Int. J. Solids Structures, 1968, Vol. 4, pp. 929 to 952. Pergamon Press. Printed in Great Britain

THEORETICAL FOUNDATIONS OF THE FINITE ELEMENT METHOD

EDUARDO R. DE ARANTES E OLIVEIRA

Applied Mathematics Division, Laboratório Nacional de Engenharia Civil, Lisboa, Portugal

Abstract—The finite element method is nowadays the most general and one of the most powerful tools for the analysis of structures.

It is also a general mathematical technique and the main concern of the paper is to present it in this light. Functional Analysis is used as the ideal frame for a general abstract formulation.

The ability to predict convergence to the exact solution of a sequence of approximate solutions obtained from patterns of finite elements with decreasing size is fundamental in the application of the method.

In case conformity between elements is obtained, the finite element method is a particular case of Ritz's method, so that convergence can be ensured as far as completeness is achieved.

A general completeness criterion is justified in the paper. Such criterion requires that the field components and all their derivatives, of order not higher than the highest order of derivative entering into the energy density expression, can take up any constant value within the element.

It is finally proved that such criterion is also a general convergence criterion, i.e. a sufficient condition for convergence even if conformity is not achieved.

NOTATION

All the symbols will be defined where they are introduced.

The following general conventions are adopted:

1. Matrices (or vectors) will be denoted by bold face symbols: $\mathbf{q}$, $\mathbf{H}$.
2. The dummy index convention will be used: $A_{ij}x_i = A_{1j}x_1 + A_{2j}x_2 + \ldots$.
3. A derivative will be denoted by a comma followed by indices indicating the variables with respect to which the function is differentiated. The order is indicated by a superscript in parentheses:

$$u^{(r)}_{i,jk\ldots l} = \frac{\partial^r u_i}{\partial x_j \partial x_k \ldots \partial x_l}.$$

4. A sequence will be denoted by its general term between braces: $\{U_n\}$.

1. INTRODUCTION

THE finite element method is nowadays the most general and one of the most powerful tools for the analysis of structures.

Although it was developed for structural analysis it is really a general mathematical technique, and the main concern of this paper is to present it in this light. Mikhlin's book [1] was used as a basis for such purpose.

The important problem of the convergence to the exact solution of a sequence of approximate solutions generated by elements with decreasing size will be given special attention. Experience seems indeed to indicate that the best control of the approximation error consists in examining the behaviour of a sequence of that kind. It has also been observed that no reasonable approximate solutions are likely to be generated if the type of finite element used is such that convergence to the exact solution is not obtainable.

Before convergence to the exact solution was given the attention it deserves, there was a tendency to make monotonic convergence play the fundamental role. Monotonic convergence is nowadays no more considered so important. It has indeed been demonstrated [2] that conformity (a condition for monotonic convergence) does not always speed up the convergence to the exact solution, i.e. less approximate solutions have been obtained for some problems when the monotonic convergence requirements were verified than when they were not.

The capacity for convergence to the exact solution of some kinds of elements has been already examined in the case where continuity is preserved [3]. The author himself presented a first proof [4] of the known criteria [2, 5] which is also valid for cases where continuity is violated.

If continuity is not violated, the finite element method becomes a particularization of the classical Ritz method. This connection with the Ritz method has been observed very often but very seldom studied in detail and explored.

It is very important to notice however that, if continuity is violated, the finite element method is not a simple application of the Ritz method. A section of this paper will be devoted to demonstrating that convergence to the exact solution is still possible even in those cases.

2. STATEMENT OF THE PROBLEM

Let A be a linear bounded operator defined for a dense linear subset M of a real Hilbert space H. Assume the operator A to be symmetric and positive definite [1].

Let (u, v) denote the scalar product of two elements of H. Let $\|u\|$ denote the norm of an element in H.

This paper is concerned with the solution of the equation

$$Au = f \tag{1}$$

that is, in the determination of the element u which the operator A transforms into f; u and f are elements of H.

Equality (1) is meaningful if element u belongs to M. It is possible however that no element of M can correspond to an arbitrary element f of H; this is what is meant by stating that equation (1) can have no solution in M.

It can be shown [1] that, if equation (1) has a solution, this will be unique. It can also be demonstrated that the solution of equation (1) minimizes the functional

$$F(u) = (Au, u) - 2(u, f) \tag{2}$$

and conversely, that the element which minimizes F in M satisfies equation (1).

If A is positive-bounded-below, and not merely positive definite, that is, if

$$(Au, u) \geq \gamma^2(u, u) \tag{3}$$

γ being a real constant, then the field of definition of A can be extended so that equation (1) has a solution for an arbitrary element f of H.

The extended field of definition belongs to a new Hilbert space, H_A, which is a dense subset of H, defined as the completion of the Hilbert space which results from associating

with the elements of M the scalar product

$$[u, v] = (Au, v). \tag{4}$$

This scalar product, which will be denoted by square brackets, is called the *energy product*. The norm in H_A is termed the *energy norm* and will be denoted by bold vertical rules:

$$|u| = \sqrt{[u, u]}. \tag{5}$$

The energy norm of the difference of two elements is the *distance* between both:

$$d(u, v) = |u - v|. \tag{6}$$

The square of the energy norm is termed *energy* [1].

If u_0 is the solution to equation (1), then

$$Au_0 = f. \tag{7}$$

Functional (2) can thus be expressed as

$$F(u) = [u, u] - 2[u, u_0]. \tag{8}$$

It can further be transformed into

$$F(u) = [(u - u_0), (u - u_0)] - [u_0, u_0] = |u - u_0|^2 - |u_0|^2. \tag{9}$$

Expression (9) makes it clear that the minimum value of F in H_A is obtained for $u = u_0$.

A sequence of elements $\{u_{an}\}$, belonging to the field of definition of a functional F, is termed minimizing [1] for F if

$$\lim_{n \to \infty} F(u_{an}) = F_0 \tag{10}$$

F_0 being the exact lower bound of F.

As

$$F_0 = F(u_0) = -|u_0|^2 \tag{11}$$

equation (10) implies

$$\lim_{n \to \infty} |u_{an} - u_0| = 0. \tag{12}$$

Equation (12) means that any sequence which is minimizing for F converges in energy to the exact solution. Energy convergence is characterized by the fact that the distance between each term of the sequence and its limit tends to zero [1].

3. PARTICULARIZATION TO VECTOR FIELDS

Let Ω be an open, connected and bounded domain with a finite number of dimensions. Let S be its boundary, which is supposed to be composed by a finite number of closed, smooth or piecewise smooth stretches.

Let $\bar{\Omega}$ be the closed domain resulting from the combination of Ω and S.

Take for space H the space of the real vector fields (with a fixed number of components) whose moduli are quadratically summable over $\bar{\Omega}$. The scalar product of a pair of elements,

u and v, will be given by the Lebesgue integral:

$$(u, v) = \int_{\Omega} \mathbf{u}^T \mathbf{v} \, d\Omega + \int_{S} \mathbf{u}^T \mathbf{v} \, dS = \int_{\bar{\Omega}} \mathbf{u}^T \mathbf{v} \, d\bar{\Omega} \tag{13}$$

$\mathbf{u}$ and $\mathbf{v}$ being column vectors containing the components of u and v.

The number of components of the vectors is independent of the number of dimensions of the domain.

Equation (1) can be written in a more explicit form:

$$\mathbf{Au} = \mathbf{f} \tag{14}$$

$\mathbf{A}$ being a matrix of operators. These operators are from now on assumed to be differential.

The fields belonging to M are not supposed to satisfy all the boundary conditions of the problem. Those which are necessarily satisfied by every field in M and by each field in H_A, are termed *principal boundary conditions.* The remaining ones are called *natural boundary conditions.*

Any field belonging to M is supposed to meet homogeneous principal boundary conditions. Besides, both the field and the derivatives involved in A must be continuous. These derivatives will not however generally be continuous for every field in H_A.

The energy product between elements belonging to M can be computed by the use of equations (13) and (4):

$$[u, v] = \int_{\bar{\Omega}} (\mathbf{Au})^T \mathbf{v} \, d\bar{\Omega} \tag{15}$$

An energy product involving elements in H_A not belonging to M can be computed as the limit of the energy product of a sequence of pairs of elements belonging to M.

It is assumed that the expression (15) for the energy product can be transformed, by suitable partial integration, into

$$[u, v] = \int_{\Omega} (\mathbf{Ru})^T \mathbf{L}(\mathbf{Rv}) \, d\Omega. \tag{16}$$

$\mathbf{L}$ is a square, symmetric and definite positive matrix, $\mathbf{R}$ a differential operator.

The energy of any element u in H_A is given by

$$[u, u] = \int_{\Omega} (\mathbf{Ru})^T \mathbf{L}(\mathbf{Ru}) \, d\Omega. \tag{17}$$

The expression under the integral sign receives the name of *energy density.*

Assume that $\mathbf{Ru}$ involves derivatives of component u_i with order not greater than p_i. The derivatives of order $(p_i - 1)$ or less are termed *principal derivatives.*

As the energy $[u, u]$ of any field u belonging to H_A must be finite, $(\mathbf{Ru})$ has to be bounded almost everywhere in Ω for every field in H_A. The field components and their principal derivatives must thus be continuous almost everywhere in Ω.

In what follows, f will be supposed such that the exact solution falls into the subset $C_0 \subset H_A$ of the fields whose components are continuous everywhere in $\bar{\Omega}$, together with their principal derivatives. These continuity properties will be referred to in the text as *principal continuity conditions.*

4. APPLICATION TO LINEAR THEORY OF ELASTICITY

Elastic theories involve three kinds of magnitudes: *stresses*, *strains* and *displacements*, whose vectors will be denoted by $\boldsymbol{\sigma}$, $\boldsymbol{\varepsilon}$ and $\mathbf{u}$.

These magnitudes are related by three kinds of field equations which can be symbolized as follows:

(a) *Equilibrium equations:*

$$\mathbf{E}\boldsymbol{\sigma} = \mathbf{X}. \tag{18}$$

(b) *Strain–displacement relations:*

$$\mathbf{D}\mathbf{u} = \boldsymbol{\varepsilon}. \tag{19}$$

(c) *Stress–strain relations:*

$$\boldsymbol{\sigma} = \mathbf{H}\boldsymbol{\varepsilon}. \tag{20}$$

$\mathbf{E}$ and $\mathbf{D}$ are differential operators, $\mathbf{X}$ is the vector of the body force density components, $\mathbf{H}$ is a symmetric positive definitive matrix.

Equations (18), (19) and (20) are valid on Ω. On the boundary S, the equilibrium equations become:

$$\mathbf{N}\boldsymbol{\sigma} = \mathbf{p}. \tag{21}$$

$\mathbf{N}$ is a matrix whose elements depend on the orientation of the normal vector at a given boundary point. $\mathbf{p}$ is the vector of the tractions applied to the boundary.

The analysis of the equilibrium of elastic bodies reduces to finding the solution of the system of field equations (18), (19) and (20) which satisfies certain boundary conditions. The simplest and most important types of boundary conditions can be expressed directly in terms of displacements or tractions applied to the boundary. Let S_1 and S_2 denote the portions of the boundary where tractions or displacements are respectively prescribed.

Operators $\mathbf{E}$ and $\mathbf{D}$ and matrix $\mathbf{N}$ are such that the following relation holds if $\mathbf{u}$ is continuous:

$$\int_\Omega \boldsymbol{\sigma}^T(\mathbf{D}\mathbf{u})\,\mathrm{d}\Omega = \int_\Omega (\mathbf{E}\boldsymbol{\sigma})^T\mathbf{u}\,\mathrm{d}\Omega + \int_S (\mathbf{N}\boldsymbol{\sigma})^T\mathbf{u}\,\mathrm{d}S. \tag{22}$$

In this relation, vectors $\boldsymbol{\sigma}$ and $\mathbf{u}$ are not necessarily related by the stress–strain relations (20).

Combining equations (18), (19) and (20), we obtain:

$$\mathbf{EHDu} = \mathbf{X}. \tag{23}$$

Combination of equations (19), (20) and (21) yields:

$$\mathbf{NHDu} = \mathbf{p}. \tag{24}$$

Comparing equations (23) and (24) with equation (14), there results:

$$\left.\begin{aligned}\mathbf{A} &= \mathbf{EHD}\\ \mathbf{f} &= \mathbf{X}\end{aligned}\right\}\text{for points in } \Omega \tag{25}$$

$$\left.\begin{aligned}\mathbf{A} &= \mathbf{NHD}\\ \mathbf{f} &= \mathbf{p}\end{aligned}\right\}\text{for points on } S. \tag{26}$$

It can easily be shown that operator A defined by equations (25) and (26) has the properties which were indicated in Section 2, if the displacement boundary conditions are enough to eliminate rigid body motion.

Using equation (15) we obtain:

$$[u, v] = \int_\Omega (\mathbf{EHDu})^T \mathbf{v}\, d\Omega + \int_S (\mathbf{NHDu})^T \mathbf{v}\, dS. \tag{27}$$

Equation (22) allows the transformation of (27) into:

$$[u, v] = \int_\Omega (\mathbf{Du})^T \mathbf{H}(\mathbf{Dv})\, d\Omega. \tag{28}$$

Operator **R** coincides thus with **D** and matrix **L** with **H**.*

Functional F becomes:

$$F = \int_\Omega \boldsymbol{\varepsilon}^T \mathbf{H}\boldsymbol{\varepsilon}\, d\Omega - 2\int_\Omega \mathbf{X}^T \mathbf{u}\, d\Omega - 2\int_S \mathbf{p}^T \mathbf{u}\, dS \tag{29}$$

i.e. twice the total potential energy, if the displacement boundary conditions are supposed to be homogeneous.

The theorem of the minimum total potential energy, which states that the exact solution is the one, from all the compatible elastic fields, which makes the total potential energy a minimum, is thus a particularization of the theorem which affirms that the solution of equation (1) makes F a minimum in the space of the fields with finite energy.

The formulation which has been presented is quite general as it is valid not only for linear two and three-dimensional elasticity but also for linear theories of plates, shells and beams.

In the case of a plate, for instance, vector **u** contains the transverse displacement and two rotations, vector $\boldsymbol{\varepsilon}$ contains the curvatures and the transverse shear strains, vector $\boldsymbol{\sigma}$ contains the bending and twisting moments and the transverse shearing forces.

Operator **D** involves derivatives of the first order. The principal derivatives are thus of order zero. This means that the elements of C_0 are elastic fields with displacement components continuous everywhere in $\bar{\Omega}$.

The principal boundary conditions, which are supposed to be homogeneous, are those involving linear combinations of the displacement components. The natural boundary conditions are expressed in terms of stresses.

A very frequent simplification in the analysis of plates, shells and beams consists in neglecting the transverse shear deformation.

This makes it possible to reduce the number of the unknowns to one (the normal displacement) in the theory of plates, and to three (the normal displacement and the tangential displacements) in the theory of thin shells.

* The equation

$$\int_\Omega (\mathbf{Ru})^T \mathbf{L}(\mathbf{Rv})\, d\Omega = \int_\Omega (\mathbf{Au})^T \mathbf{v}\, d\Omega + \int_S (\mathbf{Au})^T \mathbf{v}\, dS$$

which results from the combination of (15) and (16), performs thus in the general theory of the role of the work equation (22).

The simplified theories represent by themselves also a particularization of the general problem formulated in Section 2. The only field components are now the independent unknowns [1].

The rotations become in the simplified theory first derivatives of the normal displacement. The corresponding energy density involves thus first derivatives of the tangential displacements and second derivatives of the normal displacement, so that the principal derivatives are the derivatives of first order of the normal displacement and the derivatives of order zero of the tangential displacements.

The energy will be finite if the normal displacement and its first derivatives, as well as the tangential displacements, are continuous everywhere in $\bar{\Omega}$. As the first derivatives of the normal displacement are the rotations, the elements of C_0 are still the elastic fields with all the displacement components continuous everywhere in $\bar{\Omega}$.

Similar conclusions could easily be derived for beams.

The principal continuity conditions are thus the same both in the simplified theories and in the corresponding theories where the transverse shear deformation is not neglected.

5. EQUIVALENT PROBLEM

Consider the domain Ω subdivided into a number of subdomains, $\Omega^1, \Omega^2, \Omega^3, \ldots, \Omega^e, \ldots$ and let H^e be a real functional Hilbert space whose elements are fields defined on the general closed subdomain $\bar{\Omega}^e$. The scalar product between any pair of elements, u^e and v^e, belonging to H^e is defined by

$$(u^e, v^e) = \int_{\Omega^e} \mathbf{u}^{eT}\mathbf{v}^e \, d\Omega + \int_{S^e} \mathbf{u}^{eT}\mathbf{v}^e \, dS = \int_{\bar{\Omega}^e} \mathbf{u}^{eT}\mathbf{v}^e \, d\bar{\Omega}. \tag{30}$$

Let H_n be another Hilbert space (index n refers to a certain degree of subdivision of Ω into subdomains). Each element $u_n \in H_n$ may be regarded as a piecewise defined field. It represents, however, not truly a single field defined on $\bar{\Omega}$, but a set of fields u^e (one per subdomain), belonging to the different spaces H^e. Such fields are called subfields of u_n.

The scalar product in H_n is defined by

$$(u_n, v_n)_n = \sum_e (u^e, v^e)^e \tag{31}$$

where u^e and v^e are the subfields of u_n and v_n, and $\sum_e$ denotes a summation over the whole set of subdomains.

Let M^e be a dense linear subset of H^e. Every field in M^e is assumed to coincide, on the closed subdomain $\bar{\Omega}^e$, with an arbitrary field of M.

Let M^e be the field of definition of a linear, bounded and symmetric differential operator, defined to be such that, within Ω^e,

$$\mathbf{A}^e\mathbf{u}^e = \mathbf{A}\mathbf{u}^e \tag{32}$$

and, on S^e,

$$\int_{S^e} (\mathbf{A}^e\mathbf{u}^e)^T\mathbf{v}^e \, d\Omega = \int_{\Omega^e} (\mathbf{R}\mathbf{u}^e)^T\mathbf{L}(\mathbf{R}\mathbf{v}^e) \, d\Omega - \int_{\Omega^e} (\mathbf{A}^e\mathbf{u}^e)^T\mathbf{v}^e \, d\Omega \tag{33}$$

u^e and v^e being arbitrary fields belonging to M^e.

Make

$$[u^e, v^e]^e = \int_{\Omega^e} (\mathbf{R}\mathbf{u}^e)^T \mathbf{L} (\mathbf{R}\mathbf{v}^e)\, d\Omega = \int_{\bar{\Omega}^e} (\mathbf{A}^e \mathbf{u}^e)^T \mathbf{v}^e\, d\bar{\Omega} = (A^e u^e, v^e)^e. \tag{34}$$

Call M_n the dense linear subset of H_n whose elements have subfields belonging to the sets M^e, and consider a linear subset of M_n such that, given any pair of its elements, u_n and v_n, with subfields u^e and v^e, the magnitude

$$[u_n, v_n]_n = \sum_e [u^e, v^e]^e \tag{35}$$

can be properly chosen as their scalar product. Let H_{A^e} be the completion of the Hilbert space obtained by associating the scalar product (35) (termed energy product) to such subset.

The distance between any two elements in H_{A^e} will be given by

$$d_n(u_n, v_n) = \sqrt{[(u_n - v_n), (u_n - v_n)]_n} = |u_n - v_n|_n. \tag{36}$$

Consider now the linear subspace of H_n, H'_n, whose elements are such that their subfields u'^e, corresponding to adjacent subdomains, take equal values at points lying on the common interface.

Let H_n and H'_n be the field of definition and the range of an operator B_n such that, if

$$B_n u_n = u'_n \tag{37}$$

the sum of the values taken on the interface of two adjacent subdomains by the corresponding subfields of u_n is equal to the sum of the same values respecting u'_n. The effect of the operator B_n is thus to distribute that sum equally between the subdomains in contact.

Let v'_n belong to H'_n and let v'^e denote its subfields. We can write

$$\begin{aligned}(u_n, v'_n)_n = \sum_e (u^e, v'^e)^e = \sum_e \int_{\bar{\Omega}^e} \mathbf{u}^{eT} \mathbf{v}'^e\, d\bar{\Omega} = \sum_e \int_{\bar{\Omega}^e} \mathbf{u}'^{eT} \mathbf{v}'^e\, d\bar{\Omega} \\ = \sum_e (u'^e, v'^e)^e = (u'_n, v'_n)_n = (B_n u_n, v'_n)_n.\end{aligned} \tag{38}$$

Let M_n be the field of definition of a new operator, A_n, such that

$$A_n u_n = B_n f_n \tag{39}$$

in which f_n is an element of H_n with subfields $f^e = A^e u^e$.

By virtue of equations (38) and (39), we have

$$(A_n u_n, v'_n)_n = (B_n f_n, v'_n)_n = (f_n, v'_n)_n = \sum_e (A^e u^e, v'^e)^e = \sum_e [u^e, v'^e]^e = [u_n, v'_n]_n. \tag{40}$$

Consider now the subset of H_n whose elements fulfill the condition that all the corresponding subfields coincide, within their respective subdomains, with a given field u belonging to H_A (the same for all the subdomains). It is clear that such subset is contained in H'_n. Call H_{A_n} the Hilbert space obtained by associating the scalar product $[u_n, v_n]_n$ to

such subset. As

$$[u_n, v_n]_n = \sum_e [u^e, v^e]^e = \sum_e \int_{\Omega^e} (\mathbf{R}\mathbf{u}^e)^T \mathbf{L}(\mathbf{R}\mathbf{v}^e)\, d\Omega$$
$$= \int_\Omega (\mathbf{R}\mathbf{u})^T \mathbf{L}(\mathbf{R}\mathbf{v})\, d\Omega = [u, v] \tag{41}$$

$[u_n, v_n]_n$ is a proper scalar product. No contradiction is thus introduced if H_{A_n} is assumed to be contained in H_{A^e}.

Call C_{0n} the subset of H_{A_n} corresponding to the subset C_0 of H_A.

Consider the equation

$$A_n u_n = f'_n \tag{42}$$

The solution of equation (42) in M_n is generally not unique and the operator A_n not positive definite. If, however, the field of definition of A_n is restricted to H_{A_n}, then the operator A_n becomes positive definite. Indeed, as H_{A_n} is contained in H'_n, (40) permits to write

$$(A_n u_n, u_n)_n = [u_n, u_n]_n > 0. \tag{43}$$

Consider the functional

$$F_n(u_n) = [u_n, u_n]_n - 2(u_n, f'_n)_n. \tag{44}$$

Let u_{0n} denote the solution of (42) in H_{A_n}. Then

$$A_n u_{0n} = f'_n. \tag{45}$$

If u_n belongs to H'_n, then

$$(u_n, f'_n)_n = (u_n, A_n u_{0n})_n = [u_n, u_{0n}]_n. \tag{46}$$

Introducing (46) in (44), there results

$$F_n(u_n) = [u_n, u_n]_n - 2[u_n, u_{0n}]_n = [(u_n - u_{0n}), (u_n - u_{0n})] - [u_{0n}, u_{0n}]. \tag{47}$$

Expression (47) makes it clear that u_{0n} minimizes F_n and that any element which minimizes F_n in H_{A_n} must coincide with u_{0n}. The solution of (42) in H_{A_n} is thus unique.

Assume now that f'_n vanishes on the subdomain interfaces, and let f be any field in H which takes the same values, within the subdomains Ω^e, as the subfields of f'_n. Let u_n be the element in H_{A_n} which corresponds to the element u of H_A. Then

$$(u_n, f'_n)_n = \sum_e (u^e, f'^e)^e = \sum_e \int_{\bar{\Omega}^e} \mathbf{u}^{eT} \mathbf{f}^e\, d\bar{\Omega}$$
$$= \sum_e \int_{\Omega^e} \mathbf{u}^T \mathbf{f}\, d\Omega + \int_S \mathbf{u}^T \mathbf{f}\, dS = (u, f). \tag{48}$$

By virtue of (41) and (48),

$$F_n(u_n) = [u, u] - 2(u, f) = F(u). \tag{49}$$

Let u'_0 be the field of H_A which corresponds to u_{0n}. By virtue of (9),

$$F(u'_0) = [(u'_0 - u_0), (u'_0 - u_0)] - [u_0, u_0]. \tag{50}$$

As $F_n(u_n) \equiv F(u)$ and u_{0n} minimizes $F_n(u_n)$ in H_{A_n}, u'_0 minimizes $F(u)$ in H_A. But equation (50) shows that F can only be minimized by u'_0 if u'_0 coincides with u_0. Thus the solution of equation (42) in H_{A_n} coincides, within each subdomain, with the solution of equation (1) in H_A.

This means that the problem of the solution of equation (42) is equivalent to the problem of the solution of equation (1).

The concepts introduced along this Section represent a generalization of the concepts introduced in the preceding ones. It is convenient to interpret such generalization in terms of three-dimensional Elasticity.

Operator A^e has the same meaning for subdomain Ω^e as operator A for the global domain, Ω. While A is defined by expressions (25) and (26), respectively for points located within Ω and on S, operator A^e is defined by the same expressions for points located within Ω^e and on S^e.

f'_n represents an external force distribution acting on each subdomain of the body. Assuming that f'_n belongs to H'_n is equivalent to assume that the total force acting on the interface between two adjacent subdomains is equally distributed between both.

The value taken by f'_n on the subdomain boundaries are thus half the surface density values of the forces acting on those interfaces. These values must vanish, if the body force volume density is to be bounded everywhere in Ω.

To solve equation (42) means to determine an elastic field which verifies equation (23) within each subdomain and equation (24) on S_1 and whose stresses present, on the subdomain interfaces, the discontinuities required to equilibrate the external forces applied on such interfaces. Any solution of equation (42) equilibrates thus the external force distribution symbolized by f'_n.

The solution of equation (42) becomes also compatible if it belongs to H_{A_n}, because the displacement boundary conditions are then respected on S_2 and the continuity of the displacements is preserved across the subdomain boundaries. The corresponding subfields coincide thus, within each subdomain, with the solution of equation (1).

6. THE FINITE ELEMENT METHOD

The finite element method is a general technique of numerical analysis which provides an approximate solution for equation (1).

In this method, domain Ω is considered to be decomposed into a finite number of subdomains and families of fields are considered which have different analytical expressions inside each subdomain.

A finite element is a closed subdomain, $\bar{\Omega}^e$, together with the family of fields which are allowed to occur within it. This family is a linear combination with coefficients q_i^e of a finite number of unit modes, so that each field of the family corresponds to ascribing particular values to the parameters q_i^e.

The values of the field components and its principal derivatives, at a certain number of points on the boundary of the elements, called *nodes* or *nodal points*, are as a rule chosen as parameters.

The type of an element refers to its general shape, nodal point specification and to the allowed fields, analytically defined by expressing a general field u^e in terms of the parameters

and the coordinates with respect to a given frame:

$$\mathbf{u}^e = \boldsymbol{\varphi}^e(x_1, x_2, \ldots)\mathbf{q}^e \tag{51}$$

$\mathbf{q}^e$ is the vector of the parameters.*

Elements $\boldsymbol{\varphi}^e_{ij}$ of matrix $\boldsymbol{\varphi}^e$ are supposed to be continuous and have continuous derivatives of order $(p_i - 1)$, or less, in the closed domain $\bar{\Omega}^e$ occupied by the finite element e. The unit modes are defined by the columns of $\boldsymbol{\varphi}^e$.

We suppose that each field component depends only on its own values at the nodes and on the values taken by its principal derivatives also at the nodes. Thus, if q^e_j corresponds to the field component u^e_i or one of its derivatives at any node of the element, all the magnitudes φ^e_{kj} for which $k \neq i$ will be equal to zero.

If q^e_j corresponds to a derivative of order s of u^e_i, φ^e_{ij} will take the form

$$\varphi^e_{ij}(x_1, x_2, \ldots, l^e) = (l^e)^s \psi^e_{ij}\left(\frac{x_1}{l^e}, \frac{x_2}{l^e}, \ldots\right) \tag{52}$$

in which l^e is a typical dimension of the element, for instance its maximum diameter, and ψ^e_{ij} is a function which does not depend on the absolute dimensions of the element. This is necessary in order that equation (51) can be homogeneous.

The different finite elements are compatibilized through the specification of *reduced continuity conditions.* These require that the values of the field components and their principal derivatives be the same at coincident nodes of adjacent elements and equal the prescribed ones at the nodes located on S_2, the portion of S where the principal boundary conditions are specified.

A point of the domain is said to be a *node of the system* if it is a node for one or more elements.

Let $\mathbf{q}_n$ be the vector of the field components and their principal derivatives at every node of the system but those which are located on S_2. The reduced continuity conditions can be expressed by writing for each element

$$\mathbf{q}^e = \mathbf{T}^e\mathbf{q}_n \tag{53}$$

where matrix $\mathbf{T}^e$ depends on the topology of the system.

Equations (53) show that the knowledge of $\mathbf{q}_n$ is enough for the definition of the field within every element of the system.

The reduced continuity conditions are generally not sufficient to make the field components and their principal derivatives continuous across the element boundaries. This depends on the type of the element.

If the type is such that the reduced continuity conditions are sufficient to ensure continuity of the components and their principal derivatives across the element boundaries, the piecewise defined fields generated by the system of finite elements are said to be *conforming.* Every conforming field thus belongs to C_{0n}, i.e. to the subset of H_{A_n} corresponding to C_0.

If the continuity requirements are violated across the element boundaries, the fields are said to be *non-conforming.*

Let U_n be the subset of H_n containing the elements whose subfields are defined by equation (51) and compatibilized through the reduced continuity conditions. Any element

* The allowed fields need not be introduced by giving the expression of the field components directly in terms of their own nodal values and the nodal values of their principal derivatives. They can indeed be given in terms of equal number of arbitrary parameters which in turn can be expressed in terms of those nodal values (see [9]).

$u_n \in U_n$ can be expressed, within Ω^e, by

$$\mathbf{u}^e = \mathbf{\Phi}^e \mathbf{q}_n \tag{54}$$

where

$$\mathbf{\Phi}^e = \boldsymbol{\varphi}^e \mathbf{T}^e. \tag{55}$$

Take for space H_{A^e} (see Section 5) the space spanned by U_n and H_{A_n}. The distance between any pair of elements belonging to H_{A^e} is defined by expression (36). The discussion of completeness and convergence will be based on that concept of distance. The distance between any element u_n in U_n and any element u in H_A will indeed be measured by the distance between u_n and the element in H_{A_n} corresponding to u.

The approximate solution, u_{an}, which the finite element method provides for equation (42), and thus for equation (1), is determined by making the functional F_n stationary in U_n. Such solution could be the exact one if u_{0n} was contained in U_n. As, generally, it is not, the solution yielded by the finite element method is only approximate.

Introducing (51) and (35) in (44), we obtain:

$$F_n = \sum_e [\mathbf{q}^{e^T} \mathbf{K}^e \mathbf{q}^e - 2\mathbf{q}^{e^T} \mathbf{Q}^e] \tag{56}$$

where

$$\mathbf{K}^e = \int_{\Omega^e} (\mathbf{R}\boldsymbol{\varphi}^e)^T \mathbf{L} (\mathbf{R}\boldsymbol{\varphi}^e)\, \mathrm{d}\Omega \tag{57}$$

$$\mathbf{Q}^e = \int_{\bar{\Omega}^e} \boldsymbol{\varphi}^{e^T} \mathbf{f}'^e\, \mathrm{d}\bar{\Omega}. \tag{58}$$

Introducing now (53) in (56), we obtain

$$F_n = \mathbf{q}_n^T (\sum_e \mathbf{T}^{e^T} \mathbf{K}^e \mathbf{T}^e) \mathbf{q}_n - 2\mathbf{q}_n^T (\sum_e \mathbf{T}^{e^T} \mathbf{Q}^e). \tag{59}$$

Making

$$\mathbf{K}_n = \sum_e \mathbf{T}^{e^T} \mathbf{K}^e \mathbf{T}^e \tag{60}$$

$$\mathbf{Q}_n = \sum_e \mathbf{T}^{e^T} \mathbf{Q}^e \tag{61}$$

there results

$$F_n = \mathbf{q}_n^T \mathbf{K}_n \mathbf{q}_n - 2\mathbf{q}_n^T \mathbf{Q}_n. \tag{62}$$

The stationary conditions for F_n are obtained by equating to zero the derivatives of F_n with respect to the mutually independent parameters $\mathbf{q}_{ni}$. It results in the system of linear equations

$$\mathbf{K}_n \mathbf{q}_n = \mathbf{Q}_n. \tag{63}$$

Introducing (57) in (60) and using (55), we obtain

$$\mathbf{K}_n = \sum_e \int_{\Omega^e} (\mathbf{R}\mathbf{\Phi}^e)^T \mathbf{L} (\mathbf{R}\mathbf{\Phi}^e)\, \mathrm{d}\Omega \tag{64}$$

$$\mathbf{Q}_n = \sum_e \int_{\bar{\Omega}^e} \mathbf{\Phi}^{e^T} \mathbf{f}'^e\, \mathrm{d}\bar{\Omega}. \tag{65}$$

Matrix $\mathbf{K}_n$ is non-singular whenever the columns of $\mathbf{\Phi}^e$ are linearly independent. As $\mathbf{L}$ is definite positive, $\mathbf{K}_n$ is also definite positive.

If $\mathbf{K}_n$ is non-singular, the parameters q_{ni} can be uniquely determined by solving the system of equations (63). Let $\mathbf{q}_{0n}$ be the vector of the parameters which verify equation (63). Functional F_n can be expressed as

$$F_n = \mathbf{q}_n^T\mathbf{K}_n\mathbf{q}_n - 2\mathbf{q}_n^T\mathbf{K}_n\mathbf{q}_{0n} = (\mathbf{q}_n - \mathbf{q}_{0n})^T\mathbf{K}_n(\mathbf{q}_n - \mathbf{q}_{0n}) - \mathbf{q}_{0n}^T\mathbf{K}_n\mathbf{q}_{0n}. \tag{66}$$

As $\mathbf{K}_n$ is definite positive, the first term in the right-hand side of (66) is positive unless $\mathbf{q}_n$ equals $\mathbf{q}_{0n}$. This proves that the solution of (63) minimizes F_n in U_n.

Let now u_{1n} and u_{2n} be two elements belonging to U_n. Let $\mathbf{q}_{1n}$ and $\mathbf{q}_{2n}$, be the vectors of the corresponding parameters. The energy product of u_{1n} and u_{2n} can be given by

$$\begin{aligned}
[u_{1n}, u_{2n}]_n &= \sum_e \int_{\Omega^e} (\mathbf{R}\mathbf{u}_1^e)^T\mathbf{L}(\mathbf{R}\mathbf{u}_2^e)\,\mathrm{d}\Omega = \mathbf{q}_{1n}^T \sum_e \int_{\Omega^e} (\mathbf{R}\mathbf{\Phi}^e)^T\mathbf{L}(\mathbf{R}\mathbf{\Phi}^e)\,\mathrm{d}\Omega\,\mathbf{q}_{2n} \\
&= \mathbf{q}_{1n}^T\mathbf{K}_n\mathbf{q}_{2n} = \mathbf{q}_{1n}^T\mathbf{Q}_{2n} = \mathbf{q}_{1n}^T \sum_e \int_{\bar{\Omega}^e} \mathbf{\Phi}^{eT}\mathbf{f}_2^e\,\mathrm{d}\bar{\Omega} \\
&= \sum_e \int_{\bar{\Omega}^e} (\mathbf{\Phi}^e\mathbf{q}_{1n})^T\mathbf{f}_2^e\,\mathrm{d}\bar{\Omega} = \sum_e \int_{\bar{\Omega}^e} \mathbf{u}_1^{eT}\mathbf{f}_2^e\,\mathrm{d}\Omega \\
&= \sum_e (u_1^e, f_2'^e)^e = (u_{1n}, f_{2n}')_n = (u_{2n}, f_{1n}')_n
\end{aligned} \tag{67}$$

where f_{1n}' and f_{2n}' are the right hand sides (of equation (42)) which u_{1n} and u_{2n} correspond to (as approximate solutions).

It results from (67) that the functional F_n may take in U_n the following expression

$$\begin{aligned}
F_n(u_n) &= [u_n, u_n]_n - 2(u_n, f_n')_n = [u_n, u_n]_n - 2[u_n, u_{an}]_n \\
&= [(u_n - u_{an}), (u_n - u_{an})]_n - [u_{an}, u_{an}]_n
\end{aligned} \tag{68}$$

which makes it clear that u_{an} minimizes F_n in U_n. Such expression will be used in Section 10.

7. THE RITZ METHOD

The method just described is justified if it can generate a sequence of fields converging to u_0 (the solution of equation (1)), when successive subdivisions are considered with elements of invariant type but decreasing size.

Conditions to be met by matrix $\mathbf{\varphi}^e$ in order that this convergence may be ensured can be established if it is remarked that the finite element method is related to the well-known Ritz method [1, 6].

The Ritz method is a technique for generating a minimizing sequence for a given functional, say F. This technique, which can be used whenever H is a separable space, is based [6] on the determination of a sequence of families, $\{V_n\}$, satisfying the following conditions:

(a) the sequence is complete in energy with respect to a class $C \subset H_A$ containing u_0 (*completeness requirement*);

(b) the nth family depends on a finite number, N, of arbitrary parameters;

(c) every element which can be obtained by ascribing arbitrary values to the parameters belongs to C_0 (*conformity requirement*).

In what concerns condition (a), it is remembered that a sequence of families of elements is said to be *complete in energy* with respect to a given class $C \subset H_A$ if it is possible, for a specified $\varepsilon > 0$, to find an integer N such that, in each family with order $n > N$, there exists an element u_{cn} which satisfies the inequality.

$$d(u, u_{cn}) < \varepsilon \tag{69}$$

where u is any element of C.

The terms of the minimizing sequence $\{V_{an}\}$ are obtained by minimizing F in each family V_n.

The elements of the nth family are generally given as a linear combination with coefficients q_{ni} of N linearly independent fixed elements Ψ_{ni} which are termed coordinate elements:

$$\mathbf{u}_n = \sum_{i=1}^{N} \Psi_{n_i} q_{n_i} = \Psi_n \mathbf{q}_n \tag{70}$$

$\mathbf{q}_n$ being the vector of the coefficients q_{ni} and Ψ_n the matrix with columns Ψ_{ni}.

Family V_n becomes thus a linear N-dimensional space. Introducing (70) in (2), we obtain:

$$F(u_n) = \mathbf{q}_n^T \mathbf{K}_n \mathbf{q}_n - 2\mathbf{Q}_n^T \mathbf{q}_n \tag{71}$$

in which

$$\mathbf{K}_n = \int_\Omega (\mathbf{R}\Psi_n)^T \mathbf{L} (\mathbf{R}\Psi_n) \, d\Omega \tag{72}$$

and vector $\mathbf{Q}_n$ is defined by

$$\mathbf{Q}_n = \int_{\bar{\Omega}} \Psi_n^T \mathbf{f} \, d\bar{\Omega}. \tag{73}$$

The values of the parameters which make F stationary can be determined by solving the system of linear equations

$$\mathbf{K}_n \mathbf{q}_n = \mathbf{Q}_n. \tag{74}$$

$\mathbf{K}_n$ is a non-singular matrix if the coordinate elements are linearly independent [1]. The system has thus a unique solution which provides the unique stationary point of F in V_n.

The finite element method can be considered as a technique for the application of the Ritz method only if the piecewise defined fields are conforming (conformity requirement). Only thus can indeed condition (c) be respected. The sets V_n are then the subsets of H_A corresponding to the subsets $U_n \subset H_{A_n}$.

In order that convergence to the exact solution may be obtained, it is thus only necessary to meet condition (a), that is completeness. It will be seen later on how this can be obtained.

Comparing (70) with (54) it can be concluded that the coordinate fields used in the finite element method are defined by

$$\Psi_n = \varphi^e \mathbf{T}^e = \Phi^e \quad \text{within element } e. \tag{75}$$

The analytical expression of the coordinate fields varies thus from element to element and this piecewise definition is the main characteristic of the finite element method.

It is also important to notice that such piecewise definition and the reduced continuity conditions allow matrix $\mathbf{K}_n$ and vector $\mathbf{Q}_n$ to be assembled from simpler matrices, $\mathbf{K}^e$ and $\mathbf{Q}^e$, connected with the finite elements themselves (see equations (60) and (61)). This is one of the most interesting features of the method.

8. MONOTONIC CONVERGENCE

Assume a sequence of families, $\{V_n\}$, fulfilling the conditions (b) and (c) stated in the preceding section and suppose that the nth family contains all the families with smaller order. As v_{an} makes F a minimum in V_n, we have:

$$F(v_{a1}) \geq F(v_{a2}) \geq F(v_{a3}) \geq \ldots \geq F(v_{an}) \geq \ldots F(u_0). \tag{76}$$

By Bolzano's theorem [7], the sequence $\{F(v_{an})\}$ converges to a limit which cannot be smaller than $F(u_0)$. It is remarked that this conclusion is valid even if condition (a) of Section 7 is not obeyed. If it is obeyed, then we know that the limit is $F(u_0)$.

As the inequality

$$F(v_{an}) - F(v_{am}) \leq 0 \tag{77}$$

holds, for $m < n$, equation (9) yields

$$|v_{an} - u_0| \leq |v_{am} - u_0|. \tag{78}$$

This means that the distance to the exact solution decreases when n increases. Convergence is said to be *monotonic*.

Monotonic convergence does not ensure convergence to the exact solution. On the other hand, convergence to the exact solution is not necessarily monotonic.

Consider now a sequence of approximate solutions generated by finite elements with decreasing size.

Conformity and the requirement that the family of fields corresponding to a given subdivision contains the families corresponding to elements with larger sizes have been proposed by Melosh [8] as sufficient conditions for monotonic convergence of such sequence.

However, as the approximate solution minimizes F_n in U_n, regardless of conformity being respected, the requirement that each family of fields contains the families corresponding to elements with larger sizes stands by itself as a sufficient condition for convergence. We can indeed write the set of inequalities (76) once this condition is fulfilled.

9. COMPLETENESS CRITERION

In what concerns convergence to the exact solution, we know that the Ritz method generates a minimizing sequence and that a minimizing sequence actually converges in energy to the exact solution.

Convergence to the exact solution can thus be ensured if conformity and completeness are both achieved. We shall see however that completeness is the truly important requirement.

Before proceeding further we remark that completeness of a sequence of families with respect to a set $C \subset H_A$ has a meaning provided we can compute the distance between every field of each family and any element in C (see Section 6).

A general criterion for completeness will be stated and justified in this section. This criterion was presented in a recent book [9] by Zienkiewicz but it has not yet been justified as far as we know.

Let (p_i-1) be the maximum order of the principal derivatives for component u_i.

We wish to demonstrate that completeness will be obtained if the general analytical expression for u_i^e, within element e (see equation (51)), is given* as a polynomial with a number of arbitrary coefficients equal to the number of unit modes corresponding to the element. Furthermore this polynomial expression must contain a complete polynomial of the p_ith degree all the terms of which are affected by independent arbitrary coefficients. The terms of higher degree can vanish whatever the values taken by those coefficients.

We remark that, if this is the case, the field component u_i^e or any of its derivatives of order p_i or less can take any arbitrary constant value throughout the element if suitable values are ascribed to the parameters. In order that the derivative $u^e_{i,rs\ldots}$ assumes an arbitrary constant value V in Ω^e, it is then indeed only necessary that the coefficient which multiplies the monomial $(x_1^r x_2^s \ldots)$ in u_i^e be equal to $V/r(r-1)\ldots s(s-1)\ldots$, all the remaining coefficients being equal to zero.

The right hand side of equation (1) has been constrained in Section 3 to be such that solution u_0 belongs to C_0, so that the derivatives of order p_i of solution u_0 are bounded but not necessarily continuous.

In the next Sections we assume furthermore that the exact solution falls into a subset of C_0, C_1, such that the derivatives of order (p_i+1) of the field component u_i are continuous within each element. Discontinuities of the p_ith and (p_i+1)th derivatives are still allowed at points which always remain on element boundaries as the size of the elements is progressively reduced.

Let C_{1n} be the subset of H_{A_n} corresponding to the subset C_1 or H_A.

Any field u_n belonging to C_{1n} can thus be represented inside Ω^e by the following Taylor's expansion of its subfield components:

$$u_i^e = u_i^e(O)+u^e_{i,j}(O)\,.\,(x_j-x_j^0)+\ \ldots\ +\frac{1}{p_i!}u^{e(p_i)}_{i,j\ldots k}(O)\,.\,(x_j-x_j^0)\ldots(x_k-x_k^0)+ \\ +\frac{1}{(p_i+1)!}u^{e(p_i+1)}_{i,jk\ldots l}(O_i)\,.\,(x_j-x_j^0)(x_k-x_k^0)\ldots(x_l-x_l^0) \tag{79}$$

O and O_i are points of Ω^e. O_i depends on the coordinates of the point where u_i^e is to be determined.

Let us consider now a polynomial field with components

$$u_i^{te} = u_i^e(O)+u^e_{i,j}(O)\,.\,(x_j-x_j^0)+\ \ldots\ +\frac{1}{p_i!}u^{e(p_i)}_{i,j\ldots k}(O)\,.\,(x_j-x_j^0)\ldots(x_k-x_k^0) \tag{80}$$

which we call tangent field to u at O.

As all the derivatives of order (p_i+1) are bounded inside Ω^e, (79) and (80) yield:

$$|u_i^e-u_i^{te}| < \frac{d}{(p_i+1)!}\,.\,V_1\,.\,(l^e)^{p_i+1} \tag{81}$$

* See footnote at page 939.

in which V_1 is an upper bound for all the (p_i+1)th derivatives and l^e is the maximum diameter of element e. d is the total number of the (p_i+1)th derivatives.

By considering similar expansions for the derivatives of u_i, it is possible to derive the following general inequality:

$$|u^{e(r)}_{i,j\ldots k} - u^{te(r)}_{ni,j\ldots k}| < \frac{d}{(p_i-r+1)!} \cdot V_1 \cdot (l^e)^{p_i-r+1} \tag{82}$$

for $r \leqq p_i$.

As operator $\mathbf{R}$ involves derivatives of u^e_i of order p_i or less, we have:

$$[\mathbf{R}(\mathbf{u}^e - \mathbf{u}^{te}_n)]^T \mathbf{L} [\mathbf{R}(\mathbf{u}^e - \mathbf{u}^{te}_n)] < V_2 (l^e)^2 \tag{83}$$

for l^e sufficiently small. V_2 is a positive number.

Thus:

$$[(u^e - u^{te}), (u^e - u^{te})]^e < V_2 (l^e)^2 \Omega^e. \tag{84}$$

If tangent fields u^{te} are considered for every subdomain Ω^e, piecewise defining a field u^t_n in Ω, we obtain, by using (35),

$$(|u_n - u^t_n|_n)^2 = [(u - u^t_n), (u - u^t_n)]_n < V_2 l_n^2 \Omega \tag{85}$$

in which l_n denotes the maximum value of l^e in the whole set of elements.

This means that the distance between any field in C_1 and the tangent field u^t_n, piecewise defined by (80), tends to zero with the size of the finite elements.

Consider now a type of finite element generating a sequence of families of fields whose completeness is to be investigated.

Call u^{fe} the field within the finite element e such that the values of its components and their principal derivatives at the nodal points are respectively equal to the values of the components of the field $u_n \in C_{1n}$ and corresponding partial derivatives at the same points.

Suppose the general criterion to be satisfied. u^{te} can thus be one of the fields which can occur within the finite element. Let this field correspond to values q^{te}_i of the parameters:

$$\mathbf{u}^{te} = \boldsymbol{\varphi}^e \mathbf{q}^{te}. \tag{86}$$

On the other hand

$$\mathbf{u}^{fe} = \boldsymbol{\varphi}^e \mathbf{q}^{fe}. \tag{87}$$

From (86) and (87) we obtain

$$|u^{te}_i - u^{fe}_i| = |\varphi^e_{ij}(q^{te}_j - q^{fe}_j)| \tag{88}$$

or, considering (52),

$$|u^{te}_i - u^{fe}_i| = \sum_j (l^e)^s \cdot |\psi^e_{ij}(q^{te}_j - q^{fe}_j)| \tag{89}$$

in which s is the order of the field derivative to which parameter q_j corresponds.

But

$$\psi^e_{ij} = \psi^e_{ij}\left(\frac{x_1}{l^e}, \frac{x_2}{l^e}, \ldots\right) \tag{90}$$

and

$$\psi^{e(r)}_{ij,k\ldots l} = \frac{1}{(l^e)^r} \cdot \frac{\partial^r \psi^e_{ij}}{\partial(x_k/l^e)\ldots\partial(x_l/l^e)}. \tag{91}$$

For this reason,

$$\begin{aligned} |u^{te(r)}_{i,k\ldots l} - u^{fe(r)}_{i,k\ldots l}| &= |\varphi^{e(r)}_{ij,k\ldots l}(q^{te}_j - q^{fe}_j)| \\ &\leq \sum_j (l^e)^{s-r} \left| \frac{\partial^r \psi^e_{ij}}{\partial(x_k/l^e)\ldots\partial(x_l/l^e)}(q^{te}_j - q^{fe}_j) \right|. \end{aligned} \tag{92}$$

As the absolute dimensions of the element do not appear explicitly in the functions ψ^e_{ij}, these functions remain bounded as the size of the element decreases.

The same happens to the derivatives $\partial^r \psi^e_{ij}/\partial(x_k/l^e)\ldots\partial(x_l/l^e)$, for $r \leq p_i$, because the functions φ^e_{ij} and their derivatives of order $(p_i - 1)$ or less were in Section 5 supposed to be continuous. Assume the moduli of all these magnitudes remain below a positive number V_3.

Then

$$|u^{te(r)}_{i,k\ldots l} - u^{fe(r)}_{i,k\ldots l}| < \sum_j (l^e)^{s-r} V_3 \,.\, |q^{te}_j - q^{fe}_j|. \tag{93}$$

On the other hand, as the components of u^{fe} and their principal derivatives take the same values at the nodes as the corresponding magnitudes in u^e, equation (82) permits us to write:

$$|q^{te}_j - q^{fe}_j| < \frac{d}{(p_i - s + 1)!} \,.\, V_1 \,.\, (l^e)^{p_i - s + 1} \tag{94}$$

when parameter q^e_j corresponds to a derivative of order s.

Equations (93) and (94) hold even if the p_ith derivatives of u^{fe}_i are discontinuous in Ω^e. This is an important remark because sometimes [10] the element itself is considered subdivided into parts and the field admits different analytical expressions within each part. Our proof remains valid however even if the derivatives of order p_i are not continuous across the internal boundaries of the element.

As the parameters cannot correspond to derivatives of order larger than $(p_i - 1)$, s cannot be larger than $(p_i - 1)$ and equation (94) yields

$$|q^{te}_j - q^{fe}_j| < \frac{d}{2!} V_1 (l^e)^{p_i - s + 1}. \tag{95}$$

Introducing equations (95) in (93) we obtain:

$$|u^{te(r)}_{i,k\ldots l} - u^{fe(r)}_{i,k\ldots l}| < V_3 \frac{V_1}{2!} (l^e)^{p_i - r + 1} dN^e \tag{96}$$

N^e being the total number of parameters corresponding to element e.

This equation is still valid for $r = 0$, if the derivatives of order zero are interpreted as the field components themselves.

The similarity between equations (96) and (82) allows a jump straight to the inequality:

$$(|u^t_n - u^f_n|_n)^2 < V_4 l^2_n \Omega \tag{97}$$

V_4 being a positive number and u_n^f denoting the piecewise defined field which coincides with u^{fe} within a general element e.

Equation (97) means that the distance between u_n^I and u_n^f tends to zero with l_n. Combining (97) and (85) we conclude that the distance between u_n and u_n^f tends also to zero when the size of the element decreases, so that, as u_n is an arbitrary element of C_{1n}, the completeness proof is finally achieved.

10. CONVERGENCE DISCUSSION

Consider any type of finite element which can generate a sequence $\{U_n\}$ of families of generally non-conforming fields complete in energy with respect to C_1.

We wish to investigate if the sequence of approximate solutions $\{u_{an}\}$ obtained by minimizing F_n in each family U_n converges in energy to the exact solution.

We know already that completeness implies convergence to the exact solution if it is associated with conformity. It will be concluded in this Section that completeness with respect to C_1 is a sufficient condition for convergence, regardless of conformity being obtained.

Let u_{en} be the field in U_n which presents the same nodal values of the field components and corresponding principal derivatives as u_{0n} (the solution of equation (42) in H_{A_n}). As completeness is ensured, it is possible to determine N such that, for $n > N$,

$$d_n(u_{on}, u_{en}) < \varepsilon \tag{98}$$

ε being a positive and arbitrarily small number.

As F_n is continuous, we can find ε such that

$$F_n(u_{en}) = F_n(u_{0n}) \pm \varepsilon' \tag{99}$$

ε' being also positive and arbitrarily small.

As u_{en} belongs to U_n, and u_{an} (the approximate solution yielded by the finite element method) minimizes F_n in U_n,

$$F_n(u_{an}) \leqq F_n(u_{en}) \tag{100}$$

and

$$F_n(u_{an}) \leqq F_n(u_{0n}) \pm \varepsilon'. \tag{101}$$

Let now f'_{an} be an element of H'_n defined by

$$f'_{an} = A_n u_{an}. \tag{102}$$

As u_{an} is an approximate solution to equation (42), f'_{an} generally does not coincide with f'_n.

Let u_{cn} denote the solution of the equation

$$A_n u_n = f'_{an} \tag{103}$$

in H_{A_n}. Assume that u_{cn} belongs to C_{1n}. This assumption will later be discussed.

Let u_{bn} be the field in U_n whose components and corresponding principal derivatives take the same nodal values as u_{cn}. As u_{cn} belongs to C_{1n}, and the sequence $\{U_n\}$ is complete

with respect to C_1, it is possible to find N_1 such that, for $n > N_1$,

$$d_n(u_{cn}, u_{bn}) < \varepsilon'' \tag{104}$$

ε'' being a positive number, arbitrarily small.

As A_n is a continuous operator, it is also possible to determine ε'' such that (104) implies

$$\| A_n u_{cn} - A_n u_{bn} \| < \varepsilon''' \tag{105}$$

ε''' being positive and arbitrarily small.

Let

$$f'_{bn} = A_n u_{bn}. \tag{106}$$

As u_{cn} is a solution to equation (103),

$$A_n u_{cn} = f'_{an}. \tag{107}$$

The inequality (105) can thus be transformed into

$$\| f'_{an} - f'_{bn} \| < \varepsilon'''. \tag{108}$$

By virtue of (67), as u_{bn} and u_{an} both belong to U_n,

$$\begin{aligned} [d_n(u_{an}, u_{bn})]^2 &= [(u_{an} - u_{bn}), (u_{an} - u_{bn})]_n \\ &= ((u_{an} - u_{bn}), (f'_{an} - f'_{bn}))_n. \end{aligned} \tag{109}$$

As, by Cauchy's inequality,

$$((u_{an} - u_{bn}), (f'_{an} - f'_{bn}))_n \leq \| u_{an} - u_{bn} \|_n \, \| f'_{an} - f'_{bn} \|_n < \varepsilon''' \| u_{an} - u_{bn} \|_n \tag{110}$$

we obtain

$$d_n(u_{an}, u_{bn}) < \sqrt{(\varepsilon''' \| u_{an} - u_{bn} \|_n)}. \tag{111}$$

Combining (104) with (111), there results

$$d_n(u_{an}, u_{cn}) \leq d_n(u_{an}, u_{bn}) + d_n(u_{bn}, u_{cn}) < \varepsilon^{\mathrm{iv}} \tag{112}$$

where

$$\varepsilon^{\mathrm{iv}} = \varepsilon'' + \sqrt{(\varepsilon''' \| u_{an} - u_{bn} \|_n)} \tag{113}$$

As F_n is a continuous functional, it is then possible, given $\varepsilon^{\mathrm{v}} > 0$, to determine $\varepsilon^{\mathrm{iv}}$ such that

$$F_n(u_{cn}) = F_n(u_{an}) \pm \varepsilon^{\mathrm{v}}. \tag{114}$$

As u_{cn} belongs to H_{A_n},

$$F_n(u_{cn}) \geq F_n(u_{0n}). \tag{115}$$

Thus

$$F_n(u_{0n}) \mp \varepsilon^{\mathrm{v}} \leq F_n(u_{an}). \tag{116}$$

Combining (99), (100) and (116), there results

$$F_n(u_{an}) \leq F_n(u_{en}) \leq F_n(u_{an}) \pm \varepsilon' \pm \varepsilon^{\mathrm{v}} \tag{117}$$

and thus

$$F_n(u_{en}) = F_n(u_{an}) + \varepsilon^{vi} \tag{118}$$

in which $\varepsilon^{vi} < \varepsilon' + \varepsilon^{v}$.

But, as u_{an} and u_{en} both belong to U_n, (68) permits to write

$$F_n(u_{en}) - F_n(u_{an}) = [(u_{an} - u_{en}), (u_{an} - u_{en})]_n = [d_n(u_{an}, u_{en})]^2 = \varepsilon^{vi}. \tag{119}$$

Combining (119) and (98), we obtain finally

$$d_n(u_{an}, u_{0n}) \leqslant d_n(u_{an}, u_{en}) + d_n(u_{en}, u_{0n}) < \varepsilon + \sqrt{\varepsilon^{vi}}. \tag{120}$$

Expression (120) shows that u_{an} converges to u_{0n}.

It remains to prove that u_{cn} belongs to C_{1n}. This assumption was indeed used to obtain (104).

Our reasoning will be based on a theorem which is known to be valid for Poisson's equation

$$\Delta u = f \tag{121}$$

Such theorem states that u has continuous second order derivatives in a domain Ω whenever f is a Hölder continuous function in Ω [12].

A corresponding theorem is lacking which refers to the general problem with which the present paper is concerned. We are thus not sure that the derivatives involved in the operator A are continuous whenever f is Hölder continuous. It seems however very reasonable to expect the theorem to be true, at least as far as linear elastic theories are concerned. It will thus be admitted that, at least in case of Elasticity, the Hölder continuity of the body force density implies the continuity of the displacement derivatives involved in the operator.

u_{cn} denotes the solution to equation (103) in H_{A_n}. This means, in terms of Elasticity, that u_{cn} represents the compatible field which equilibrates the same external forces as u_{an}. Such forces are of two kinds: body forces distributed within each subdomain and forces distributed on the subdomain interfaces and on S. u_{cn} will belong to C_{1n}, i.e. its (p_i+1)th derivatives will be continuous within Ω^e, if the derivatives involved in the operator are continuous within Ω^e, and thus if the body force density corresponding to u_{an} is Hölder continuous within Ω^e.

The problem now consists in proving that the body force density corresponding to u_{an} is Hölder continuous within Ω^e, no matter how large is n. We shall not attempt to investigate the general conditions in which such a statement is true.

Sometimes, however, the proof is trivial. This is namely the case if the type of the element is such that the body force density vanishes or is forced to a prescribed bounded and continuous polynomial variation within each element, no matter the values of the parameters.

If such a common situation arises, there remains no doubt that the completeness criterion is also a convergence criterion, even if conformity is not achieved. But, if it is achieved, the finite element method becomes a particularization of the Ritz method, and completeness will ensure convergence in any case.

Our reasoning can be adapted to cases [10] in which the elements are subdivided into parts and the allowed fields have different analytical expressions within each part. The body force density is then generally not continuous within the element taken as a whole. Convergence will however easily be proved if each part is treated as a separate element.

11. CONCLUSIONS

The finite element method has been presented as an analytical technique which can be applied to a very broad class of problems.

Functional Analysis provides the frame for an abstract formulation in which some generalized concepts, like energy and distance, with a physical or geometrical origin, play a fundamental role. Particularly, the definition of distance between two fields, as the square root of the energy of their difference, is an extremely convenient basis for the discussion of convergence.

The description of the finite element method also required the introduction of the concept of principal derivatives. Continuity of the field components and their principal derivatives is necessary if the energy density is to be finite. In two and three-dimensional Elasticity, for instance, as the continuity of the displacement components is enough to ensure a finite energy, the principal derivatives are of order zero.

The fact that the finite element method is based on the decomposition of the global domain into subdomains, made it convenient to transform the initial problem into an equivalent one. In Elasticity, for instance, the initial problem consists in the determination of an elastic field which verifies the field equations everywhere within the domain, while the transformed problem consists in the determination of a set of fields respecting the field equations within each subdomain and verifying compatibility and equilibrium conditions on the subdomain interfaces. In both cases the conditions imposed on the external boundary must be fulfilled.

In the finite element method, the principal continuity requirements are replaced by reduced continuity conditions which may imply the fulfillment of the principal continuity conditions everywhere in the domain. If they do, the piecewise defined fields are said to be conforming but non-conforming if it happens otherwise.

This is an opportunity to remark that conformity has been with difficulty obtained for plate and shell elements [10]. Such difficulty results from the fact that rotations are usually regarded as derivatives of the transverse displacements. This has not however to be so (see Section 4), and conformity can be easily obtained if the rotations are considered as true displacements.

When conformity is achieved, the finite element method becomes a particularization of the Ritz method. Such particularization is characterized by the piecewise definition of the field, which, together with the reduced continuity conditions, allows matrix $\mathbf{K}_n$ and vector $\mathbf{Q}_n$ (the stiffness matrix and the force vector in elastic problems) to be assembled from simpler matrices and vectors which refer to each finite element.

The matrix analysis which is developed in this paper for the determination of $\mathbf{K}_n$ and $\mathbf{Q}_n$ is a generalization of the displacement method of Structural Analysis. The force method [4] could also be used.

Completeness is a sufficient condition for convergence to the exact solution in the Ritz method, i.e. if conformity is achieved.

A general completeness criterion is justified in Section 9. It is proved that such criterion ensures completeness with respect to a set C_1 containing the fields whose derivatives of order up to (p_i+1) are continuous within the subdomains corresponding to the elements. The principal derivatives are of order p_i-1, so that the criterion does not ensure completeness with respect to the set of all the fields with finite energy.

In two and three-dimensional Elasticity, the (p_i+1)th derivatives are second order

derivatives. Their continuity implies the continuity of the body force distribution density. The continuity of the body force distribution density is not however a very strong restriction, as discontinuities of the first and second derivatives of the displacements are still admitted at points which remain on element boundaries when the size of the elements is progressively reduced. The solution of problems in which external forces are distributed on element interfaces is thus not excluded.

Completeness with respect to C_1 is a sufficient condition for convergence whenever conformity is achieved. However, it was shown in Section 10, that completeness implies convergence in any case, i.e., even when conformity is not achieved, if the body force density remains continuous and bounded within each element as n tends to infinity.

Along the whole paper the principal boundary conditions (which correspond to displacement boundary conditions in Elasticity) were supposed to be homogeneous. If they are not homogeneous, the finite element method can however still be applied. All that must be done is to make the values of the field components and their principal derivatives coincide with the prescribed values, at the nodes which are located on S_2. As the size of the elements tends to zero, we obtain approximate solutions tending to a solution which obeys the field equation inside the domain (in case the convergence criterion has been respected) and the prescribed boundary conditions on the boundary. This is of course the exact solution.

It remains to indicate that the formulation presented in this paper is not the only possible one.

In the present paper, the nodal values of the field components and their principal derivatives are indeed chosen as parameters in terms of which the reduced continuity conditions are to be expressed.

It is however possible, in elastic problems, to take as parameters the resultants and moments of the forces distributed on the element boundaries. The analysis starts then from reduced equilibrium conditions, which are directly expressed in terms of such parameters [11]. The interest of this second formulation is that it can generate equilibrated solutions while the first formulation leads to compatible ones (if conformity is achieved).

This second formulation may be generalized, as well as the first, to cover the general problem with which the present paper is concerned. It will be shown in a next paper how this can be done and which criterion may be used to ensure convergence to the exact solution.

REFERENCES

[1] S. G. MIKHLIN, *Variational Methods of Mathematical Physics*. Pergamon Press (1963).
[2] G. P. BAZELEY, Y. K. CHEUNG, B. M. IRONS and O. C. ZIENKIEWICZ, Triangular elements in plate bending. Conforming and nonconforming solutions. *Conf. Matrix Methods in Structural Mechanics*, Wright-Patterson AFB, Ohio, 1965.
[3] S. W. KEY, A convergence investigation of the direct stiffness method. Ph.D. Dissertation, University of Washington (1966).
[4] E. R. ARANTES OLIVEIRA, *Introdução à teoria das estruturas*. Lisboa, Author's edition (Dec. 1966).
[5] B. IRONS and K. DRAPER, Inadequacy of nodal connections in a stiffness solution for plate bending. *AIAA Jnl* 3, 965 (1963).
[6] L. V. KANTOROVICH and V. I. KRYLOV, *Approximate Methods of Higher Analysis*. Noordhoff (1958).
[7] E. T. WHITTAKER and G. N. WATSON. *A Course of Modern Analysis*. Cambridge University Press (1965).
[8] R. J. MELOSH, Basis for the derivation of matrices for the direct stiffness method. *AIAA Jnl* 1, 1631 (1963).

[9] O. C. ZIENKIEWICZ and Y. K. CHEUNG, *The Finite Element Method in Structural and Continuum Mechanics*. McGraw-Hill (1967).
[10] R. W. CLOUGH and J. L. TOCHER, Finite element stiffness matrices for analysis of plate bending. *Conf. Matrix Methods in Structural Mechanics*, Wright-Patterson AFB, Ohio, 1965.
[11] E. R. ARANTES OLIVEIRA, Formulações básicas do método dos elementos finitos. 2.ªs *Jornadas Luso-Brasileiras de Engenharia Civil, Rio de Janeiro-S. Paulo* (Agosto 1967).
[12] R. COURANT and D. HILBERT, *Methods of Mathematical Physics*, Vol. II. Interscience (1962).

(*Received* 13 *June* 1967; *revised* 22 *March* 1968)

Абстракт—Метод конечного элемента явлтется в настоящее время наиболее общим и одним из мощных способов расчета конструкций.

Он оказывается также общим математическим методом. Главной задачей работы является представление именно этого аспекта проблемы. Используется функциональный анализ как идеальное орудие для общей абстрактной формулировки.

Основным ддя применения этого метода является определение сходимости к точному решении последовательности приближенных решений, полученных на основе моделей конечных элементов, при уменьшании размера.

Для случая соответствия мехду элементами, метод конечного элемента оказывается частным случаем метода Ритца, так что сходимость можно гарантировать настолько насколько достигнута полнота.

В работе обснавыается общий критерий полноты. Этот критерий требует, чтобы компоненты поля и все их производные, порядка не выше старшей производной, входящей в выражение для плотности энергии, могли принимать какое либо постоянное значение в пределах элемента.

Наконец доказывается, что такой критерий йвляется также общим критерием сходимости, то есть достаточным услоием сходимости, даже если не достигается соответствия.