REVUE FRANÇAISE D'AUTOMATIQUE, INFORMATIQUE, RECHERCHE OPÉRATIONNELLE. ANALYSE NUMÉRIQUE

ALAN E. BERGER

The truncation method for the solution of a class of variational inequalities

Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique, tome 10, n° 1 (1976), p. 29-42.

<http://www.numdam.org/item?id=M2AN_1976__10_1_29_0>

R.A.I.R.O.
ANALYSE NUMÉRIQUE
(Vol. 10, n° 3, mars 1976, p. 29 à 42)

THE TRUNCATION METHOD FOR THE SOLUTION OF A CLASS OF VARIATIONAL INEQUALITIES (*)

par Alan E. BERGER (1)

Communiqué par G. STRANG

Abstract. – *A concise numerical method for a class of variational inequality problems is presented, and numerical results are given. Sufficient conditions for stability are derived, and an error estimate is obtained for the steady state problem.*

I. INTRODUCTION

We consider a concise numerical method, called the truncation method, for the solution of a class of variational inequalities. The truncation method was developed for approximating the solution of a specific parabolic diffusion-consumption problem with a moving free surface in Berger, Ciment, and Rogers [2]. Since this diffusion-consumption problem is equivalent to a parabolic variational inequality problem (Lewy and Stampacchia [10]), the truncation method is hence seen to be applicable to this type of variational inequality.

The results of numerical experiments will be presented which indicate an $0(\Delta t + \Delta x^2)$ error in L^2 for the parabolic problem. Numerical solution of an elliptic variational inequality problem by relaxing the corresponding time dependent problem toward the steady state will also be analyzed. Stability conditions will be given under which the truncation method solution will

(*) This research has been supported by the Naval Surface Weapons Center Independent Research Fund and under ONR Research Project Number RR 014-03-01 (NR 044-453).

(1) Mathematical Analysis Division Naval Surface Weapons Center Silver Spring, Maryland.

approach a limiting value. An error estimate will be obtained by comparing this limit with the Ritz solution, for which an $0(\Delta x)$ error estimate in H^1 has been proven by Mosco and Strang [15] and Falk [6].

II. DESCRIPTION OF THE FREE BOUNDARY PROBLEM AND THE TRUNCATION METHOD

For simplicity we state the equations in one space dimension; wherever a second spatial dimension affects the analysis of the truncation method it will be pointed out. The type of problem considered in [2] was to find the concentration $c(x, t) \geqslant 0$ of a substance which diffuses and is consumed at unit rate wherever it is present. Letting $s(t)$ denote the location of the interface between the region where $c > 0$ and the region where $c = 0$, an example of the establishment of a steady state concentration is given by

$$c_t = c_{xx} - 1 \qquad 0 < x < s(t) \qquad t > 0 \tag{1a}$$

$$c(0, t) = 1/2 \qquad t \geqslant 0 \tag{1b}$$

$$c(s(t), t) = c_x(s(t), t) = 0 \qquad t > 0 \tag{1c}$$

$$c(x, 0) = c_0(x) \equiv \left\{ \begin{array}{ll} 9(x - 1/3)^2/2 & x \in [0, 1/3] \\ 0 & x \geqslant 1/3 \end{array} \right\} \tag{1d}$$

$$s(0) = 1/3. \tag{1e}$$

The steady state solution $\sigma(x)$ is

$$\sigma(x) = \left\{ \begin{array}{ll} (1 - x)^2/2 & 0 \leqslant x \leqslant s \\ 0 & x \geqslant s \end{array} \right\} \text{where } s = 1. \tag{2}$$

The truncation method for (1) works in the following way. One chooses a fixed interval $J = (0, b)$ such that $s(t) \leqslant b$ (by numerical experiment if necessary). Having approximate solution values C^n at time t^n, one obtains intermediate values $\tilde{C}^{n+1}$ for time $t^{n+1} = t^n + \Delta t$ by advancing

$$c_t = c_{xx} - 1 \qquad x \in J \tag{3a}$$

$$c(0, t) = 1/2 \tag{3b}$$

$$c(b, t) = 0 \tag{3c}$$

$$c(t = t^n) = C^n \tag{3d}$$

one time step from t^n to t^{n+1} using some finite difference or finite element scheme for (3). At each node point p, the approximate solution for (1) at time t^{n+1} is then given by

$$C^{n+1}(p) = \max (0, \tilde{C}^{n+1}(p)). \tag{4}$$

The truncation method does not depend in any way on explicitly tracking the front $s(t)$ (whose approximate location is taken to be the first node point p at which C^n vanishes). Tracking multiple fronts or fronts in two space dimensions thus presents no computational difficulty, since the method works by advancing a linear parabolic problem (3) on a fixed domain J (or a two dimensional analog) followed by the simple operations (4). In [2] the truncation method with the alternating-direction finite difference scheme was used to solve such a two dimensional problem on a rectangle.

The discussion in Lewy and Stampacchia [10] (see their Problem I) and in Brézis [3] (in particular pages 99-101) demonstrates that (1) is equivalent to a parabolic variational inequality problem which will be described below. A two dimensional steady state free boundary problem arising in semiconductor theory, which corresponds to the two dimensional problem considered in [2], is formulated as a variational inequality by Hunt and Nassif in [8].

III. FORMULATION OF THE VARIATIONAL INEQUALITY

For an open set Ω, we let $H^1(\Omega)(H^2(\Omega))$ denote the Sobolev space of functions with one (two) L^2 derivatives in Ω, and we let $H_0^1(\Omega) \subset H^1(\Omega)$ be those functions which vanish on the boundary Γ of Ω (in the sense of the Trace Theorem). We also use the notation $W_\infty^2(\Omega)$ for the functions in $H^2(\Omega)$ whose derivatives up through order two are bounded on Ω. For $v, w \in L^2(\Omega)$ we set $(v, w) = \int_\Omega vw$ and for $v, w \in H^1(\Omega)$ we set $a(v, w) = \int_\Omega \nabla v \cdot \nabla w$. In the remainder of this section Ω will be a C^2 domain in R^2 or an open interval in R^1.

If $\theta \in H^1(\Omega)$, one has the convex subset $K(\theta) \subset H^1(\Omega)$ consisting of all $v \in H^1(\Omega)$ satisfying

$$\theta \leqslant v \qquad \text{a.e. (almost everywhere) on } \Omega.$$

Note if $\theta \leqslant 0$ a.e. on Γ, $K_0(\theta) \equiv K(\theta) \cap H_0^1(\Omega)$ is nonvoid. Let $f \in L^2(\Omega)$, $u_0 \in H_0^1(\Omega) \cap W_\infty^2(\Omega)$, and $g, \psi \in C^2(\bar{\Omega})$ with $\theta \equiv \psi - g \leqslant 0$ on Γ. Then a general type of parabolic inequality is to find $u(x, t)$ such that for almost all $t > 0$,

(5a) $u(x, t) \in K_0(\theta)$

(5b) $(v - u, u_t) + a(v - u, u) \geqslant (v - u, f)$ for $v \in K_0(\theta)$

(5c) $u(x, 0) = u_0(x)$.

Under the assumptions we have imposed, one has (Brézis [3]) :

(6a) $u : [0, \infty) \to H_0^1(\Omega)$ is continuous

(6b) $u(t) \in H^2(\Omega)$ and $u_t \in L^\infty(\Omega)$ for each $t > 0$

and for almost all $t > 0$

$$u_t = u_{xx} + f \qquad \text{a.e. on } \{ x/u > \theta \}, \tag{6c}$$

and

$$u_t = 0 \qquad \text{a.e. on } \{ x/u = \theta \} \tag{6d}$$

Weaker hypotheses on the data yield weaker results on the regularity of u [3], [5].

The problem (5) corresponds to (1) if we let $c = u + g$ where : $g \in C^2(\bar{J})$ and $g(0) = 1/2$, $g(b) = 0$; $\psi = 0$; $f = -1 + g_{xx}$, and $u_0 = c_0 - g$ ([10]). Indeed, noting the results (6), (1) is equivalent to the following variational inequality with inhomogeneous boundary data : for almost all $t > 0$

(7a) $c(x, t) \in K(\psi)$ and $c(x, t) - g(x) \in H_0^1(\Omega)$

(7b) $(w - c, c_t) + a(w - c, c) \geqslant (w - c, f)$ for all w such that $w \in K(\psi)$ and $w - g \in H_0^1(\Omega)$

(7c) $c(x, 0) = c_0(x)$

where $\psi = 0$, $g(0) = 1/2$ and $g(b) = 0$, and $f = -1$.

In analogy with (3), (4), we define the truncation method for (7) as follows. Given (approximate) solution values C^n at time t^n, one obtains intermediate values C^{n+1} at time $t^{n+1} = t^n + \Delta t$ by advancing

$$c_t = c_{xx} + f \tag{8a}$$

$$c = g \text{ on } \Gamma \tag{8b}$$

$$c(t = t^n) = C^n \tag{8c}$$

from t^n to t^{n+1} by any appropriate numerical scheme, and then for each grid point p,

$$C^{n+1}(p) = \max (\psi(p), \tilde{C}^{n+1}(p)). \tag{9}$$

Note in the numerical implementation, f and g can just as well depend on t. For the sake of simplicity we henceforth take $g = 0$.

IV. NUMERICAL EXPERIMENTS WITH A PARABOLIC PROBLEM

Numerical calculations were done for the parabolic problem (7) with data

(10a) $\Omega = (0, 1)$

(10b) $g = 0$

(10c) $\psi = 2x(1 - x)$

(10d) $f(x, t) = (40xt - 20x - 40) \exp (x + t^2 - 1) - 40xt + 44.$

The initial data is $u(x, 0)$ where for $t \in [0, 1]$ the exact solution is

$$u(x, t) = \begin{cases} 20x(-.x - t^2 + \exp(x + t^2 - 1)) + \psi(x), & 0 \leqslant x \leqslant 1 - t^2, \\ \psi(x) & , \quad 1 - t^2 \leqslant x \leqslant 1 \end{cases} \tag{11}$$

In order to advance (8) forward in time, we used the finite element method with the subspace $S_{\Delta x}$ of piecewise linear functions on a uniform Δx mesh, and Crank-Nicolson or a purely implicit method in time (e.g., [4] or [18]). The values of the approximate solution U at the node points were then found using (9). The true solution and some approximate solution values are plotted in Figure 1.

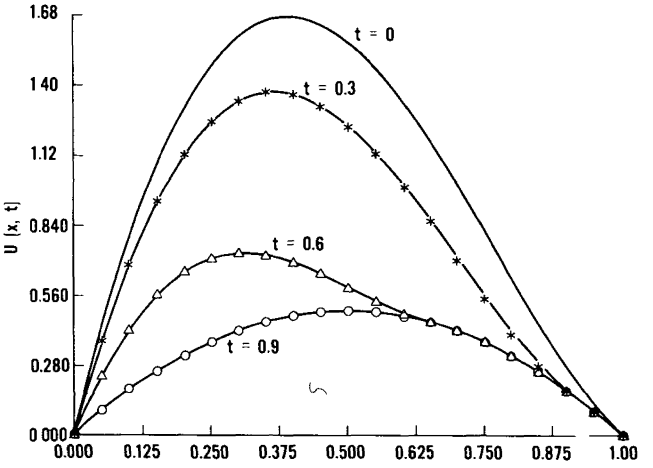

Figure 1

Exact solution (solid lines); and approximate solution values (points) obtained using the truncation method. Linear finite elements and Crank-Nicolson in time were used with

$$\Delta x = .05, \qquad \Delta t = .00375$$

A similar family of numerical methods for the parabolic problem has been investigated by Lions [11]. Instead of the pointwise operations (9), C^{n+1} is set equal to the L^2 projection of $\tilde{C}^{n+1}$ into $\{ v \in S_{\Delta x} / v(p) \geqslant \psi(p)$ at each node $p \}$. A stability result and weak convergence are demonstrated for this method in [11].

At each time step we calculated the L^2 error as

$$\left\{ \sum_{\substack{\text{elements} \\ [x_i, x_{i+1}]}} \int_{x_i}^{x_{i+1}} (u - U)^2 \right\}^{1/2}$$

using four point Gaussian quadrature [19] to compute the integral over each element. If $s \equiv 1 - t^2$ was interior to an element $[x_j, x_{j+1}]$, it was

subdivided into $[x_j, s]$ and $[s, x_{j+1}]$ and the four point Gaussian quadrature was done on each subinterval (hence guaranteeing the integrand to be a smooth function over each quadrature interval). Calculations were done in single precision on a CDC 6400 (14 significant digits).

Tables 1-2 indicate that with both the implicit and Crank-Nicolson methods in time, reducing Δx by a factor of 2 leads to a reduction of error by a factor of 4. This is consistent with a spatial error of $0(\Delta x^2)$. The time discretization error seems to behave like $0(\Delta t)$ after the presence of the « obstacle » $\psi(x)$ makes itself felt (Tables 3-4). The irregular behavior of the error for smaller times is probably due to interaction of the discretization error in advancing (8) with the error from the truncation operation (9). For the specific problem considered in [2], the error due entirely to the truncation (done at every point of Ω, and assuming (8) to be advanced « analytically » on Ω with no error) was shown to be $0(\Delta t)$.

Note that for a fixed $t \in (0, 1)$, the solution (11) as a function of x has a jump in the second derivative at $x = 1 - t^2$ and so it is not in $H^3(\Omega)$. In general, as pointed out by Falk in [6] and Strang in [17], use of higher order elements (beyond quadratics) would not be expected to result in a better *global* spatial error due to the limited smoothness of the solution.

V. THE TRUNCATION METHOD FOR AN ELLIPTIC PROBLEM

We consider the elliptic variational inequality corresponding to the « steady state » of (7), for convenience taking homogeneous boundary data ($g = 0$) :

$$u(x) \in K_0(\psi) \tag{12 a}$$

$$a(v - u, u) \geqslant (v - u, f) \qquad \text{for} \qquad v \in K_0(\psi). \tag{12 b}$$

If $f \in L^2(\Omega)$, $\psi \in C^2(\bar{\Omega})$, and Ω is an open interval in R^1 or a C^2 domain in R^2, the solution of (12) lies in $H^2(\Omega)$ [3]. By Sobolev's inequality u is continuous on $\bar{\Omega}$ [1], and on the open set D where $u > \psi$, u satisfies

$$-u_{xx} = f \qquad \text{a.e.}$$

while [10]

$$-u_{xx} = -\psi_{xx} \qquad \text{a.e. on} \qquad \Omega - D.$$

The numerical solution of (12) by relaxing the corresponding parabolic problem (7) toward the steady state using the truncation method will be analyzed. The « initial guess » $u_0(x)$ can be arbitrary and the method considered for the advancing of (8) is the finite element technique with linear elements. In two space dimensions we assume that Ω is convex, so the triangulated region is contained in Ω. The stability and convergence of the truncation method applied in this way will be examined.

Table 1

L^2 Error for $\Delta t = .0001875$, Crank-Nicolson

Time	.00	.15	.30	.45	.60	.75	.90
$\Delta x = .1$	1.566E-2	1.540E-2	1.587E-2	1.176E-2	9.339E-3	6.293E-3	3.994E-3
$\Delta x = .05$	3.922E-3	3.926E-3	3.478E-3	3.164E-3	2.533E-3	1.708E-3	1.014E-3

Table 2

L^2 Error for $\Delta t = .0001875$, Implicit

Time	.00	.18	.36	.45	.63	.72	.90
$\Delta x = .1$	1.566E-2	1.538E-2	1.519E-2	1.194E-2	9.057E-3	7.264E-3	3.997E-3
$\Delta x = .05$	3.922E-3	4.142E-3	3.412E-3	3.259E-3	2.398E-3	1.882E-3	1.023E-3

Table 3

L^2 Error for $\Delta x=.01$, Crank-Nicolson

Time	.0	.2	.4	.5	.6	.7	.8	.9
$\Delta t=.00625$	1.570E-4	2.888E-3	1.096E-2	1.137E-2	9.657E-3	6.985E-3	4.102E-3	1.416E-3
$\Delta t=.003125$	1.570E-4	2.316E-3	6.199E-3	5.910E-3	4.838E-3	3.477E-3	2.069E-3	7.702E-4

Table 4

L^2 Error for $\Delta x=.01$, Implicit

Time	.0	.2	.4	.5	.6	.7	.8	.9
$\Delta t=.00625$	1.570E-4	4.276E-3	1.818E-2	2.052E-2	1.835E-2	1.349E-2	7.802E-3	2.282E-3
$\Delta t=.003125$	1.570E-4	3.345E-3	1.115E-2	1.135E-2	9.507E-3	6.830E-3	4.032E-3	1.405E-3

Let $\varphi_1, \ldots, \varphi_m$ be the piecewise linear basis functions, with node points $P_1, \ldots, P_m$; the Gram or mass matrix is $M_{ij} = (\varphi_i, \varphi_j)$, the stiffness matrix is $K_{ij} = a(\varphi_i, \varphi_j)$, the load vector is $F_j = (\varphi_j, f)$, and the *obstacle vector* is $\Psi_j = \psi(P_j)$. If the approximate solution at time t^n is $U^n(x) = \sum_{j=1}^{m} q_j^n \varphi_j(x)$, we let Q^n be the (column) vector $(q_1^n, \ldots, q_m^n)$. For $V \in R^m$ we write $V \geqslant \Psi$ if $V_j \geqslant \Psi_j$ for $j = 1, \ldots, m$, and for $V \in R^m$ we define $\overline{V} = \sum_{j=1}^{m} V_j \varphi_j \in H_0^1(\Omega)$.

The advance step (8) of the truncation method is then

$$M(\tilde{Q}^{n+1} - Q^n)/\Delta t + K(\alpha \tilde{Q}^{n+1} + (1 - \alpha)Q^n) = F \tag{13}$$

where $\alpha = 1, .5, 0$ for the implicit, Crank-Nicolson, and explicit time discretizations, respectively. Solving for $\tilde{Q}^{n+1}$,

$$\tilde{Q}^{n+1} = (I + \alpha\, \Delta t M^{-1} K)^{-1}(I - (1 - \alpha)\, \Delta t M^{-1} K)Q^n + \\ + (M/\Delta t + \alpha K)^{-1} F \tag{14}$$

where I is the $m \times m$ identity matrix. We denote the operator on R^m taking Q^n into $\tilde{Q}^{n+1}$ by P and the matrix multiplying Q^n in (14) by A. The truncation operation (9) is $Q^{n+1} = T(\tilde{Q}^{n+1})$ where the j component of $T(\tilde{Q}^{n+1})$ is $\max(\psi_j, \tilde{Q}_j^{n+1})$ and so

$$Q^{n+1} = LQ^n \qquad \text{where} \qquad L = T \circ P.$$

If L is shown to be a contraction mapping on R^m with respect to some norm, then Q^n will converge to the fixed point of L which we call Q, and stability follows. Convergence will be evaluated by comparing $\overline{Q}$ with the solution of the Ritz approximating problem

$$\min_{W \geqslant \Psi} a(\overline{W}, \overline{W}) - 2(\overline{W}, f) \tag{15}$$

or

$$\min_{W \geqslant \Psi} W^t K W - 2W^t F \tag{16}$$

which is equivalent to (see e.g. [12])

$$W \geqslant \Psi \quad \text{and for any} \quad V \geqslant \Psi, \quad (V - W)^t K W \geqslant (V - W)^t F. \tag{17}$$

The error estimate

$$\|u - \overline{W}\|_{H_0^1(\Omega)} = 0 \text{ (mesh size)} \tag{18}$$

has been demonstrated by Mosco and Strang [15] and Falk [6]. We note that

the following iterative « truncation type » algorithm is used for direct approximation of (16) (Lions, Trémolières, Glowinski [13], [14]) :

(19 a)

$$\tilde{W}_i^{n+1} = \left(F_i - \sum_{j<i} K_{ij} W_j^{n+1} - \sum_{j>i} K_{ij} W_j^n \right) / K_{ii} \qquad i = 1, ..., m$$

(19 b)

$$W_i^{n+1} = \max (\Psi_i, (1 - \omega) W_i^n + \omega \tilde{W}_i^{n+1}) \qquad i = 1, ..., m, \; 0 < \omega < 2.$$

VI. STABILITY

In this section we prove that L is a contraction mapping under various assumptions. For a uniform Δx mesh on a line it is easy to verify that the matrices M and K (and hence M^{-1}) commute (this is not true for an irregular mesh or in two dimensions). One has

Theorem 1. — Assume M and K commute and impose the usual Euclidean norm on R^m. Then the operator L is a contraction for Δt sufficiently small when $\alpha \in [0, 1/2)$, and for all Δt when $\alpha \in [1/2, 1]$.

Proof. — We use the notation $|V|^2 = |V|_{l^2}^2 = \sum_i V_i^2$. The operator T is then obviously nonexpansive, i.e.,

$$|TV_1 - TV_2| \leqslant |V_1 - V_2| \qquad \text{for} \qquad V_1, V_2 \in R^m,$$

and so

$$|LV_1 - LV_2| \leqslant |PV_1 - PV_2| = |A(V_1 - V_2)| \leqslant \|A\|_{l^2} |V_1 - V_2|.$$

The consideration of P (i.e., the l^2 norm of the matrix A) follows the usual type of stability argument for difference schemes for parabolic equations.

From their definition, M and K are symmetric, and are positive definite since

$$V^t M V = \|\overline{V}\|_{L^2(\Omega)}^2 \qquad \text{and} \qquad V^t K V = a(\overline{V}, \overline{V}) \qquad \text{for} \qquad V \in R^m.$$

Using the assumption that M and K commute, we may simultaneously unitarily diagonalize M^{-1} and K, and find that

$$\|A\|_{l^2} = \max_i \left| \frac{1 - (1 - \alpha) \Delta t k_i / m_i}{1 + \alpha \Delta t k_i / m_i} \right| \tag{20}$$

where k_i, m_i are the eigenvalues (in appropriate order) of K and M respectively. The result follows.

For a uniform Δx mesh on an interval, the stability condition when $\alpha = 0$ is comparable to that on the standard explicit finite difference method which is

$$\Delta t / \Delta x^2 \leqslant 1/2 \tag{21}$$

(e.g. page 189 of [16]). For a uniform grid, M and K have the form

$$M = \frac{\Delta x}{6}\begin{pmatrix} 4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & 4 & 1 \\ & & 1 & 4 \end{pmatrix} \qquad K = \frac{1}{\Delta x}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & 2 & -1 \\ & & -1 & 2 \end{pmatrix}$$

Using the Gerschgorin circle theorem (e.g. [20]), one has

$$k_i \leqslant 4/\Delta x \qquad \text{while} \qquad m_i \geqslant \Delta x/3.$$

The stability condition from (20) is then

$$\Delta t/\Delta x^2 \leqslant 1/6. \tag{22}$$

If the trapezoidal rule $\int_a^b f\,dx \sim (b-a)(f(b)+f(a))/2$ is used to compute M_{ij}, then $M = \Delta x I$ and the left side of (13) is just Δx times the standard difference scheme for $u_t - u_{xx}$, in which case (20) yields (21) for $\alpha = 0$. Similarly with a regular triangularization of a rectangular region (uniform Δx, Δy and all diagonals drawn in the same direction), and the vertex rule to integrate on each triangle abc

$$\int_{abc} f\,dx\,dy \sim (f(a)+f(b)+f(c))\ \text{area}\ (abc)/3,$$

the matrix M becomes $\Delta x\,\Delta y I$. The left side of (13) is then $\Delta x\,\Delta y$ times the usual difference scheme for $u_t - u_{xx} - u_{yy}$ whose stability condition for $\alpha = 0$ is ([16])

$$\Delta t/\Delta x^2 + \Delta t/\Delta y^2 \leqslant 1/2. \tag{23}$$

We extend theorem 1 to the case where M and K do not commute, but under the restrictive hypothesis that M is a (positive definite) diagonal matrix (which is the case with any grid if the trapezoidal or vertex quadrature methods have been used to form M). We now use the inner product

$$(V_1, V_2) = V_1^t M V_2 \qquad \text{for} \qquad V_1, V_2 \in R^m, \tag{24}$$

with which T is again nonexpansive (it is for this, and only this, that we need to assume M is diagonal). It remains to show that for some $l \in (0, 1)$

$$|AV|_M \leqslant l\,|V|_M \qquad \text{for} \qquad V \in R^m, \tag{25}$$

which is a usual way stability is demonstrated for the finite element method [18]. From the theory of pencils of quadratic forms [7], the eigenvalue problem

$$KV = \lambda MV \qquad \text{or} \qquad M^{-1}KV = \lambda V \tag{26}$$

has m linearly independent eigenvectors E_i which are orthogonal with respect to the inner product (24), and the eigenvalues λ_i of (26) are positive. We then have

$$AE_i = \frac{1 - (1 - \alpha)\,\Delta t\lambda_i}{1 + \alpha\,\Delta t\lambda_i} E_i, \tag{27}$$

and so the conclusion of theorem 1 holds using the norm (24).

We note that if the method of Lions [11] is used, in which $\overline{T(V)}$ is then defined to be the L^2 projection of $\overline{V}$ into

$$\left\{ \sum_{j=1}^{m} r_j \varphi_j \;/\; r_j \geqslant \psi_j \right\} \subset L^2(\Omega);$$

it is no longer necessary that M be « lumped » into a diagonal matrix for the conclusion of Theorem 1 to hold. This is because for T thus defined, $T(V)$ is the solution of

$$\min_{S \geqslant \Psi} |S - V|_M^2 \text{ or } \qquad \min_{S \geqslant \Psi} (S, S)_M - 2(S, V)_M$$

and so ([12])

$$|TV_1 - TV_2|_M \leqslant |V_1 - V_2|_M. \tag{28}$$

We also note that with a regular grid, $\alpha = 0$, M lumped into a constant times I, and with $\Delta t = \Delta x^2/2$ in one dimension or $\Delta t = (\Delta x^2 + \Delta y^2)/2$ in two dimensions, (14) becomes the « Jacobi » form of (19 a) : W_j^{n+1} is replaced by W_j^n on the right hand side.

VII. ERROR ESTIMATES

In the case that M has been « lumped » into a (positive definite) diagonal matrix, and the explicit method is used ($\alpha = 0$ in (13)), Q turns out to be exactly the solution W of the Ritz problem (15)-(17). In this case, by (18),

$$\|u - \overline{Q}\|_{H_0^1(\Omega)} = 0 \text{ (mesh size)}. \tag{29}$$

To show that $Q = W$, we use the property that Q is the fixed point of L;

$$Q = T(\tilde{Q}) = T(Q - \Delta t\, M^{-1}KQ + \Delta t\, M^{-1}F), \tag{30}$$

which immediately implies that

$$\begin{aligned} &Q \geqslant \Psi \text{ and} \\ &\text{if } Q_i > \Psi_i \qquad (M^{-1}KQ)_i = (M^{-1}F)_i \\ &\text{if } Q_i = \Psi_i \qquad (M^{-1}KQ)_i \geqslant (M^{-1}F)_i. \end{aligned} \tag{31}$$

The problem (31) is quickly seen to be equivalent to the discrete variational inequality

$$Q \geqslant \Psi \text{ and } (V - Q)^t M^{-1}KQ \geqslant (V - Q)^t M^{-1}F \text{ for all } V \text{ with } V \geqslant \Psi. \tag{32}$$

Because M is diagonal and positive definite, there is no change in the inequalities (31) – or in the equivalent problem (32) – if we replace M by the identity matrix I. Since this replacement leads back to (17), the steady state Q in the approximate parabolic problem is identical with the Ritz approximation W. The same is true for Lions' scheme of projection instead of truncation, since it is equivalently defined by ([11] pages 314-316) :

$$(33)\quad \left(\overline{V} - \bar{Q}^{n+1}, \frac{\bar{Q}^{n+1} - \bar{Q}^{n}}{\Delta t}\right) + a(\overline{V} - \bar{Q}^{n+1}, \bar{Q}^{n}) \geqslant (\overline{V} - \bar{Q}^{n+1}, f)$$

$$\text{for} \quad V \geqslant \Psi.$$

For $\alpha > 0$, and assuming $(M^{-1}K + KM^{-1})$ is positive definite, we have as yet only been able to prove that $|Q - W| \leqslant C_{\Delta x} \cdot \Delta t$, where $C_{\Delta x} \to \infty$ as $\Delta x \to 0$.

ACKNOWLEDGEMENTS

The author thanks Professor Bruce Kellogg for several helpful discussions, and in particular for posing the question of when the truncation method solution is the Ritz solution, and Professor George Fix for his useful suggestions.

REFERENCES

[1] AGMON S., *Lectures on elliptic boundary value problems*, Princeton : D. Van Nostrand Company, Inc. 1965.

[2] BERGER A. E., CIMENT M. and ROGERS J. C. W., *Numerical solution of a diffusion consumption problem with a free boundary*, SIAM J. Num. Anal. 12, 646-672 (1975).

[3] BRÉZIS H., *Problèmes unilatéraux*, J. Math. Pures et Appl. 51, 1-168 (1972).

[4] DOUGLAS J., Jr. and DUPONT T., *Galerkin methods for parabolic equations*, SIAM J. Num. Anal. 7, 575-626 (1970).

[5] DUVAUT G. and LIONS, J. L., *Les inéquations en mécanique et en physique*, Paris : Dunod 1972.

[6] FALK R., *Error estimates for the approximation of a class of variational inequalities*, Math of Comp. 28, 963-971 (1974).

[7] GANTMACHER F. R., *The theory of matrices*, Vol. 1. New York : Chelsea Publishing Company 1959.

[8] HUNT C. and NASSIF N., *Inéquations variationnelles et détermination de la charge d'espace de certains semi-conducteurs*, C. R. Acad. Sc. Paris, A 278, 1409-1412 (1974).

[9] ISAACSON E. and KELLER H. B., *Analysis of numerical methods*, New York : John Wiley & Sons, Inc., 1966.

[10] LEWY H. and STAMPACCHIA G., *On the regularity of the solution of a variational inequality*, Comm. on Pure and Appl. Math. 22, 153-188 (1969).

[11] LIONS J. L., Approximation numérique des inéquations d'évolution, *Constructive Aspects of Functional Analysis* edited by G. Geymonat, II Ciclo 1971-Centro Internazionale Matematico Estivo, Roma (1973).

[12] LIONS J. L. and STAMPACCHIA G., *Variational inequalities*, Comm. on Pure and Appl. Math. 20, 493-519 (1967).

n° mars 1976.

[13] LIONS J L , TREMOLIERES R and GLOWINSKI R , *Methodes generales d'approximation des problemes d inequations stationnaires*, Institut de Recherche d'Informatique et d'Automatique (March 1971)

[14] LIONS J L , TREMOLIERES R and GLOWINSKI R , *Algorithmes d optimisation*, Institut de Recherche d'Informatique et d'Automatique (July 1971)

[15] MOSCO U and STRANG G , *One-sided approximation and variational inequalities*, Bull Amer Math Soc 80, 308-312 (1974)

[16] RICHTMYER R D and MORTON K W , *Difference methods for initial-value problems* New York Interscience Publishers, 1967

[17] STRANG G , *The finite element method-linear and nonlinear applications*, to appear in the Proceedings of the International Congress of Mathematicians, Vancouver, Canada 1974

[18] STRANG G and FIX G , *An analysis of the finite element method*, Englewood Cliffs Prentice-Hall, Inc 1973

[19] STROUD A H and SECREST D , *Gaussian quadrature formulas* Englewood Cliffs Prentice-Hall, Inc 1966

[20] VARGA R S , *Matrix iterative analysis*, Englewood Cliffs Prentice-Hall, Inc 1965