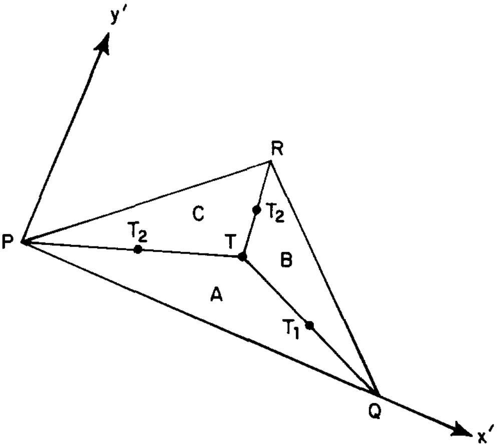

Fig. 3. Triangular element used in plate and shell problems.

partial derivatives are required to equal to that of w at the vertices of ΔA. So far only nine conditions are imposed on the a's. Let us consider a local orthogonal coordinate x', y' for ΔA as shown in Fig. 3. Thus by requiring that $\partial W_A/\partial y'$ is only a linear function of x' on the x'-axis, the a's of equations (3.15) for W_A are uniquely determined in terms of the values of w and its partial derivatives at the vertices of triangle A. Write W_A in a slightly different form as

$$\begin{aligned} W_A = {} & w(P)\varphi_P^A(x_1, x_2) + w(Q)\varphi_Q^A(x_1, x_2) + w(T)\varphi_T^A(x_1, x_2) \\ & + \sum_{\alpha=1}^{2} [w_\alpha(P)\varphi_{P\alpha}^A(x_1, x_2) + \ldots + w_\alpha(T)\varphi_{T\alpha}^A(x_1, x_2)] \end{aligned} \tag{3.16}$$

where the functions φ are independent of the value of w. Similarly W_B and W_C are defined in the corresponding subtriangles ΔB and ΔC. For a sufficiently smooth function w, it is easy to see

$$\begin{aligned} w &= W_A + \varepsilon^3 E_A(x_1, x_2) \quad && \text{in } \Delta A \\ &= W_B + \varepsilon^3 E_B(x_1, x_2) && \text{in } \Delta B \\ &= W_C + \varepsilon^3 E_C(x_1, x_2) && \text{in } \Delta C \end{aligned} \tag{3.17}$$

where E_A, E_B and E_C are functions of the same order as w.

Let $\partial/\partial n_{MN}$ denote the normal derivative across the line segment MN. Then

$$\frac{\partial W_A}{\partial n_{QT}} = \frac{\partial W_B}{\partial n_{QT}} \quad \text{at } Q \text{ and } T \tag{3.18}$$

$$\frac{\partial W_B}{\partial n_{RT}} = \frac{\partial W_C}{\partial n_{RT}} \quad \text{at } R \text{ and } T \tag{3.19}$$

$$\frac{\partial W_C}{\partial n_{PT}} = \frac{\partial W_A}{\partial n_{PT}} \quad \text{at } P \text{ and } T \tag{3.20}$$

for all values of $w(P)$, $w(Q)$, $w(R)$, $w(T)$, $w_\alpha(P)$, ... $w_\alpha(T)$. In the scheme by Clough and Tocher the displacement functions which are denoted by W_A^*, W_B^* and W_C^* are independent of $w(T)$, $w_1(T)$ and $w_2(T)$. Indeed, they are obtained by replacing $w(T)$, $w_1(T)$ and $w_2(T)$ in the expressions of W_A, W_B and W_C by θ, θ_1 and θ_2 respectively. θ, θ_1, and θ_2 are defined by requiring that equations (3.18) and (3.19) are also satisfied at points T_1, and T_2, and T_3 of QT, RT, and PT respectively. Since $\partial W_A^*/\partial n$ ... are only quadratic equations, equations (3.18)(3.20) are satisfied on QT, RT, and PT for W_A^*, W_B^* and W_C^*; the values of θ, θ_1, θ_2 are independent of the choice of T_1, T_2, and T_3.

By equations (3.16) and (3.17), one has

$$\begin{aligned}
&[\theta - w(T)]\left[\frac{\partial\varphi_T^A}{\partial n_{QT}} - \frac{\partial\varphi_T^B}{\partial n_{QT}}\right] + \sum_{\alpha=1}^{2}\,[\theta_\alpha - w_\alpha(T)]\left[\frac{\partial\varphi_{T\alpha}^A}{\partial n_{QT}} - \frac{\partial\varphi_{T\alpha}^B}{\partial n_{QT}}\right] = \varepsilon^3\left[\frac{\partial E_A}{\partial n_{QT}} - \frac{\partial E_B}{\partial n_{QT}}\right] \quad \text{at } T_1 \\
&[\theta - w(T)]\left[\frac{\partial\varphi_T^B}{\partial n_{RT}} - \frac{\partial\varphi_T^C}{\partial n_{RT}}\right] + \sum_{\alpha=1}^{2}\,[\theta_\alpha - w_\alpha(T)]\left[\frac{\partial\varphi_{T\alpha}^B}{\partial n_{RT}} - \frac{\partial\varphi_{T\alpha}^C}{\partial n_{RT}}\right] = \varepsilon^3\left[\frac{\partial E_B}{\partial n_{RT}} - \frac{\partial E_C}{\partial n_{RT}}\right] \quad \text{at } T_2 \\
&[\theta - w(T)]\left[\frac{\partial\varphi_T^C}{\partial n_{PT}} - \frac{\partial\varphi_T^A}{\partial n_{PT}}\right] + \sum_{\alpha=1}^{2}\,[\theta_\alpha - w_\alpha(T)]\left[\frac{\partial\varphi_{T\alpha}^C}{\partial n_{PT}} - \frac{\partial\varphi_{T\alpha}^A}{\partial n_{PT}}\right] = \varepsilon^3\left[\frac{\partial E_C}{\partial n_{PT}} - \frac{\partial E_A}{\partial n_{PT}}\right] \quad \text{at } T_3
\end{aligned} \tag{3.21}$$

Evidently

$$\begin{aligned}
\theta - w(T) &= O(\varepsilon^3) \\
\theta_\alpha - w_\alpha(T) &= O(\varepsilon^3) \qquad (\alpha = 1, 2)
\end{aligned} \tag{3.22}$$

Or, in turn, one has

$$\begin{aligned}
W_A - W_A^* &= O(\varepsilon^3) \qquad \text{in } \triangle A \\
W_B - W_B^* &= O(\varepsilon^3) \qquad \text{in } \triangle B \\
W_C - W_C^* &= O(\varepsilon^3) \qquad \text{in } \triangle C
\end{aligned} \tag{3.23}$$

Write W_A^* in the form

$$\begin{aligned}
W_A^* = {} & w(P)f_P^A(x_1, x_2) + w(Q)f_Q^A(x_1, x_2) \\
& + \sum_{\alpha=1}^{2}\,[w_\alpha(P)g_{P\alpha}^A(x_1, x_2) + w_\alpha(Q)g_{Q\alpha}^A(x_1, x_2)]
\end{aligned} \tag{3.24}$$

and write W_B^* and W_C^* in the similar form. Then one can define the interpolation functions f_p and $g_{P\alpha}$ by letting

$$\begin{aligned} f_P &= f_P^A(x_1, x_2) && \text{in } \triangle A \\ &= f_P^B(x_1, x_2) && \text{in } \triangle B \\ &= f_P^C(x_1, x_2) && \text{in } \triangle C \\ g_{P\alpha} &= g_{P\alpha}^A(x_1, x_2) && \text{in } \triangle A \\ &= g_{P\alpha}^B(x_1, x_2) && \text{in } \triangle B \\ &= g_{P\alpha}^C(x_1, x_2) && \text{in } \triangle C \end{aligned}$$

of the triangle PQR. Similarly f_P and $g_{P\alpha}$ are defined in other triangles of S_P. Both f_P and $g_{P\alpha}$ are set to be zero outside S_P. It can be easily seen that the so defined interpolation functions f_P and $g_{P\alpha}$ have continuous first partial derivatives for all points in S and satisfy equations (3.8)–(3.12).

In the case of rectangular elements, one can easily show the corrected version of the interpolation functions constructed in [9] also satisfy equation (3.8)–(3.12). Using that kind of interpolation function in the finite element scheme will also guarantee the convergence of its solution. In the case of axial symmetric structures, e.g. shells of revolution, or circular plates, the numerical scheme introduced in [10] and [11] will also converge to the exact solution.

4. EVALUATION OF STIFFNESS MATRIX

There are many ways to construct the interpolation functions that satisfy the sufficient conditions for the mean square convergence. The question is for a given subdividing region which interpolation functions will give the better approximation. Khanna [12] has postulated that of two element stiffness matrices the one which gives the greater strain energy under all load vectors will give consistently better results in the numerical analysis. He also pointed out that the comparison of the strain energy difference can be accompanied by examining the inverse difference $[\mathbf{k}_2^{-1} - \mathbf{k}_1^{-1}]$, where $\mathbf{k}_1$ and $\mathbf{k}_2$ are the element stiffness matrices. Let us re-examine this problem.

As we know from the minimum potential energy theorem, for all the functions $\underline{u}$ which satisfy the restrained boundary condition and have piecewise continuous first partial derivatives (in the case of plate or shell, the normal displacement must have piecewise continuous second partial derivatives), the exact solution corresponding to the minimum of the potential energy providing the strain energy of the elastic body is positive definite.

Let $\underline{U}_i$'s be different approximate solutions with its associated stiffness matrices $\mathbf{K}_i$ which are obtained based on continuous displacements at the interelement boundary. Then

$$\mathbf{K}_i \mathbf{U}_i = \mathbf{T}_i \tag{4.1}$$

(For the problem with only stress prescribed over the boundary, $\mathbf{K}_i$ denotes the modified matrix so that $\mathbf{K}_i$ is nonsingular.) By equations (2.27) and (2.28)

$$\begin{aligned}\int_V (\underline{U}_i-\underline{u})^2\,\mathrm{d}V &\leq \frac{1}{\lambda}[D(\underline{U}_i-\underline{u})+O(\varepsilon)] \\ &= \frac{2}{\lambda}[\Pi(\underline{U}_i)-\Pi(\underline{u})]+O(\varepsilon)]\end{aligned} \tag{4.2}$$

with $\lambda > 0$. In the case of no body force and no external force acting in the interior of the region considered, T_i depends only on the interpolation function over the boundary. The term $O(\varepsilon)$ in equation (4.2) depends only on the interpolation functions over the boundary where restrained boundary conditions are prescribed; and if the restrained conditions are homogeneous, the term $O(\varepsilon)$ is identically zero. If the bound of

$$\int_V (\underline{U}_i-\underline{u})^2\,\mathrm{d}V$$

is used as a criterion to justify the approximations, of all the approximations for a problem with homogeneous restrained boundary conditions, or of all those approximations having the same interpolation functions over the restrained boundary, the approximation with smaller $\Pi(\underline{U}_i)$ is the better. By equations (4.1) and (2.22)

$$\begin{aligned}\Pi(\underline{U}_i) &= \tfrac{1}{2}\mathbf{U}_i^T\mathbf{K}_i\mathbf{U}_i-\mathbf{U}_i^T\mathbf{T}_i+A \\ &= -\tfrac{1}{2}\mathbf{U}_i^T\mathbf{K}_i\mathbf{U}_i+A = -\tfrac{1}{2}\mathbf{U}_i^T\mathbf{T}_i+A\end{aligned} \tag{4.3}$$

Thus one may say the larger strain energy $\frac{1}{2}\mathbf{U}_i\mathbf{K}_i\mathbf{U}_i$ is the better.

If all the approximations have the same interpolation functions over the entire boundary, and no load is applied in the interior of the region, one has

$$\mathbf{T}_i = \mathbf{T}$$

equation (4.3) becomes

$$\Pi(\underline{U}_i) = -\tfrac{1}{2}\mathbf{T}^T\mathbf{K}_i^{-1}\mathbf{T}+A \tag{4.4}$$

Take $i = 1$ and 2, one has

$$\Pi(\underline{U}_1)-\Pi(\underline{U}_2) = \tfrac{1}{2}\mathbf{T}^T(\mathbf{K}_2^{-1}-\mathbf{K}_1^{-1})\mathbf{T} \tag{4.5}$$

Thus if $\mathbf{K}_2^{-1}-\mathbf{K}_1^{-1}$ is positive definite or semidefinite

$$\Pi(\underline{U}_1) \geq \Pi(\underline{U}_2) \tag{4.6}$$

In fact $\mathbf{K}_2^{-1}-\mathbf{K}_1^{-1}$ is symmetric and positive definite or semidefinite only if $\mathbf{K}_1-\mathbf{K}_2$ is so. For symmetric matrices $\mathbf{K}_1$ and $\mathbf{K}_2$, it is obvious that if $\mathbf{K}_1-\mathbf{K}_2$ is symmetric, so is $\mathbf{K}_2^{-1}-\mathbf{K}_1^{-1}$. If

$$\mathbf{x}^T(\mathbf{K}_1-\mathbf{K}_2)\mathbf{x} \geq 0$$

for all $\mathbf{x} \not\equiv 0$, then

$$\mathbf{x}^T(\mathbf{K}_1-\mathbf{K}_2)\mathbf{x} \geq C\mathbf{x}^T\mathbf{K}_2\mathbf{x} \tag{4.7}$$

where C is nonnegative and is equal to the ratio of the smallest eigenvalue of $\mathbf{K}_1 - \mathbf{K}_2$ to the largest eigenvalue of $\mathbf{K}_2$. Equation (4.7) implies that all the eigenvalues μ of the equation

$$\mathbf{V}^T(\mathbf{K}_1 - \mathbf{K}_2) = \mu \mathbf{V}^T \mathbf{K}_2 \tag{4.8}$$

is nonnegative. If $\mathbf{V}$ satisfies equation (4.8), one has

$$\mathbf{V}^T(\mathbf{K}_1 \mathbf{K}_2^{-1} - \mathbf{I}) = \mu \mathbf{V}^T$$

or

$$(\mathbf{K}_1 \mathbf{K}_2^{-1} - \mathbf{I})\mathbf{V} = \mu \mathbf{V}$$

or

$$(\mathbf{K}_2^{-1} - \mathbf{K}_1^{-1})\mathbf{V} = \mu \mathbf{K}_1^{-1} \mathbf{V} \tag{4.9}$$

Therefore we may say that all the eigenvalues of equation (4.9) are nonnegative. By equations (4.7)–(4.9), one concludes that

$$\mathbf{x}^T(\mathbf{K}_2^{-1} - \mathbf{K}_1^{-1})\mathbf{x} \geq C \mathbf{x}^T \mathbf{K}_1^{-1} \mathbf{x} \tag{4.10}$$

is nonnegative. In fact, if $\mathbf{K}_1 - \mathbf{K}_2$ is positive definite, C is positive, i.e. $\mathbf{K}_2^{-1} - \mathbf{K}_1^{-1}$ is also positive definite. Since $(\mathbf{K}_i^{-1})^{-1} = \mathbf{K}_i$, the properties of $\mathbf{K}_2^{-1} - \mathbf{K}_1^{-1}$ will imply those of $\mathbf{K}_1 - \mathbf{K}_2$. Thus one may justify which of the approximations is better by simply comparing their stiffness matrices. Of course, if $\mathbf{K}_i - \mathbf{K}_j$ is indefinite, one cannot say anything in general.

The result above, strictly speaking, can only apply to compare those approximations which have the same $\mathbf{T}$ (see equation 4.1). But if the size of the elements is sufficiently small, the difference in $\mathbf{T}$ for different approximations is small, and we may still use the properties of stiffness matrices as a crude justification.

It is easy to see that

$$\mathbf{U}^T(\mathbf{K}_1 - \mathbf{K}_2)\mathbf{U} = \sum_{\substack{\text{all} \\ \text{elements}}} \mathbf{u}^T(\mathbf{k}_i - \mathbf{k}_2)\mathbf{u} \tag{4.11}$$

$\mathbf{k}_i$, called the elemental stiffness matrix, is the stiffness matrix of a single element of the body associated with $\mathbf{K}_i$. $\mathbf{k}_i$ is symmetric and in general positive semidefinite. However, there is at least one elemental matrix which is positive definite, e.g. the elemental matrices of the elements adjacent to the rigid boundary. In the case stress prescribed problem, some rows and columns of some elemental stiffness matrices are removed to make the stiffness of the body to be nonsingular, some of those elemental matrices are then positive definite. $\mathbf{u}$ is a column vector with components to be the generalized coordinates of that particular element. Evidently, if $\mathbf{k}_1 - \mathbf{k}_2$ of all the elements is positive semidefinite, then $\mathbf{K}_1 - \mathbf{K}_2$ is positive semidefinite. This would be a very useful observation. If, for a problem, the interpolation functions, which satisfy the displacement compatibility at the interelement boundaries, are the same for all elements, one need only to examine one single elemental stiffness matrix instead of handling the larger stiffness matrix of the whole body. It appears that only in this restricted case can one apply Khanna's criterion for evaluating the finite element methods.

Special attention must be taken that the above proof is based on the minimum potential energy theorem. In the case that the stiffness matrix is not derived by using assumed displacement functions [7, 13], the functional used in the construction of the stiffness matrix is no longer positive definite [14]. Then the strain energy obtained from the approximate method can be either larger or smaller than that of the exact solution. Therefore, the above criteria can no longer be applied.

5. CONCLUSION

The sufficient condition for the finite element method to be convergent is established. The procedure used can be extended to derive higher order accurate approximations. All our discussions, so far, on finite element methods are restricted to the displacement model. Similar procedure can be extended to the so-called equilibrium model by the use of the complementary energy. The use of an equilibrium model will provide an upper bound to the strain energy. From the upper and lower bounds of the strain energy probably more can be said about the real state of convergence of the solution. In the case of a mixed model, e.g. Pian's method [7, 13], and those methods using Reissner's principle, the functional is no longer positive definite, but the conditions for the convergence of the solution can still be established [14].

Acknowledgements—This work was supported by United States Air Force Office of Scientific Research through Contract AF49(638)1364 with the California Institute of Technology and Grant OSR-347-65 with the Massachusetts Institute of Technology. Both authors were affiliated with the California Institute of Technology during the academic year 1965–1966. The authors also wish to express their appreciation to Professor Y. C. Fung for his many useful suggestions and comments.

REFERENCES

[1] R. J. MELOSH, *AIAA Jnl* **1**, 1631 (1963).
[2] J. F. BESSELING, The complete analogy between the matrix equations and the continuous field equations of structural analysis. *Proc. Int. Symp. on Analogue and Digital Techniques Applied to Aeron.*, Liege, 1963, pp. 22–242.
[3] B. FRAEIJS DE VEUBEKE, Displacement and equilibrium models in the finite element method. *Stress Analysis*, edited by O. C. ZIENKIEWICZ and G. S. HOLISTER, pp. 145–197. Wiley (1965).
[4] J. L. SYNGE, *The Hypercircle in Mathematical Physics*. Cambridge University Press (1957).
[5] Y. C. FUNG, *Foundations of Solid Mechanics*, chapter 10. Prentice-Hall.
[6] R. COURANT and D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, p. 247. Interscience (1953).
[7] T. H. H. PIAN, *AIAA Jnl* **2**, 1333 (1964).
[8] R. W. CLOUGH and J. L. TOCHER, Finite element stiffness matrices for analysis of plate bending. *Proc. Conf. on Matrix Methods in Structural Mechanics*, AFFDL-TR-66-80, Air Force Flight Dynamics Lab., Nov. 1966, pp. 515–545.
[9] F. K. BOGNER, R. L. FOX and L. A. SCHMIT, JR.: The generalization of interelement-compatible stiffness and mass matrices by the use of interpolation formulas. *Proc. Conf. on Matrix Methods in Structural Mechanics*, AFFDL-TR-66-80, Air Force Flight Dynamics Lab., Nov. 1966, pp. 397–443.
[10] J. H. PERCY, T. H. H. PIAN, S. KLEIN and D. R. NAVARATNA, *AIAA Jnl* **3**, 2138 (1965).
[11] P. TONG, Liquid sloshing in an elastic container. Ph.D. Thesis, California Institute of Technology (1966).
[12] J. KHANNA, *AIAA Jnl* **3**, 1976 (1965).
[13] T. H. H. PIAN, Element stiffness matrices for boundary compatibility and for prescribed boundary stresses. *Proc. Conf. on Matrix Methods in Structural Mechanics*, AFFDL-TR-66-80, Air Force Flight Dynamic Lab., Nov. 1966, pp. 457–477.
[14] P. TONG and T. H. H. PIAN, On the convergence of a finite element method based on assumed stress distribution. (In preparation.)

(*Received* 5 *December* 1966; *revised* 9 *March* 1967)

Résumé—Cet exposé présente un développement théorique pour montrer les conditions suffisantes qui assureront que l'analyse du déplacement d'un élément limité converge vers les solutions de déplacement exactes lorsque la taille des éléments est progressivement réduite. L'ordre d'une telle convergence est aussi évaluée. Le développement est relatif au problème d'élasticité à trois dimensions et au problème de la flexion des plaques. Une étude est faite pour déterminer les moyens possibles d'évaluer les avantages de différentes matrices de rigidité à être utilisées dans l'analyse de l'élément limité.

Zusammenfassung—Diese Arbeit gibt eine theoretische Entwicklung die die genügenden Bedingungen anzeigt, die eine Analyse in endlichen Verschiebungselementen zu genauen Verschiebungen konvergieren wenn die Elementen grössen dauernd abnehmen. Die Konvergierungs-Ordnung wird auch geschätzt. Die Entwicklung ist im Zusammenhang mit dem dreidimensionalen Elastizitätsproblem und mit dem Plattenbiegungsproblem. Eine Untersuchung der Bestimmungsmöglichkeiten, die verschiedenem Steifigkeitsmatrtzen die in der endlichen Analyse verwendet werden sollen wird auch unternommen.

Абстракт—В настоящей работе дается теоретический вывод для указания достаточных условий, которые обеспечивают сходимость расчета перемещений конечного злемента со строгими решениями в перемещениях в случае, когда размер злементов прогрессивно уменьшается. Приводится также оценку такой сходимости. Вывод связан з задачей трехмерной теории упругости и с задачей изгиба пластинки. Делается попытка, целью которой является указание возможных значений оценки достоинств матриц разных жесткостей. Эти матрицы используются при расчете конечного элемента.