František Melkes
The finite element method for non-linear problems

Aplikace matematiky, Vol. 15 (1970), No. 3, 177–189

THE FINITE ELEMENT METHOD FOR NON-LINEAR PROBLEMS[1])

FRANTIŠEK MELKES

(Received February 5, 1969)

The finite element method, which is in its essence the generalized Ritz method with a special choice of the basis functions, has come forward lately. Among the publications dealing with this method let us mention [2], [3] where the method is applied to some class of non-linear ordinary differential equations, and [1], [8] where linear partial differential equations are solved by the finite element method. Further literature devoted to the subject is mentioned in [8].

In the present paper the finite element method is applied to non-linear operator equations. The attained results are used to solve the general quasilinear equation.

NON-LINEAR OPERATORS

In this section we shall deal with the solution of the operator equation

$$F(x) = \theta\,, \tag{1}$$

$\|\theta\| = 0$ where F is generally a non-linear operator defined on the whole real Banach space E. Throughout the whole section we shall suppose that the operator F is potential and hence its range is in the adjoint Banach space E^*. Conditions for the operator F, either differentiable or not, to be potentional, are given in [7].

We shall limit our considerations to the class of monotonous operators. The operator F will be called, in accordance with [4], monotonous on the space E if for arbitrary elements $x_1, x_2 \in E$ it fulfils the inequality

$$(x_1 - x_2, F(x_1) - F(x_2)) \geqq 0\,. \tag{2}$$

[1]) While the present paper was being prepared for publication, the paper P. G. CIARLET, M. H. SCHULTZ, R. S. VARGA: Numerical Methods of High-Order Accuracy for Nonlinear Boundary Value Problems, V. Monotone Operator Theory, Numer. Math. 13, 51—77 (1969) appeared which deals with similar problems.

Since the finite element method belongs to the variational methods, we shall solve a certain variational problem instead of the equation (1). The equivalence of both problems is guaranteed by the following

Lemma 1. *Let the monotonous potential operator $F(x)$ be defined on the whole Banach real space E,* $\operatorname{grad} f(x) = F(x)$. *Then the element $x^* \in E$ which minimizes the functional $f(x)$ on the space E fulfils the equation* (1). *Inversely, the solution x^* of the equation* (1) *minimizes the functional $f(x)$ on the space E.*

Proof. The first assertion is proved in [7]; the other follows from the Lagrange formula for the potential. In fact, if $F(x^*) = \theta$ then for an arbitrary element $x \in E$ there is

$$f(x) - f(x^*) = \int_0^1 (x - x^*, F(x^* + t(x - x^*)))\,\mathrm{d}t =$$

$$= \int_0^1 (x - x^*, F(x^* + t(x - x^*)) - F(x^*))\,\mathrm{d}t \geqq 0\,.$$

It turns out that the monotony of the potential operator F by itself is not sufficient for the proof of the existence and unicity of the corresponding variational problem. It is necessary that the expression on the left-hand side of the inequality (2) be suitably bounded from below. A sufficient condition for the existence and unicity of the solution of the problem is given by Theorem 2.7 in [4]. However, the course of the proof makes it possible to modify the theorem in a certain way. In view of the fact that we are going to use this modified assertion in the sequel, we introduce its full wording. We shall require that the operator F fulfils the following condition of boundedness:

1° given arbitrary elements $x_1, x_2 \in E$, the inequality

$$(3) \qquad (x_1 - x_2, F(x_1) - F(x_2)) \geqq \alpha(\|x_1 - x_2\|)$$

holds, $\alpha(t)$ being a non-negative function of the non-negative argument such that the function $\bar{\alpha}(R) = \int_0^1 \alpha(Rt)\,\mathrm{d}t/t$ is continuous and increasing for $R \geqq 0$, $\bar{\alpha}(0) = 0$ and $\lim_{R\to\infty} \bar{\alpha}(R)/R = \infty$.

Lemma 2. *Let the potential operator $F(x)$,* $\operatorname{grad} f(x) = F(x)$ *satisfying Condition 1° be defined on the real Banach space. Let $M \subset E$ be an arbitrary closed or weakly closed convex set. Then there exists one and only one element $\bar{x} \in M$ minimizing the functional $f(x)$ on the set M. Each sequence $\{x_n\} \subset M$ satisfying* $\lim_{n\to\infty} f(x_n) = \inf_{x\in M} f(x)$ *converges strongly to the element $\bar{x}$.*

Proof. Since the proof of Lemma 2 is essentially coincident with that of the above mentioned Theorem, we shall introduce it just in outline. If $x_0 \in E$ is a fixed element, it follows from the Lagrange formula for the potential and from the condition (3)

$$f(x) = f(x_0) + \int_0^1 (x - x_0, F(x_0 + t(x - x_0)))\,\mathrm{d}t \geqq$$
$$\geqq f(x_0) + \bar{\alpha}(\|x - x_0\|) - \|F(x_0)\|\,\|x - x_0\|$$

where $x \in E$ is an arbitrary element. Owing to Condition 1° there exists $R_0 > 0$ such that for all $R > R_0$ the function $\bar{\alpha}(R) - \|F(x_0)\|\,R$ is positive. Since this function is bounded from below on the interval $\langle 0, R_0\rangle$ in view of its continuity, it is bounded from below on the whole positive semi-axis. The functional $f(x)$ is consequently bounded from below on the whole space E and thus, all the more, on the set M. Hence there exists $d = \inf_{x \in M} f(x)$. For any two elements $x, y \in E$ there is

$$\tfrac{1}{2}f(x) + \tfrac{1}{2}f(y) - f\left(\frac{x + y}{2}\right) =$$
$$= \frac{1}{4}\int_0^1 \left(x - y, F\left(\frac{x + y}{2} + t\,\frac{x - y}{2}\right) - F\left(\frac{x + y}{2} - t\,\frac{x - y}{2}\right)\right) \mathrm{d}t \geqq$$
$$\geqq \tfrac{1}{4}\bar{\alpha}(\|x - y\|)\,.$$

Let us now choose an arbitrary sequence $\{x_n\} \subset M$, $\lim_{n\to\infty} f(x_n) = d$. For any $\varepsilon > 0$ and for m, n sufficiently large we have

$$f(x_n) < d + \varepsilon\,, \quad f(x_m) < d + \varepsilon\,,$$
$$f\left(\frac{x_n + x_m}{2}\right) \geqq d\,.$$

Hence

$$\tfrac{1}{4}\bar{\alpha}(\|x_n - x_m\|) \leqq \tfrac{1}{2}f(x_n) + \tfrac{1}{2}f(x_m) - f\left(\frac{x_n + x_m}{2}\right) \leqq$$
$$\leqq \frac{d + \varepsilon}{2} + \frac{d + \varepsilon}{2} - d = \varepsilon$$

and thus

$$\lim_{m,n\to\infty} \bar{\alpha}(\|x_n - x_m\|) = 0\,.$$

Condition 1° guarantees that $\lim_{m,n\to\infty} \|x_n - x_m\| = 0$ as well. In view of the completeness of the space E there is an element $\bar{x} \in E$ to which the sequence $\{x_n\}$ converges strongly and, all the more, weakly. Since the set M is closed or weakly closed, $\bar{x} \in M$ holds. The potential $f(x)$ of the monotonous operator $F(x)$ is weakly semi-continuous from below and hence

$$d \leqq f(\bar{x}) \leqq \underline{\lim}_{n\to\infty} f(x_n) = d$$

which implies $f(\bar{x}) = d$. If there existed two different elements $\bar{x}$, $\bar{\bar{x}} \in M$ satisfying $f(\bar{x}) = f(\bar{\bar{x}}) = d$ then according to (4) it would be

$$f\left(\frac{\bar{x} + \bar{\bar{x}}}{2}\right) < \tfrac{1}{2}f(\bar{x}) + \tfrac{1}{2}f(\bar{\bar{x}}) = d$$

which is a contradiction, for $\frac{1}{2}(\bar{x} + \bar{\bar{x}}) \in M$.

If in particular $M = E$ then Lemma 2 guarantees in the space E the unique existence of the minimum of the functional $f(x)$ and thus of the solution of the equation (1) as well. Any minimizing sequence converges strongly to this solution.

The difference between the mentioned Lemma and Theorem 2.7 in [4] consists partly in the existence of the minimum of the functional $f(x)$ being guaranteed not only on the whole space E but even on its closed or weakly closed convex subset, partly in the fact that Condition 1° is a little more general than the analogous condition in the Theorem.

Approximate variational methods consist in solving the variational problem not on the whole space E but only on its subset $M \subset E$. We shall require that this subset should fulfil the assumptions of the preceding Lemma, i.e. that it should be a closed or a weakly closed convex set. The element $\bar{x} \in M$ which minimizes the functional $f(x)$ on the set M and which exists uniquely according to Lemma 2 will be called an approximate solution of the equation (1). Let us deal now with the estimate of the error caused by replacing the exact solution x^* of the equation (1) by the approximate solution $\bar{x}$. To this purpose it will be necessary for the operator F to fulfil some further condition of boundedness:

2° given arbitrary elements $x_1, x_2 \in E$, the inequality

$$(5) \qquad (x_1 - x_2, F(x_1) - F(x_2)) \leqq \beta(\|x_1 - x_2\|)$$

holds, $\beta(t)$ being a non-negative function of the non-negative argument such that the function $\bar{\beta}(R) = \int_0^1 \beta(Rt)\, \mathrm{d}t/t$ is continuous and increasing for $R \geqq 0$, $\bar{\beta}(0) = 0$.

An estimate of the error of the solution is given by the following

Theorem 1. *Let a potential operator $F(x)$,* grad $f(x) = F(x)$ *fulfilling Conditions* 1° *and* 2° *be defined on the real Banach space E. Let $M \subset E$ be a closed or weakly closed convex set. Denote by $x^* \in E$ the element for which $f(x^*) = \min_{x\in E} f(x)$ and $\bar{x} \in M$ the element for which $f(\bar{x}) = \min_{x\in M} f(x)$. Then there holds for any $x \in M$*

$$(6) \qquad \|\bar{x} - x^*\| \leqq \gamma(\|x - x^*\|)$$

where $\gamma(R)$ is a certain increasing non-negative function of the non-negative argument such that $\gamma(0) = 0$.

Proof. Since $F(x^*) = \theta$ in view of Lemma 1, we can write

$$f(x) - f(x^*) = \int_0^1 (x - x^*, F(x^* + t(x - x^*)) - F(x^*))\,dt \tag{7}$$

for any $x \in E$. Applying the inequality (3) to this relation we get

$$\bar{\alpha}(\|\bar{x} - x^*\|) \leqq f(\bar{x}) - f(x^*) .$$

The right-hand side may be increased on the set M since the definition of the element $\bar{x} \in M$ implies the inequality $f(\bar{x}) \leqq f(x)$ for all $x \in M$ and hence

$$\bar{\alpha}(\|\bar{x} - x^*\|) \leqq f(x) - f(x^*) .$$

If we use again (7) and the inequality (5) we obtain

$$\bar{\alpha}(\|\bar{x} - x^*\|) \leqq \bar{\beta}(\|x - x^*\|) .$$

Since the function $\bar{\alpha}(R)$ is positive, continuous and increasing on the whole positive semi-axis, it has on the whole semi-axis a continuous inverse function $\bar{\alpha}^{-1}$ which is increasing as well and $\bar{\alpha}^{-1}(0) = 0$.

With regard to the last inequality we have

$$\|\bar{x} - x^*\| \leqq \gamma(\|x - x^*\|)$$

where $\gamma(R) = \bar{\alpha}^{-1}[\bar{\beta}(R)]$. The function $\gamma(R)$ is obviously continuous and increasing for $R \geqq 0$ and $\gamma(0) = 0$.

Thus, if we succeed in finding a single element $\tilde{x} \in M$ which in the norm of the space E differs only little from the exact solution x^*, then Theorem just proved guarantees that the error of the solution is sufficiently small as well. Owing to (7) and to Condition 2° the relation

$$0 \leqq f(\bar{x}) - f(x^*) \leqq f(x) - f(x^*) \leqq \bar{\beta}(\|x - x^*\|) \tag{8}$$

holds for all $x \in M$ expressing the fact that the error of the approximation is also small. The construction of the element $\tilde{x} \in M$ sufficiently close to the exact solution x^* depends on the choice of the space E as well as of the set M. We shall show later some practical examples of the choice of this element.

Let us now choose a finite dimensional subspace which is closed and convex as the set M. Denote by n its dimension and by $x_1, \ldots, x_n$ its arbitrarily chosen linearly independent elements. Any element $x \in M$ can be written in the form

$$x = \sum_{i=1}^{n} c_i x_i \tag{9}$$

where $c_1, \ldots, c_n$ are suitable real numbers. The functional $f(x)$ on the subspace M can be then considered as a function of real variables $c_1, \ldots, c_n$, i.e.

$$\varphi(c_1, \ldots, c_n) = f(\sum_{i=1}^{n} c_i x_i)\,.$$

If the requirement $\bar{\beta}(0) = 0$ from Condition 2° is replaced by a stronger one

$$\lim_{R \to 0_+} \frac{\bar{\beta}(R)}{R} = 0 \tag{10}$$

then also the condition $\lim_{R \to 0_+}[\bar{\alpha}(R)/R] = 0$ is fulfilled owing to the inequality $\bar{\alpha}(R) \leqq \bar{\beta}(R)$ and the function $\varphi(c_1, \ldots, c_n)$ has partial derivatives of the first order with respect to all variables c_j, $j = 1, \ldots, n$. In fact, if x is in the form (9), then

$$\frac{\partial \varphi}{\partial c_j} = \lim_{s \to 0} \frac{f(x + s x_j) - f(x)}{s} = \lim_{s \to 0} \int_0^1 (x_j, F(x + t s x_j))\, dt =$$
$$= (x_j, F(x)) + \lim_{s \to 0} \int_0^1 (x_j, F(x + t s x_j) - F(x))\, dt\,.$$

For the second term it holds with regard to Conditions 1° and 2°

$$\lim_{s \to 0_+} \frac{1}{s}\ \bar{\alpha}(s \|x_j\|) \leqq \lim_{s \to 0_+} \int_0^1 (x_j, F(x + t s x_j) - F(x))\, dt \leqq$$
$$\leqq \lim_{s \to 0_+} \frac{1}{s}\ \bar{\beta}(s \|x_j\|)$$

and hence it vanishes. It would be possible to show analogously that the second term vanishes for $s \to 0_-$ as well. Partial derivatives of the first order of the function $\varphi(c_1, \ldots, c_n)$ hence exist and are given by

$$\frac{\partial \varphi}{\partial c_j} = \left(x_j, F\left(\sum_{i=1}^{n} c_i x_i\right)\right). \tag{11}$$

The coefficients $\bar{c}_1, \ldots, \bar{c}_n$ of the element $\bar{x} \in M$ for which the functional $f(x)$ attains its minimum can be determined either by the gradient method or by solving a generally non-linear system of equations

$$(x_j, F(\sum_{i=1}^{n} c_i x_i)) = 0\,, \quad j = 1, \ldots, n$$

which has precisely one solution owing to Lemma 2.

In practice, the problem of solving the operator equation (1) often occurs, with the operator F satisfying the following conditions:

3° for any elements x, $h \in E$ there exists the Gateaux derivative $F'_x h$ (the linear Gateaux differential);

4° the functional $(h_1, F'_x h_2)$ is continuous with respect to x on an arbitrary hyperplane passing through x for any elements $h_1, h_2 \in E$;

5° for arbitrary elements $x, h_1, h_2 \in E$ it is $(h_1, F'_x h_2) = (h_2, F'_x h_1)$;

6° there exist positive constants α_0, β_0 such that for arbitrary x, $h \in E$ it is

$$\alpha_0^2 \|h\|^2 \leqq (h, F'_x h) \leqq \beta_0^2 \|h\|^2 . \tag{12}$$

Conditions 3° to 5° guarantee that the operator F is potential (cf. [7]). Making use of the Lagrange formula for the operator, we obtain for arbitrary elements $x_1, x_2 \in E$ the identity

$$(x_1 - x_2, F(x_1) - F(x_2)) = \int_0^1 (h, F'_x h) \, \mathrm{d}t \tag{13}$$

with $h = x_1 - x_2$, $x = x_2 + th$ which makes it possible to verify Conditions 1° and 2°. From (12) and (13) it follows

$$(x_1 - x_2, F(x_1) - F(x_2)) \geqq \alpha_0^2 \|x_1 - x_2\|^2 .$$

The function $\alpha(t)$ from Condition 1° is defined in the following way:

$$\alpha(t) = \alpha_0^2 t^2 .$$

The corresponding function

$$\bar{\alpha}(R) = \int_0^1 \alpha_0^2 (Rt)^2 \frac{\mathrm{d}t}{t} = \tfrac{1}{2} \alpha_0^2 R^2$$

is obviously continuous and increasing and $\bar{\alpha}(0) = 0$,

$$\lim_{R \to \infty} \frac{\bar{\alpha}(R)}{R} = \tfrac{1}{2} \alpha_0^2 \lim_{R \to \infty} R = \infty .$$

Condition 1° is thus fulfilled. We show analogously that 2° is fulfilled as well. Hence, if 3° to 6° are fulfilled, then according to our preceding considerations there exists precisely one solution x^* of the equation (1). This solution minimizes on E its potential. On an arbitrary finite dimensional subspace M the solution x^* of (1) can be replaced by the approximate solution $\bar{x}$ which also exists uniquely. Since, as we can verify easily by a direct computation, the function $\gamma(R)$ occuring in the assertion of Theorem 1 is given by the relation $\gamma(R) = \beta_0 \, . \, R/\alpha_0$ the error of the solution is in its order equal to the distance of the chosen element of the set M from the exact solution.

The function $\bar{\beta}(R)$ satisfies also the suplementary condition (10) and hence the function defined above, $\varphi(c_1, \ldots, c_n)$, has all partial derivatives of the first order. Condition 3° guarantees that all these derivatives are continuous. We shall show that in this case, the function $\varphi(c_1, \ldots, c_n)$ has even all partial derivatives of the second order, all being continuous functions. Let us choose arbitrarily $j, k = 1, \ldots, n$, consider x in the form (9) and compute

$$\frac{\partial^2 \varphi}{\partial c_j \, \partial c_k} = \lim_{s \to 0} \frac{1}{s} \left[(x_j, F(x + s x_k)) - (x_j, F(x)) \right] =$$

$$= \lim_{s \to 0} \int_0^1 (x_j, F'_y x_k) \, \mathrm{d}t = (x_j, F'_x x_k)$$

where we put $y = x + s t x_k$. The last limiting process may be performed owing to Condition 4°. This condition guarantees also the continuity of the second partial derivatives.

QUASILINEAR EQUATIONS

The results of the previous section will be now applied to the solution of the quasilinear partial differential equation in the divergence form which is solved in [4].

In the n-dimensional space R^n with the general point $x = (x_1, \ldots, x_n)$ let an open bounded set Ω be given with a sufficiently smooth boundary. Denote $D^\mu = \partial^{|\mu|}/\partial x_1^{\mu_1} \ldots \partial x_n^{\mu_n}$ where $\mu = (\mu_1, \ldots, \mu_n)$, $|\mu| = \mu_1 + \ldots + \mu_n$. All the derivatives are considered in the generalized sense. The scalar product in the space $W_2^{(m)}$ will be denoted by $(u, v)_m$, the corresponding norm by $\|u\|_m^2 = (u, u)_m$; in particular, by $(u, v)_0$ we shall denote the scalar product in the space L_2.

Consider the quasilinear partial differential equation of the order $2m$, $m \geqq 1$ in the form

$$\sum_{|\mu| \leqq m} (-1)^{|\mu|} D^\mu \, a_\mu(x, u, \ldots, D^m u) = g \tag{14}$$

where $g \in L_2(\Omega)$. The solution of this equation will be sought for in the space E satisfying $\mathring{W}_2^{(m)}(\Omega) \subset E \subset W_2^{(m)}(\Omega)$. The coefficients a_μ are supposed to satisfy the following condition

7° all coefficients a_μ are real continuous functions of all their arguments and for all $u \in W_2^{(m)}(\Omega)$, $x \in \Omega$ they satisfy the inequality

$$|a_\mu(x, u(x), \ldots, D^m u(x))| \leqq \varphi(\|u\|_m) \left[\sum_{|\nu| \leqq m} |D^\nu u(x)| + 1 \right] \tag{15}$$

where $\varphi(R)$ is a continuous non-negative function of the non-negative variable.

To the differential equation (14), the non-linear Dirichlet form

(16) $$A(u, v) = \sum_{|\mu| \leqq m} (a_\mu(x, u, \ldots, D^m u), D^\mu v)_0$$

having sense for all $u, v \in W_2^{(m)}(\Omega)$ will be adjoined. The form will be supposed to satisfy the following condition:

8° there is a positive constant α_0 such that for all $u, v \in E$ the following inequality holds:

$$\alpha_0^2 \|u - v\|_m^2 \leqq A(u, u - v) - A(v, u - v)\,.$$

Making use of the Hölder inequality and of (15) we find out that (16) is a linear functional bounded with respect to v. Consequently, to each $u \in E$ it is possible to determine uniquely an element $G(u) \in E$ so that for all $v \in E$

(17) $$(v, G(u))_m = A(u, v)\,.$$

The function $u^* \in E$ will be called the weak solution of (14) corresponding to the space E if for all $v \in E$

$$A(u^*, v) = (g, v)_0\,.$$

It is shown in [4] that if Conditions 7° and 8° are fulfilled then for any $g \in L_2(\Omega)$ there exists precisely one weak solution $u^* \in E$ of the equation (14) corresponding to the space E. Moreover, the element $u^* \in E$ is the weak solution if and only if it satisfies the equation

(18) $$F(u) \equiv G(u) - w = \theta$$

where $w \in E$ is uniquely determined by the relation

(19) $$(w, v)_m = (g, v)_0$$

which is valid for all $v \in E$.

If we want to use the finite element method to determine the solution of the equation (18) – and thus also the weak solution of the equation (14) – we have to add some supplementary assumptions. To this purpose, note that the Dirichlet form (16) is a functional of two variables. Let us denote by $A_u^{\cdot}(h_1, h_2)$ its Gateaux derivative with respect to the first variable, i.e. let us put

$$A_u^{\cdot}(h_1, h_2) = \lim_{s \to 0} \frac{1}{s} [A(u + sh_1, h_2) - A(u, h_2)]$$

for arbitrary elements $u, h_1, h_2 \in E$. Further, the fulfilment of the following conditions will be required:

9° for all elements u, h_1, $h_2 \in E$ there exists the Gateaux derivative $A'_u(h_1, h_2)$, it is continuous with respect to u on any hyperplane passing through u and $A'_u(h_1, h_2) = A'_u(h_2, h_1)$;

10° there exists a positive constant β_0 so that for all u, $v \in E$

$$A(u, u - v) - A(v, u - v) \leqq \beta_0^2 \|u - v\|_m^2 .$$

We shall show that under these assumptions the operator $F(u)$ given by (18) satisfies Conditions 1° and 2°, it is potential and its potential is of the form

$$f(u) = \int_0^1 A(tu, u)\, \mathrm{d}t - (u, w)_m \tag{20}$$

w being determined by (19). The fulfilment of Conditions 1° and 2° follows immediately from 8° and 10° since we have the equality

$$(u - v, F(u) - F(v))_m = A(u, u - v) - A(v, u - v)$$

for all u, $v \in E$.

Let us now compute the gradient of the functional (20). It is

$$\lim_{\tau \to 0} \frac{f(u + \tau h) - f(u)}{\tau} =$$

$$= \lim_{\tau \to 0} \int_0^1 \frac{A(tu + \tau th, u + \tau h) - A(tu, u)}{\tau}\, \mathrm{d}t - (h, w)_m =$$

$$= \lim_{s \to 0} \int_0^1 \left[A(tu + sh, h) + t\, \frac{A(tu + sh, u) - A(tu, u)}{s} \right] \mathrm{d}t -$$

$$- (h, w)_m = \int_0^1 [A(tu, h) + t\, A'_{tu}(h, u)]\, \mathrm{d}t - (h, w)_m =$$

$$= \int_0^1 [A(tu, h) + t\, A'_{tu}(u, h)]\, \mathrm{d}t - (h, w)_m =$$

$$= \int_0^1 \left[A(tu, h) + t\, \frac{\mathrm{d}}{\mathrm{d}t} A(tu, h) \right] \mathrm{d}t - (h, w)_m =$$

$$= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} [t\, A(tu, h)]\, \mathrm{d}t - (h, w)_m = A(u, h) - (h, w)_m =$$

$$= (h, G(u))_m - (h, w)_m = (h, F(u))_m$$

which proves that (20) is the potential of the operator F. All assumptions of Lemma 1 and 2 as well as those of Theorem 1 are fulfilled and therefore we can replace the weak solution u^* of the equation (14) on an arbitrary finite dimensional subspace

$M \subset E$ by an approximate solution $\bar{u} \in M$ which minimizes the functional (20) on the set M. For the error of the solution it is with respect to (6)

$$\|\bar{u} - u^*\|_m \leqq \frac{\beta_0}{\alpha_0} \|\tilde{u} - u^*\|_m \tag{21}$$

where $\tilde{u} \in M$ is a suitably chosen element. One way of choosing this element is given in [1], another one in [8]. In both cases the considerations are made without expressing the basis functions of the finite dimensional space M explicitly and are restricted at most to a two-dimensional space, the reasoning in a general n-dimensional space being too complicated.

In [1], the two-dimensional region Ω is assumed to be a polygon whose sides are parallel to the coordinate axes. Every such polygon can be expressed as a union of rectangles $R_i = \langle a_i, b_i \rangle \times \langle c_i, d_i \rangle$, $i = 1, \ldots, k$ any two of them being either disjoint or having a part of the boundary in common. On every rectangle let us define a partition ϱ_i:

$$a_i = x_0^i < x_1^i < \ldots < x_{N_i}^i = b_i ,$$
$$c_i = y_0^i < y_1^i < \ldots < y_{N'_i}^i = d_i .$$

A partition of the whole region Ω is such partition which is defined on each rectangle R_i by means of ϱ_i. A system of such partitions let us denote by C. We say that this system is regular if there exist such positive constants σ, τ, η that for all i, $1 \leqq i \leqq k$ and for all $\varrho \in C$ there holds

$$\sigma \bar{\pi}_i \leqq \pi_i , \quad \sigma \bar{\pi}'_i \leqq \pi'_i , \quad \eta \leqq \bar{\pi}'_i / \bar{\pi}_i \leqq \tau$$

where

$$\bar{\pi}_i = \max_j \left(x_{j+1}^i - x_j^i\right) , \quad \bar{\pi}'_i = \max_j \left(y_{j+1}^i - y_j^i\right) ,$$
$$\pi_i = \min_j \left(x_{j+1}^i - x_j^i\right) , \quad \pi'_i = \min_j \left(y_{j+1}^i - y_j^i\right) .$$

As the finite dimensional subspace M on which the approximate solution is sought for we take the set $M = E \cap H^{(m)}(\varrho, \Omega)$ where $H^{(m)}(\varrho, \Omega)$ for any natural m and for any choice of $\varrho \in C$ is the set of all real functions u defined on the set Ω, satisfying the condition $D^{(i,j)} u \in C^0(\Omega)$ for all i, j for which $0 \leqq i, j \leqq m - 1$, and being a polynomial of the degree $2m - 1$ on each elementary rectangle of which the above described rectangle R_i consists.

If the solution $u^* \in S^{p,r}(\Omega)$, $p \geqq 2m$, $r \geqq 2$ where $S^{p,r}(\Omega)$ is the set of all functions $u \in W_r^{(p)}(\Omega)$ satisfying $D^\mu u \in C^0(\Omega)$, $|\mu| < p$, then in the quality of $\tilde{u}$ we take the element of the set $H^{(m)}(\varrho, \Omega)$ forming the $H^{(m)}(\varrho, \Omega)$-approximation of the element u^*. It is shown in [1] that if C is a regular system of partitions of the region Ω then there exists a constant K independent of the choice of the partition $\varrho \in C$ so that it holds

$$\|\tilde{u} - u^*\|_m \leqq K \varkappa^m ,$$

$\varkappa = \max_i (\bar{\pi}_i, \bar{\pi}'_i)$. If $\tilde{u} \in E$ then $\tilde{u} \in M$ and making use of the last inequality we get the result that when replacing the weak solution u^* of the equation (14) by the approximate solution $\bar{u} \in M$ we make an error estimated by

$$\|\bar{u} - u^*\|_m \leqq K \frac{\beta_0}{\alpha_0} \varkappa^m .$$

Another choice of the element $\tilde{u}$ is introduced in [8]. The region Ω may be more general, viz. an arbitrary polygon. On this polygon we perform a triangulation, i.e. we express it in the form of a union of triangles T_i each two of them either being disjoint or having in common a vertex or a side. If we denote by $\varkappa_i$ the largest side and by ϑ_i the smallest angle of the triangle T_i then each triangulation is characterized by the quantities $\varkappa = \max_i \varkappa_i$, $\vartheta = \min_i \vartheta_i$. A system C of triangulations will be called regular if there exists a constant $\vartheta_0 > 0$ such that $\vartheta \geqq \vartheta_0$. In the quality of the set M we take $M = E \cap H^{(l)}(\Omega)$ where $H^{(l)}(\Omega)$ is the system of functions being polynomials of two variables of the degree l on each triangle T_i and satisfying some conditions at vertices, centres of sides or at centres of gravity of the triangles (cf. [8]). If the function u^* is $(l - m - 1)$-times continuously differentiable and if it has bounded derivatives of the $(l + 1)$-st order, then the element of the set $H^{(l)}(\Omega)$ satisfying the above mentioned conditions at the vertices, centres of sides or centres of gravity with parameters given by the exact solution u^* can be taken for $\tilde{u}$. For $m = 1$, $l = 2, 3$ and $m = 2$, $l = 5$ it is shown in [8] that

$$\|\tilde{u} - u^*\|_m \leqq K \frac{\varkappa^{l-m+1}}{\sin^m \vartheta}$$

where the constant K is independent of the choice of the partition $\varrho \in C$. If $\tilde{u} \in E$ and thus $\tilde{u} \in M$ then we get in these cases the following estimate for the error of the solution:

$$\|\bar{u} - u^*\|_m \leqq K \frac{\beta_0}{\alpha_0} \frac{\varkappa^{l-m+1}}{\sin^m \vartheta} .$$

E.g. when solving the Dirichlet problem $u|_{\dot{\Omega}} = 0$ for the equation (14), there is $E = \mathring{W}_2^{(m)}(\Omega)$ and using any of the two mentioned ways of dividing the region Ω the element $\tilde{u}$ selected above belongs to E.

In conslusion I would like to express my gratitude to Prof. M. Zlámal who read the manuscript carefully and made many valuable comments.

References

[1] *Birkhoff G., Schultz M. H., Varga R. S.:* Piecewise Hermite Interpolation in One and Two Variables with Applications to Partial Differential Equations. Numer. Math. 11, (1968), 232–256.

[2] *Ciarlet P. G., Schultz M. H., Varga R. S.:* Numerical Methods of High-Order Accuracy for Nonlinear Boundary Value Problems, I. One Dimensional Problem. Numer. Math. 9 (1967), 394–430.

[3] *Ciarlet P. G., Schultz M. H., Varga R. S.:* Numerical Methods of High-Order Accuracy for Nonlinear Boundary Value Problems, II. Nonlinear Boundary Conditions. Numer. Math. 11, (1968), 331–345.

[4] *Качуровский Р. И.*: Нелинейные монотонные операторы в Банаховых пространствах. Успехи мат. наук XXIII (1968), 2 (140), 121—168.

[5] *Михлин С. Г.*: Численная реализация вариационных методов. Москва 1966.

[6] *Йосида К.*: Функциональный анализ, Москва 1967.

[7] *Вайнберг М. М.*: Вариационные методы исследования нелинейных операторов, Москва 1956.

[8] *Zlámal M.:* On the Finite Element Method. Numer. Math. 12 (1968), 394–409.

Souhrn

METODA KONEČNÝCH PRVKŮ PRO NELINEÁRNÍ PROBLÉMY

FRANTIŠEK MELKES

Práce pojednává o metodě konečných prvků, která je v podstatě zobecněnou Ritzovou metodou se speciálním výběrem bázových funkcí. Metoda konečných prvků byla různými autory aplikována na nelineární obyčejné diferenciální rovnice i na lineární parciální diferenciální rovnice. V předložené práci je tato metoda použita při řešení nelineární operátorové rovnice. Operátor stojící na levé straně zmíněné rovnice je potenciální a splňuje jisté podmínky ohraničenosti. Z těchto předpokladů vyplývá jednoznačná existence jak přesného tak přibližného řešení rovnice i jistý odhad chyby řešení. Dosažené výsledky jsou využity při řešení obecné kvasilineární rovnice.

Author's address: RNDr. *František Melkes,* Výzkumný a vývojový ústav elektrických strojů točivých, Mostecká 26, Brno 14.