RAIRO. ANALYSE NUMÉRIQUE

MILOŠ ZLAMAL

A finite element solution of the monlinear heat equation

RAIRO. Analyse numérique, tome 14, nº 2 (1980), p. 203-216

<http://www.numdam.org/item?id=M2AN_1980__14_2_203_0>

R A I R O Analyse numérique/Numerical Analysis
(vol 14 n° 2 1980 p 203 à 216)

A FINITE ELEMENT SOLUTION OF THE NONLINEAR HEAT EQUATION (*)

by Miloš ZLAMAL ([1])

Communiqué par P G CIARLET

Abstract — *Transformation of dependent variables as, e g , the Kirchhoff transformation, is a classical tool for solving nonlinear partial differential equations In* [1] *this approach was used in connection with the finite element method and applied to the solution of nonlinear heat conduction problems, of degenerate parabolic equations and of multidimensional Stefan problems Here we give a justification of the method for the nonlinear heat equation in case that the discretization is carried through by piecewise-linear polynomials in space and by the implicit Euler method in time An estimate of the discretization error in the maximum norm is introduced and the convergence rate of the nonlinear Gauss-Seidel method is investigated*

Résume — *Une transformation portant sur les variables dependantes, par exemple la transformation de Kirchhoff, est un outil classique pour resoudre des equations aux dérivées partielles non linéaires On a combiné en* [1] *cette approche avec la methode des eléments finis pour résoudre des problèmes non lineaires de transmission de chaleur, des équations paraboliques dégénérées, et des problèmes de Stefan à plusieurs dimensions On donne dans cet article une justification de la méthode dans le cas de l'équation non linéaire de la chaleur, lorsque la discrétisation correspond à des polynômes linéaires par morceaux en variable d'espace, et à une méthode d'Euler implicite en variable de temps On introduit une estimation de l'erreur de discretisation dans la norme du maximum, et on etudie la vitesse de convergence de la méthode de Gauss-Seidel non lineaire*

1. THE PROBLEM, THE METHOD AND THE RESULTS

Let Ω be a bounded two or three-dimensional domain with a Lipschitz boundary Γ. We consider the following initial-boundary value problem

$$\left\{\begin{aligned} & c(u)\frac{\partial u}{\partial t}=\nabla.(k(u)\nabla u)+q(u,x,t), \\ & \qquad x\in\Omega, \quad t\in]0,T[, \quad T<\infty, \end{aligned}\right. \tag{1.1}$$

$$u(x,0)=u^0(x), \qquad x\in\Omega, \tag{1.2}$$

$$\left\{\begin{aligned} & u=\varphi(x,t), \qquad x\in\Gamma^1, \quad t\in]0,T[, \\ & -k(u)\frac{\partial u}{\partial \nu}=\psi(u,x,t), \qquad x\in\Gamma^2, \quad t\in]0,T[. \end{aligned}\right. \tag{1.3}$$

(*) Received May 1979

([1]) L P S Technical University Brno Tchecoslovaquie

R A I R O Analyse numerique/Numerical Analysis, 0399-0516/1980/203/$ 5 00

Here x is the point $(x_1, \ldots, x_N)$ and $N=2, 3$, $c(u)$ and $k(u)$ are piecewise continuously differentiable functions bounded from below and from above by positive constants,

$$0<c_1 \leqq c(u) \leqq c_2, \qquad 0<k_1 \leqq k(u) \leqq k_2, \qquad \forall u \in]-\infty, \infty[, \tag{1.4}$$

the function $q(u, x, t)$ satisfies

$$\left.\begin{aligned} &|q(u_1, x, t)-q(u_2, x, t)| \leqq L|u_1-u_2|, \\ &|q(u, x, t_1)-q(u, x, t_2)| \leqq L|t_1-t_2|, \\ &x \in \overline{\Omega}, \quad t \in]0, T[, \quad u_1, u_2, u \in]-\infty, \infty[, \end{aligned}\right\} \tag{1.5}$$

$u^0 \in H^2(\Omega)$, $\Gamma^1 \cup \Gamma^2 = \Gamma$, φ is continuous on $\overline{\Omega} \times]0, T[$, ν is the outward normal to Γ^2, ψ is continuous for $u \in]-\infty, \infty[$, $x \in \overline{\Omega}$, $t \in]0, T[$ and

$$\psi(u_2, x, t)-\psi(u_1, x, t) \geqq 0, \qquad \forall u_1, u_2, \quad u_2 \geqq u_1. \tag{1.6}$$

We assume that the problem (1.1)-(1.3) has a unique solution.

REMARK: As soon as we know an *a priori* bound for u in the maximum norm it is sufficient that the assumptions introduced above hold for u from a bounded interval.

The following notation is used

$$H^m(\Omega)=\{v \in L^2(\Omega);\ D^\alpha v \in L^2(\Omega),\ \forall |\alpha| \leqq m\},$$

$$W^{m,\infty}(\Omega)=\{v \in L^\infty(\Omega);\ D^\alpha v \in L^\infty(\Omega),\ \forall |\alpha| \leqq m\},$$

$$V=\{v \in H^1(\Omega),\ v|_{\Gamma_1}=0\},$$

$$(u, v)=\int_\Omega uv\,dx,$$

$$\langle u, v \rangle=\int_{\Gamma^2} uv\,d\sigma, \qquad a(u, v)=\int_\Omega \nabla u . \nabla v\,dx.$$

If X is a Banach space, $L^\infty(0, T; X)=L^\infty(X)$ denotes the space of functions $v(t) : t \in]0, T[\to v(t) \in X$, which are measurable and such that

$$\operatorname*{ess\,sup}_{t \in]0, T[} \|v(t)\|_X = \|v\|_{L^\infty(X)} < \infty.$$

Further, we introduce the enthalpy

$$H(u)=\int_0^u c(s)\,ds \tag{1.7}$$

and the Kirchhoff transformation

$$G(u)=\int_0^u k(s)\,ds. \tag{1.8}$$

If u is sufficiently smooth [(1.25) is sufficient] then multiplying (1.1) by a function $v\in V$ and using Green's theorem we get the identity

$$(\dot{H}, v)+a(G, v)+\langle \psi, v\rangle=(q, v), \qquad \forall v\in V, \quad t\in]0, T[, \tag{1.9}$$

where

$$\dot{H}=\frac{\partial}{\partial t}H(u), \qquad G=G(u), \qquad \psi=\psi(u, x, t), \qquad q=q(u, x, t).$$

We consider a family $\mathcal{T}_h$ of triangulations consisting of triangles and tetrahedrons, respectively, with vertices lying in $\overline{\Omega}$. Let K (a closed set) denote an element of $\mathcal{T}_h$, $h_K=\text{diam}(K)$

$$\rho_K=\sup\{\text{diam}(S);\ S \text{ is a ball contained in } K\},$$
$$\overline{\Omega}_h=\bigcup_{K\in\mathcal{T}_h} K \text{ (in general, } \Omega_h\neq\Omega),$$

$\Gamma_h=\partial\Omega_h$. Let Γ_h^i $(i=1, 2)$ be the parts of Γ_h corresponding to Γ^i and let Γ_h^1 be such that it is a closed set the boundary of which consists of vertices of triangles and of edges of tetrahedrons, respectively. Hence, Γ_h^2 is open and a boundary side and face, respectively, belongs either to Γ_h^1 or to $\overline{\Gamma}_h^2$. Finally, we assume that the family $\mathcal{T}_h$ is regular in the following sense (Ciarlet [2], p. 132) :

(*a*) there exists a constant σ such that

$$\frac{h_K}{\rho_K}\leqq\sigma, \qquad \forall K\in\bigcup_h \mathcal{T}_h,$$

(*b*) the quantity

$$h=\max_{K\in\mathcal{T}_h} h_K$$

approaches zero.

To each triangulation $\mathcal{T}_h$ we associate the finite dimensional space

$$V_h=\{v\in C^0(\overline{\Omega}_h);\ v \text{ is piecewise linear on } \mathcal{T}_h \text{ and } v|_{\Gamma_h^1}=0\}. \tag{1.10}$$

Also the space

$$W_h=\{v\in C^0(\overline{\Omega}_h);\ v \text{ is piecewise linear on } \mathcal{T}_h\}, \tag{1.11}$$

will be needed.

Let $\{x^j\}_{j=1}^{r}$ be the set of all nodes of $\mathscr{T}_h$ and $\{x^j\}_{j=1}^{p}$ be the set of nodes from $\Omega_h \cup \Gamma_h^2$ (hence $\{x^j\}_{j=p+1}^{r}$ are nodes lying on Γ_h^1). Let $v_j(x)$ be the basis function associated to the node x^j

$$(v_j(x) \in W_h,\ v_j(x^k)=0 \text{ for } k \neq j,\ v_j(x^j)=1).$$

We consider a uniform partition of the interval $[0, T]$:

$$0<t_1<\ldots<t_M, \qquad t_i=i\,\Delta t \quad (i=1, \ldots, M), \qquad M<\frac{T}{\Delta t}.$$

The value u^i of the exact solution $u(x, t)$ at the time $t=t_i$ will be approximated by

$$U^i=\sum_{j=1}^{r} U_j^i v_j(x), \qquad U_j^i=\varphi(x^j, t_i), \quad j=p+1, \ldots, r. \tag{1.12}$$

[REMARK: The condition $U_j^i=\varphi(x^j, t_i)$, $j=p+1, \ldots, r$ is equivalent to $U^i|_{\Gamma_h^1}=\varphi_h^i|_{\Gamma_h^1}$ where for the approximation φ_h^i of $\varphi(x, t_i)$ we take the interpolate of $\varphi(x, t_i)$].

For approximations of $H^i=H(u^i)$, $G^i=G(u^i)$, $\psi^i=\psi(u^i, x, t_i)$ and $q^i=q(u^i, x, t_i)$ we take

$$\left.\begin{aligned} W^i&=\sum_{j=1}^{r} H(U_j^i)v_j(x), & Y^i&=\sum_{j=1}^{r} G(U_j^i)v_j(x), \\ \Psi^i&=\sum_{j=1}^{r} \psi(U_j^i, x^j, t_i)v_j(x), & Q^i&=\sum_{j=1}^{r} q(U_j^i, x^j, t_i)v_j(x). \end{aligned}\right\} \tag{1.13}$$

Evidently, the functions W^i, etc. are interpolates of the functions $H(U^i)$, etc.

We also approximate the bilinear forms (w, v), $a(w, v)$, $\langle w, v\rangle$. To this end the following quadrature formula for a N-dimensional simplex S is chosen:

$$I_s(F)=\frac{1}{N+1}\,\text{meas}\,(S)\sum_{i=1}^{N+1} F(x^i) \tag{1.14}$$

[in (1.14) x^i are vertices of S]. Then we set

$$\left.\begin{aligned} (w, v)_h&=\sum_{K\in\mathscr{T}_h} I_K(wv), \qquad a_h(w, v)=\sum_{K\in\mathscr{T}_h} I_K(\nabla w.\nabla v), \\ \langle w, v\rangle_h&=\sum_{K'\in\overline{\Gamma_h^2}} I_{K'}(wv). \end{aligned}\right\} \tag{1.15}$$

Here K' denotes a side of a triangle and a face of a tetrahedron, respectively. Obviously,

$$I_K(\nabla w.\nabla v)=\int_K \nabla w.\nabla v\,dx, \qquad \forall\, w, v\in V_h;$$

therefore

$$a_h(w, v) = \int_{\Omega_h} \nabla w . \nabla v \, dx, \qquad \forall w, v \in V_h. \tag{1.16}$$

Now we can derive the discrete analog of (1.9) by means of which the approximate solution will be defined. We put $t = t_{i+1}$ in (1.9), we replace $\dot{H}^{i+1}$ by $\Delta t^{-1}(W^{i+1} - W^i)$, G^{i+1} by Y^{i+1}, ψ^{i+1} by Ψ^{i+1}, q^{i+1} by Q^i, the forms (w, v), $a(w, v)$, $\langle w, v \rangle$ are replaced by $(w, v)_h$, $a_h(w, v)$, $\langle w, v \rangle_h$ and u^0 by the interpolate u_I^0. We get

$$\left.\begin{aligned} &(W^{i+1} - W^i, v)_h + \Delta t a_h(Y^{i+1}, v) + \Delta t \langle \Psi^{i+1}, v \rangle_h = \Delta t (Q^i, v)_h, \\ &\qquad \forall v \in V_h, \qquad i = 0, \ldots, M-1, \\ &\qquad\qquad U^0 = u_I^0. \end{aligned}\right\} \tag{1.17}$$

We prove that there exists just one set $\{U^i\}_{i=1}^M$ of functions of the form (1.12) satisfying (1.17). U^i is the value of the approximate solution U of the problem (1.1)-(1.3) at the time $t = t_i$.

Let us introduce the following notations: $\{x^j\}_{j=1}^{p'}$ are nodes from Ω_h:

$$\xi_j = \mathrm{U}_j^{i+1}, \qquad j = 1, \ldots, p,$$

$$\mathbf{H}(\xi) = (H(\xi_1), \ldots, H(\xi_p))^T, \qquad \mathbf{G}(\xi) = (G(\xi_1), \ldots, G(\xi_p))^T,$$

$$\psi(\xi) = (0, \ldots, 0, \psi(\xi_{p'+1}, x^{p'+1}, t_{i+1}), \ldots, \psi(\xi_p, x^p, t_{i+1}))^T,$$

$$M = \{(v_i, v_j)_h\}_{i,j=1}^p, \qquad K = \{a_h(v_i, v_j)\}_{i,j=1}^p,$$

$$B = \{\langle v_i, v_j \rangle_h\}_{i,j=1}^p.$$

The matrices M, K, B are constant $p \times p$ band matrices, M and K are positive definite and B is positive semidefinite. In addition, owing to the choice (1.13) of the quadrature formula the matrices M and B are diagonal (the engineers speak about lumping).

Suppose now that U^i has been computed. Setting $v = v_j$, $j = 1, \ldots, p$ in (1.17), transferring all given or computed terms to the right-hand side and denoting it by $\mathbf{f}$ we see that the computation of U^{i+1} is equivalent to the solution of the nonlinear system

$$M \mathbf{H}(\xi) + \Delta t \mathrm{K} \mathbf{G}(\xi) + \Delta t B \boldsymbol{\Psi}(\xi) = \mathbf{f}. \tag{1.18}$$

We introduce the new variables

$$\zeta_j = G(\xi_j), \qquad j = 1, \ldots, p. \tag{1.19}$$

Due to the assumption (1.4), the mapping (1.19) maps R^p one-to-one on R^p. We set (m_j and b_j are diagonal elements of M and B, hence $m_j>0$, $b_j \geqq 0$, $j=1, \ldots, p$):

$$\sigma_j(s)=m_j H(s)+\Delta t b_j \psi(s, x^j, t_{i+1}),$$
$$F_j=\sigma_j . G^{-1}, \qquad \mathbf{F}(\zeta)=(F_1(\zeta_1), \ldots, F_p(\zeta_p))^T.$$

Then (1.18) is equivalent to the system

$$\mathbf{F}(\zeta)+\Delta t K \zeta=\mathbf{f}. \tag{1.20}$$

(1.20) is a necessary condition for the minimum of the functional J :

$$\zeta=(\zeta_1, \ldots, \zeta_p)^T \in R^p \to J(\zeta),$$
$$J(\zeta)=\sum_{j=1}^{p} \int_0^{\zeta_j} F_j(s)\, ds+\frac{1}{2} \Delta t \zeta^T K \zeta-\mathbf{f}^T \zeta. \tag{1.21}$$

The G-derivative of J is a uniformly monotone mapping on R^p (*see* Ortega and Rheinboldt [5], p. 141) due to the assumptions (1.4), (1.6) and to the fact that M and K are positive definite matrices, B is positive semidefinite and M, B are diagonal matrices. Therefore J is uniformly convex on R^p (*see* [5], 3.4.5) and subsequently (*see* [5], 4.37) it has a unique global minimizer. If J attains the minimum at $\zeta=\zeta^0$ then $\xi_j^0=G^{-1}(\xi_j^0)$ is the only solution to (1.18).

Several Galerkin-type methods leading to the solution of linear systems were proposed for the nonlinear heat equation. Let us mention the predictor-corrector method and the Crank-Nicolson extrapolation by Douglas and Dupont [3]. These methods are certainly good when applied to mildly nonlinear problems. The method discussed here leads to the solution of nonlinear systems. This difficulty is compensated by three things: 1) The method gives very good results (even when Δt is not very small) also in case that rapid variation of heat capacity occurs within a narrow temperature range and the boundary condition is highly nonlinear. 2) It is not necessary to recompute the matrices M, K, B at every time step as in the methods leading to linear systems. 3) We shall prove that the nonlinear Gauss-Seidel method applied to (1.18) converges at least so fast as the linear Gauss-Seidel method in case of the linear heat equation with $c(u)=c_1$, $k(u)=k_2$.

Consider the following linear elliptic boundary value problem: Find z such that $z-\varphi^* \in V$ and

$$a(z, v)+\langle \psi^*, v \rangle=(f, v), \qquad \forall v \in V. \tag{1.22}$$

Here $\varphi^*(x)\in C^0(\overline{\Omega})\cap H^1(\Omega)$, $\psi^*(x)\in C^0(\overline{\Omega})$ and $f(x)\in L^2(\Omega)$. The approximate solution $z_h\in W_h$ is defined by

$$\left.\begin{aligned} a_h(z_h, v)+\langle \psi^*, v\rangle_h&=(f, v)_h, \qquad \forall v\in V_h,\\ z_h(x^j)&=\varphi^*(x^j), \qquad \forall x^j\in\Gamma_h^1. \end{aligned}\right\} \tag{1.23}$$

(REMARK: The discrete boundary condition is, in fact, the discrete analog of $z-\varphi^*\in V$ because it is equivalent to $z_h-\varphi_h^*\in V_h$ where for the approximation φ_h^* of φ^* we take the interpolate of φ^*). We will assume that the following error bound in the maximum norm is valid: If $z\in W^{2,\infty}(\Omega)$ then

$$\|z-z_h\|_{L^\infty(\Omega\cap\Omega_h)}\leqq C\|z\|_{W^{2,\infty}(\Omega)}\vartheta(h). \tag{1.24}$$

Before introducing the main result of the paper we need one more definition: The triangulation $\mathcal{T}_h$ is called of acute type if all angles of the triangles and all angles made by adjacent faces and edges of tetrahedrons, respectively, are not greater than $(1/2)\pi$.

THEOREM. — *Let the triangulations $\mathcal{T}_h$ be of acute type and let the solution u of the problem* (1.1)-(1.3) *be sufficiently smooth, i.e.*

$$\left.\begin{aligned} G(u)\in L^\infty(W^{2,\infty}(\Omega)), \qquad \frac{\partial}{\partial t}G(u)\in L^\infty(W^{2,\infty}(\Omega)),\\ \frac{\partial^2}{\partial t^2}H(u)\in L^\infty(L^\infty(\Omega)). \end{aligned}\right\} \tag{1.25}$$

Then

$$\|u^i-U^i\|_{L^\infty(\Omega\cap\Omega_h)}\leqq C[\vartheta(h)+\Delta t], \qquad i=1,\ \ldots,\ M, \tag{1.26}$$

where the constant C does not depend on $\vartheta(h)$, Δt and i.

REMARK: Several papers contain error bounds in the maximum norm for solutions of elliptic boundary value problems. Nevertheless, we do not know such error bounds for the general formulation (1.23). Usually, Γ is supposed to be a polygon and a polyhedron, respectively (hence $\Omega=\Omega_h$), $\Gamma^1=\Gamma$ and $\varphi^*=0$ or $\Gamma^2=\Gamma$ and $\psi^*=0$ and numerical integration is not taken into account. We refer to the papers by J. A. Nitsche [4] and R. Scott [6] where in two dimensions there are proved error bounds of the form (1.24) with $\vartheta(h)=h^2|\lg h|$.

2. PROOF OF THE ERROR ESTIMATE

If $v\in W_h$ attains a local maximum (minimum) at a point from $\overline{\Omega}_h$ then evidently it attains this maximum (minimum) also at a node. v attains a local maximum (minimum) at a node x^j iff the values of v at the neighbouring nodes are not greater (smaller) then $v(x^j)$.

LEMMA: *Let the triangulation $\mathcal{T}_h$ be of acute type. If $v \in W_h$ attains a local maximum (minimum) at the node x^j then*

$$a_h(v, v_j) \geqq 0 (\leqq 0), \tag{2.1}$$

where v_j is the basis function associated with the node x^j. Further,

$$k_{il} = a_h(v_i, v_l) \leqq 0 \qquad \text{if} \quad i \neq l. \tag{2.2}$$

Proof: We restrict ourselves to the two-dimensional case and to the maximum. $a_h(v, v_j)$ is equal to the sum of integrals $\int_K \nabla v . \nabla v_j \, dx$ where K is any triangle with the vertex x^j. Consider $\int_K \nabla v . \nabla w \, dx$. Any displacement, rotation and reflection does not change the expression $\nabla v . \nabla w$. As the Jacobian of such transformations is equal to ± 1 we have (writing, for a moment, x, y instead of x_1, x_2):

$$\int_K \nabla v . \nabla w \, dx \, dy = \int_{K'} \nabla v' . \nabla w' \, d\xi \, d\eta .$$

We take such transformations that the vertices of the resulting K' are the points $(0, 0)$, $(\xi_2, 0)$, (ξ_3, η_3) with $\xi_2 > 0$, $\xi_3 \geqq 0$, $\eta_3 > 0$. By elementary computations we get

$$\begin{aligned} \int_K \nabla v . \nabla w \, dx \, dy = \frac{1}{2 \xi_2 \eta_3} \{ & \eta_3^2 (v^2 - v^1)(w^2 - w^1) \\ & + [-\xi_3 (v^2 - v^1) + \xi_2 (v^3 - v^1)] \\ & \times [-\xi_3 (w^2 - w^1) + \xi_2 (w^3 - w^1)] \}, \end{aligned} \tag{2.3}$$

where v^i, w^i $(i = 1, 2, 3)$ are the values of v and w at the vertices $(0, 0)$, $(\xi_2, 0)$ and (ξ_3, η_3), respectively. If v attains a local maximum at x^j we choose the transformations so that $v^1 \leqq v^2 \leqq v^3 = v(x^j)$. Then for $w = v_j$ we have $w^1 = w^2 = 0$, $w^3 = 1$ and

$$\int_K \nabla v . \nabla v_j \, dx \, dy = \frac{1}{2 \eta_3} [-\xi_3 (v^2 - v^1) + \xi_2 (v^3 - v^1)].$$

All angles of K are not greater than $(1/2)\pi$. Therefore $\xi_2 \geqq \xi_3$, hence $\int_K \nabla v . \nabla v_j \, dx \, dy \geqq 0$ which proves (2.1).

If x^i and x^l are not neighbours then $a_h(v_i, v_l)=0$. If they are neighbours, then $a_h(v_i, v_l)$ is a sum of two integrals over triangles which both have x^i and x^l for vertices. Consider any of these triangles. We set $v=v_i$, $w=v_l$ in (2.3) and choose the transformations in such a way that $v^1=0, v^2=1, v^3=0, w^1=w^2=0, w^3=1$. We get

$$\int_K \nabla v_i \nabla v_l \, dx\, dy = -\frac{1}{2}\frac{\xi_3}{\eta_3} \leqq 0$$

which proves (2.2).

REMARK : An easy consequence of the lemma is a discrete maximum principle. Take $\Gamma_h^1=\Gamma_h$ and let S be the set of ordered couples (k, l), $k=1, \ldots, r$ (r is as before the number of all nodes of the triangulation $\mathcal{T}_h$), $l=0, \ldots, M-1$, such that either $x^k \in \Gamma_h$ and $l=0, \ldots, M-1$ or $x^k \in \Omega_h$ and $l=0$. Let $\{U_i\}_{i=1}^M$ be the functions from W_h satisfying

$$(W^{i+1}-W^i\ v)_h+\Delta t a_h(Y^{i+1}, v)=\Delta t(Q^i, v)_h,$$
$$\forall v \in V_h, \quad i=0, \ldots, M-1.$$

If $q \leqq 0$ ($q \geqq 0$) then it holds

$$U_j^i > \max_{(k\ l)\in S} U_k^l \left(\geqq \min_{(k\ l)\in S} U_k^l\right), \qquad j=1, \ldots, r; \quad i=1, \ldots, M. \tag{2.4}$$

Proof of the theorem. In the sequel, C will denote a generic constant, not necessarily the same in any two places, which does not depend on h, Δt, i.

From (1.9) and (1.3) if follows that $G^i=G(u^i)=G(u(x, t_i))$ satisfies

$$a(G^i, v)+\langle \psi^i, v\rangle=(q^i-\dot{H}^i, v), \qquad \forall v \in V,$$
$$G^i|_{\Gamma_1}=G(\varphi^i)|_{\Gamma_1}.$$

Let $y^i \in W_h$ be the approximate solution of the above problem:

$$\left.\begin{aligned} &a_h(y^i, v)+\langle \psi^i, v\rangle_h=(q^i-\dot{H}^i, v)_h \qquad \forall v \in V_h, \\ &y^i(x^j)=G(\varphi^i(x^j)), \qquad \forall x^j \in \Gamma_h^1. \end{aligned}\right\} \tag{2.5}$$

From (1.24) and from the first requirement in (1.25) we have for $i=1, \ldots, M$:

$$\|G^i-y^i\|_{L^\infty(D)} \leqq C\vartheta(h), \qquad D=\Omega \cap \Omega_h. \tag{2.6}$$

Also

$$\|G^{i+1}-G^i-(y^{i+1}-y^i)\|_{L^\infty(D)}$$
$$\leqq C\|G^{i+1}-G^i\|_{W^{2,\infty}(\Omega)}\vartheta(h) \leqq C\,\Delta t\vartheta(h). \tag{2.7}$$

We derive a relation which will play a fundamental role in the error estimation. First, notice that $(f, v)_h = (f_I, v)_h$, $\langle f, v \rangle_h = \langle f_I, v \rangle_h$ for $v \in W_h$ and for any function f defined on $\overline{\Omega_h}$. Therefore using (2.5) we get

$$\begin{aligned}(H_I^{i+1} - H_I^i, v)_h + \Delta t a_h(y^{i+1}, v) \\ + \Delta t \langle \psi_I^{i+1}, v \rangle_h = \Delta t (\Delta t^{-1} [H^{i+1} - H^i] - \dot{H}^{i+1}, v)_h \\ + \Delta t (q^{i+1}, v)_h = \Delta t (r^i, v)_h + \Delta t (q^{i+1}, v)_h, \qquad \forall v \in V_h,\end{aligned} \tag{2.8}$$

$$|r_j^i| \leqq C \Delta t, \qquad j = 1, \ldots, r, \quad i = 0, \ldots, M-1 \tag{2.9}$$

[the subscript denotes always the node at which the corresponding value is taken, e. g. $r_j^i = r(x^j, t_i)$]. (2.9) follows from three facts: (*a*) all nodes of $\mathcal{T}_h$ lie in $\overline{\Omega}$; (*b*) the implicit Euler method is of order one; (*c*) we assume

$$\frac{\partial^2}{\partial t^2} H(u) \in L^\infty (L^\infty(\Omega)).$$

Set

$$\omega^i = H_I^i - W^i, \qquad \varepsilon^i = y^i - Y^i, \qquad \eta^i = \psi_I^i - \Psi^i, \qquad e^i = u_I^i - U^i. \tag{2.10}$$

Subtracting (2.8) from (1.17) one obtains.

$$\begin{aligned}(\omega^{i+1} - \omega^i, v)_h + \Delta t a_h(\varepsilon^{i+1}, v) + \Delta t \langle \eta^{i+1}, v \rangle_h \\ = \Delta t (Q^i - q^{i+1}, v)_h - \Delta t (r^i, v)_h \qquad \forall v \in V_h.\end{aligned} \tag{2.11}$$

We estimate $Q^i - q^{i+1}$ by means of (1.5):

$$\begin{aligned}q(U_j^i, x^j, t_i) - q(u_j^{i+1}, x^j, t_{i+1}) = q(U_j^i, x^j, t_i) \\ - q(u_j^i, x^j, t_i) + q(u_j^i, x^j, t_i) \\ - q(u_j^i, x^j, t_{i+1}) + q(u_j^i, x^j, t_{i+1}) \\ - q(u_j^{i+1}, x^j, t_{i+1}) = O(|e_j^i| + \Delta t).\end{aligned}$$

As

$$\varepsilon_j^i = y_j^i - G(u_j^i) + G(u_j^i) - G(U_j^i) = O(\vartheta(h)) + k(\tau_j^i) e_j^i$$

by the Mean-Value theorem and by (2.6), we see that

$$|e_j^i| \leqq C[\vartheta(h) + |\varepsilon_j^i|], \tag{2.12}$$

hence

$$|Q_j^i - q(u_j^{i+1}, x^j, t_{i+1})| \leqq C[\vartheta(h) + \Delta t + |\varepsilon_j^i|]$$

and (2 11) is equivalent with

$$\left.\begin{aligned}(\omega^{i+1} - \omega^i, v)_h + \Delta t a_h(\varepsilon^{i+1}, v) + \Delta t \langle \eta^{i+1}, v \rangle_h = \Delta t (r^i, v)_h, \\ \forall v \in V_h, \qquad |r_j^i| \leqq C(\delta + |\varepsilon_j^i|), \qquad \delta = \vartheta(h) + \Delta t.\end{aligned}\right\} \tag{2.13}$$

We use the notation $\Delta\omega_j^i = \omega_j^{i+1} - \omega_j^i$, etc. and we express $\Delta\omega_j^i$ by means of $\Delta\varepsilon_j^i$. First, by (2.7):

$$\Delta\varepsilon_j^i = \Delta y_j^i - \Delta G(u_j^i) + \Delta G(u_j^i) - \Delta G(U_j^i) = O(\Delta t\, \vartheta(h)) + \Delta G(u_j^i) - \Delta G(U_j^i).$$

By the Mean-Value theorem

$$\Delta G(u_j^i) - \Delta G(U_j^i) = k(\xi_j^i)\,\Delta u_j^i - k(\zeta_j^i)\,\Delta U_j^i$$

$$= k(\zeta_j^i)\,\Delta e_j^i + [k(\xi_j^i) - k(\zeta_j^i)]\,\Delta u_j^i = k_j^i\,\Delta e_j^i + [k(\xi_j^i) - k(\zeta_j^i)]\,O(\Delta t),$$

$$\xi_j^i \in]u_j^i,\, u_j^{i+1}[, \qquad \zeta_j^i \in]U_j^i,\, U_j^{i+1}[, \qquad k_j^i = k(\zeta_j^i).$$

The numbers ξ_j^i, ζ_j^i are of the form

$$\xi_j^i = (1-\alpha)\,u_j^i + \alpha\, u_j^{i+1}, \qquad 0 < \alpha < 1,$$

$$\zeta_j^i = (1-\beta)\,U_j^i + \beta\, U_j^{i+1}, \qquad 0 < \beta < 1,$$

therefore

$$\xi_j^i - \zeta_j^i = (1-\beta)\,e_j^i + \beta\, e_j^{i+1} + (\alpha - \beta)\,\Delta u_j^i,$$

hence

$$\Delta\varepsilon_j^i = k_j^i\,\Delta e_j^i + \Delta t\, O(\delta + |e_j^i| + |e_j^{i+1}|).$$

Similarly,

$$\Delta\omega_j^i = c_j^i\,\Delta e_j^i + \Delta t\, O(\Delta t + |e_j^i| + |e_j^{i+1}|).$$

From the last two equations and from (2.12) and (1.4) it follows

$$\left.\begin{aligned} &\Delta\omega_j^i = c_j^i\,\Delta\varepsilon_j^i + \Delta t\, O(\delta + |\varepsilon_j^i| + |\varepsilon_j^{i+1}|), \\ &0 < \frac{c_1}{k_2} \leqq d_j^i \leqq \frac{c_2}{k_1}. \end{aligned}\right\} \tag{2.14}$$

We come to the estimation of $\|\varepsilon^i\|_{L^\infty(\Omega_h)}$. As ε^i is piecewise linear it is sufficient to estimate $\max\limits_k |\varepsilon_k^i|$ We denote

$$\varepsilon^{i+1} = (\varepsilon_1^i, \ldots, \varepsilon_r^i)^T, \qquad \|\varepsilon^{i+1}\|_\infty = \max_k |\varepsilon_k^i|.$$

Let $\|\varepsilon^{i+1}\|_\infty = |\varepsilon_j^{i+1}|$ and let first $\varepsilon_j^{i+1}\, e_j^{i+1} \geqq 0$. If $\varepsilon_j^{i+1} > 0$ then $e_j^{i+1} \geqq 0$ and, due to (1.6), $\eta_j^{i+1} \geqq 0$. We put $v = v_j$ in (2.13) and use (2.1) and (2.14). We get easily

$$m_j\, d_j^i(\varepsilon_j^{i+1} - \varepsilon_j^i) \leqq C\, m_j\, \Delta t(\delta + \|\varepsilon^i\|_\infty + \varepsilon_j^{i+1}),$$

consequently

$$\|\varepsilon^{i+1}\|_\infty \leqq (1+C\Delta t)\|\varepsilon^i\|_\infty + C\Delta t\,\delta. \tag{2.15}$$

(2.15) can be proved in the same way if $\varepsilon_j^{i+1}<0$ and if $\varepsilon_j^{i+1}=0$ then $\|\varepsilon^{i+1}\|_\infty=0$. Let now $\varepsilon_j^{i+1}e_j^{i+1}<0$. If $\varepsilon_j^{i+1}>0$ then $e_j^{i+1}<0$ and, as

$$\varepsilon_j^{i+1}=O(\vartheta(h))+k(\tau_j^{i+1})e_j^{i+1}$$

[*see* the line preceeding to (2.12)], it holds $e_j^{i+1}>-C\vartheta(h)$. Because e_j^{i+1} is negative it follows $e_j^{i+1}=O(\vartheta(h))$ and

$$\|\varepsilon^{i+1}\|_\infty \leqq C\vartheta(h). \tag{2.16}$$

(2.16) can be proved in the same way if $\varepsilon_j^{i+1}<0$, $e_j^{i+1}>0$. As $\|\varepsilon^0\|_\infty \leqq C\vartheta(h)$ we *see* from (2.15) and (2.16) that

$$\left.\begin{aligned} &\|\varepsilon^0\|_\infty \leqq C\delta,\\ &\|\varepsilon^{i+1}\|_\infty \leqq \max\{(1+C\Delta t)\|\varepsilon^i\|_\infty + C\Delta t\,\delta,\ C\delta\}\\ &i=0,\ \ldots,\ M-1. \end{aligned}\right\} \tag{2.17}$$

To finish the proof we set $\alpha^0=C\delta$, $\alpha^{i+1}=\gamma\alpha^i+C\Delta t\,\delta$, $\gamma=1+C\Delta t$. Evidently, $\alpha^i \geqq C\delta$. By induction we easily prove $\|\varepsilon^i\|_\infty \leqq \alpha^i$. As

$$\alpha^i \leqq C\exp(TC)\,\delta, \qquad i=1,\ \ldots,\ M$$

we get $\|\varepsilon^i\|_\infty \leqq C\delta$ and by (2.12) $\|\mathbf{e}^i\|_\infty \leqq C\delta$, $i=1,\ \ldots,\ M$ e^i is piecewise linear, hence $\|e^i\|_{L^\infty(\Omega_h)} \leqq C\delta$. Finally, $u^i-U^i=u^i-u_I^i+u_I^i-U^i$ The first term is in the $L^\infty(D)$-norm bounded by Ch^2 (*see* [2]), the other by $C\delta$ which proves (1.26).

3. CONVERGENCE OF THE NONLINEAR GAUSS-SEIDEL ITERATION

We consider the system (1.20) which is equivalent to (1.18): If ζ^ν are Gauss-Seidel iterates for the system (1.20) then

$$(\xi_1^\nu,\ \ldots,\ \xi_p^\nu)^T=(G^{-1}(\zeta_1^\nu),\ \ldots,\ G^{-1}(\zeta_p^\nu))^T$$

are Gauss-Seidel iterates for (1.18). We assume again that the triangulations $\mathcal{T}_h$ are of acute type and instead of (1.6) we assume.

$$\frac{\partial\psi}{\partial u}\geqq 0. \tag{3.1}$$

Let

$$0<\alpha \leqq \inf_{-\infty<s<\infty} \frac{c(s)}{k(s)}. \tag{3.2}$$

As we assume (1.4) we can take

$$\alpha=\frac{c_1}{k_2}.$$

Consider the linear system

$$\alpha M \mathbf{y}+\Delta t K \mathbf{y}=\mathbf{f}, \tag{3.3}$$

which we get if we solve the linear heat equation $\alpha(\partial u/\partial t)=\Delta u+q(x, t)$. Let $\mathbf{y}^\nu$ be the Gauss-Seidel iterates for the system (3.3) and let us choose $\mathbf{y}^0$ such that it satisfies

$$\mathbf{y}^0 \leqq \mathbf{y} \tag{3.4}$$

(i. e. $y_j^0 \leqq y_j$, $j=1, \ldots, p$). It is easy to see that

$$\mathbf{y}^\nu \leqq \mathbf{y}, \qquad \nu=1, \ldots \tag{3.5}$$

In fact, let $\mathbf{y}^n \leqq \mathbf{y}$, $n=1, \ldots \nu$. The nondiagonal elements k_{ij} of K are nonpositive [*see* (2.2)]. Therefore

$$\begin{aligned}\alpha m_1 y_1^{\nu+1}+\Delta t k_{11} y_1^{\nu+1} &= -\Delta t \sum_{s>1} k_{1s} y_s^\nu+f_1 \\ &\leqq -\Delta t \sum_{s>1} k_{1s} y_s+f_1=\alpha m_1 y_1+\Delta t k_{11} y_1,\end{aligned}$$

thus $y_1^{\nu+1} \leqq y_1$. Supposing $y_s^{\nu+1} \leqq y_s$ for $s \leqq j$ we prove in the same way that $y_{j+1}^{\nu+1} \leqq y_{j+1}$. Hence, $\mathbf{y}^{\nu+1} \leqq \mathbf{y}$ which proves (3.5).

We now require that $\mathbf{y}^0$ satisfies

$$|\zeta-\zeta^0| \leqq \mathbf{y}-\mathbf{y}^0 \tag{3.6}$$

(i. e., $|\zeta_j-\zeta_j^0| \leqq y_j-y_j^0$, $j=1, \ldots, p$). For such a choice of $\mathbf{y}^0$ we prove that

$$|\zeta-\zeta^\nu| \leqq \mathbf{y}-\mathbf{y}^\nu, \tag{3.7}$$

i. e., the Gauss-Seidel iterates ζ^ν for the nonlinear system (1.20) converge in each component at least so fast as the iterates $\mathbf{y}^\nu$ for the linear system (3.3). From (3.7) it also follows that $|\mathbf{G}(\xi)-\mathbf{G}(\xi^\nu)| \leqq \mathbf{y}-\mathbf{y}^\nu$, hence

$$|\xi-\xi^\nu| \leqq \frac{1}{k_1}(\mathbf{y}-\mathbf{y}^\nu). \tag{3.8}$$

Proof of (3.7): From (3.5) it follows $\mathbf{y}-\mathbf{y}^\nu \geqq 0$. Assume that $|\zeta-\zeta^n| \leqq \mathbf{y}-\mathbf{y}^n$, $n=1, \ldots, \nu$. Set $\varphi_j(s)=F_j(s)+\Delta t k_{jj} s$. We have

$$|\zeta_1-\zeta_1^{\nu+1}|=\left|\varphi_1^{-1}\left(-\Delta t \sum_{s>1} k_{1s}\zeta_s+f_1\right)-\varphi_1^{-1}\left(-\Delta t \sum_{s>1} k_{1s}\zeta_s^\nu+f_1\right)\right|.$$

As

$$\varphi_j' = \frac{\sigma_j'}{k} + \Delta t k_{jj} \geqq \alpha m_j + \Delta t k_{jj},$$

we get by means of the Mean-Value theorem

$$\begin{aligned} |\zeta_1 - \zeta_1^{\nu+1}| &\leqq |\alpha m_1 + \Delta t k_{11})^{-1} \Big| -\Delta t \sum_{s>1} k_{1s}(\xi_s - \xi_s^{\nu}) \Big| \\ &\leqq |\alpha m_1 + \Delta t k_{11}|^{-1} \Big[-\Delta t \sum_{s>1} k_{1s}(y_s - y_s^{\nu}) \Big] = y_1 - y_1^{\nu+1} \end{aligned}$$

Let $|\zeta_s - \zeta_s^{\nu+1}| \leqq y_s - y_s^{\nu+1}$ for $s \leqq j$ Then

$$\begin{aligned} |\zeta_{j+1} - \zeta_{j+1}^{\nu+1}| &= \Big| \varphi_{j+1}^{-1} \Big(-\Delta t \sum_{s<j+1} k_{j+1s}\zeta_s - \Delta t \sum_{s>j+1} k_{j+1s}\zeta_s + f_{j+1} \Big) \\ &\quad - \varphi_{j+1}^{-1} \Big(-\Delta t \sum_{s<j+1} k_{j+1s}\zeta_s^{\nu+1} - \Delta t \sum_{s>j+1} k_{j+1s}\zeta_s^{\nu} + f_{j+1} \Big) \Big| \\ &\leqq (\alpha m_{j+1} + \Delta t k_{j+1j+1})^{-1} \Big| -\Delta t \sum_{s<j+1} k_{j+1s}(\zeta_s - \zeta_s^{\nu+1}) \\ &\qquad - \Delta t \sum_{s>j+1} k_{j+1s}(\zeta_s - \zeta_s^{\nu}) \Big| \\ &\leqq (\alpha m_{j+1} + \Delta t k_{j+1j+1})^{-1} \Big[-\Delta t \sum_{s<j+1} k_{j+1s}(y_s - y_s^{\nu+1}) \\ &\qquad - \Delta t \sum_{s>j+1} k_{j+1s}(y_s - y_s^{\nu}) \Big] = y_{j+1} - y_{j+1}^{\nu+1} \end{aligned}$$

Hence $|\boldsymbol{\zeta} - \boldsymbol{\zeta}^{\nu+1}| \leqq \mathbf{y} - \mathbf{y}^{\nu+1}$ which proves (3 7)

REMARK The mapping defined by the left-hand side of (1 20) is an M-function in the sense of Ortega, Rheinboldt [5] (p 468)

REFERENCES

1 L ČERMAK and M ZLÁMAL, *Transformation of Dependent Variables and the Finite Element Solution of Non-Linear Evolution Equations*, Int Y Numer Meth Eng, Vol 15 1980 pp 31-40

2 P G CIARLET *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam New York, Oxford, 1978

3 J DOUGLAS and T DUPONT, *Galerkin Methods for Parabolic Equations*, SIAM J Numer Anal, Vol 7, 1970, pp 575-626

4 J NITSCHE, *L^∞-Convergence of Finite Element Approximations*, *Mathematical Aspects of the Finite Element Method*, Rome, 1975, pp 261-274, Springer-Verlag, Berlin, Heidelberg, New York, 1977

5 J M ORTEGA and W C RHEINBOLDT, *Iterative Solution of Non-Linear Equations in Several Variables*, Academic Press, New York, London, 1970

6 R SCOTT, *Optimal L^∞ Estimates for the Finite Element Method on Irregular Meshes*, Math Comp, Vol 30, No 136, 1976, pp 681-697