MATHEMATICS OF COMPUTATION, VOLUME 28, NUMBER 126, APRIL 1974, PAGES 393–404

Finite Element Methods for Parabolic Equations

By Miloš Zlámal

Abstract. The initial-boundary value problem for a linear parabolic equation with the Dirichlet boundary condition is solved approximately by applying the finite element discretization in the space dimension and three types of finite-difference discretizations in time: the backward, the Crank-Nicolson and the Calahan discretization. New error bounds are derived.

1. Introduction. A number of years ago, engineers applied the finite element method to the solution of the heat conduction problem. We mention the papers by Visser [7] and by Wilson and Nickell [8]. Their idea is that in the space dimension a finite element discretization is used whereas in time a finite-difference method is applied. Recently, there appeared papers in mathematical journals where these methods were analyzed as well as new methods proposed, some of them of higher order of accuracy, and where error bounds of a different kind were derived. We mention the papers by Douglas and Dupont [3], Hlaváček [5] and Bramble and Thomée [1], [2].

The problem we are considering is the initial-boundary value problem

$$\begin{aligned} \partial u/\partial t &= Lu \quad \text{for } (x, t) \in \Omega \times (0, T), \\ u &= 0 \quad \text{on } \Gamma \times \langle 0, T\rangle, \\ u(x, 0) &= g(x) \quad \text{in } \Omega. \end{aligned} \tag{1.1}$$

Here

$$Lu = \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) - a(x)u \tag{1.2}$$

and $x = (x_1, \cdots, x_N)$ is a point of a bounded domain Ω in Euclidean N-space R^N with a smooth boundary Γ.

At this point, let us introduce some notation. The norm $||\cdot||_{L_2}$ of the space $L_2(\Omega)$ and the scalar product are denoted by $||\cdot||_0$ and $(\cdot\ ,\cdot)_0$, respectively. $H^m = W_2^{(m)}(\Omega)$, $m = 0, 1, \cdots$, denotes the Sobolev space defined by

$$||u||_{H^m} = \left(\sum_{|\nu|\leq m} ||D^\nu u||_0^2\right)^{1/2}.$$

Instead of $||\cdot||_{H^m}$, we write $||\cdot||_m$. $H_0^{\,1}$ is the closure of $\mathfrak{D}(\Omega)$, the set of infinitely differentiable functions with compact support in Ω, in the $||\cdot||_1$-norm.

The finite element discretization is considered in spaces $V_h^{\,p}$ which are finite-

Received March 15, 1973.

AMS (MOS) subject classifications (1970). Primary 65N30.

dimensional subspaces of H_0^1 and which have the following property: For any $u \in H^{p+1} \cap H_0^1$, there exists a function $\hat{u} \in V_h^p$ such that

$$(1.3) \qquad ||u - \hat{u}||_j \leqq Ch^{p+1-j}||u||_{p+1}, \qquad j = 0, 1,$$

C being a constant independent of the small parameter h and of the function u. Such spaces are well known for domains of a special form (see, e.g., the references in [9]). They were constructed in [11] (for p odd) and in the Appendix of [12] for arbitrary two-dimensional domains.

To introduce the first two approximations, we set $u^n = u(x, nk)$, $n = 0, 1, \cdots$, $n \leqq T/k$. Here $k > 0$ is the time increment. Further, we denote by $a(u, v)$ the energy functional of the operator Lu:

$$(1.4) \qquad a(u, v) = \int_\Omega \left[\sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + a(x)uv \right] dx.$$

In the case of backward discretization, the approximate values U^n of the exact solution u^n are the functions from V_h^p, determined, aside from an initial condition, by

$$(1.5) \qquad (U^{n+1} - U^n, \varphi)_0 + ka(U^{n+1}, \varphi) = 0 \quad \forall \varphi \in V_h^p.$$

The Crank-Nicolson discretization gives

$$(1.6) \qquad (U^{n+1} - U^n, \varphi)_0 + \tfrac{1}{2}ka(U^{n+1} + U^n, \varphi) = 0 \quad \forall \varphi \in V_h^p.$$

Although the reader can find the derivation of the defining equations (1.5) and (1.6), e.g. in [3], we briefly describe the method to derive (1.5) and (1.6), respectively. The variational formulation of problem (1.1) is to find, for $t > 0$, the function $u \in H_0^1$ such that, besides the initial condition, it satisfies

$$(1.7) \qquad (\dot{u}, \varphi)_0 + a(u, \varphi) = 0 \quad \forall \varphi \in H_0^1.$$

We approximate $u(x, t)$ by a function $U(x, t) \in V_h^p$:

$$(1.8) \qquad (\dot{U}, \varphi)_0 + a(U, \varphi) = 0 \quad \forall \varphi \in V_h^p.$$

If $U(x, t) = \sum_{i=1}^l \alpha_i(t)v_i(x)$, where the $v_i(x)$ form the basis of V_h^p, then (1.8) represents a linear system of ordinary differential equations for the unknown coefficients $\alpha_i(t)$. We get (1.5) when we solve this system by the simplest implicit one-step method, whereas (1.6) follows by using the trapezoidal method.

For backward discretization, the estimate of the error will be

$$||u^n - U^n||_1 \leqq C(h^p + k), \qquad 0 \leqq n \leqq T/k.$$

For the Crank-Nicolson discretization, we shall prove that

$$||u^n - U^n||_1 \leqq C(h^p + k^2), \qquad 0 \leqq n \leqq T/k.$$

Here C is a constant which does not depend on h, k or n. We want to stress that no restriction is imposed on h or k.

The estimates of Douglas and Dupont [3], which were derived for a nonlinear equation, are in a different norm whereas the order, both in h and k, is the same. Bramble and Thomée [2] consider Galerkin methods with parameters h and k tied together by the relation kh^{-2} = const. Their Theorem 2, when applied to the Crank-

Nicolson discretization, gives error bounds of the same order in k, again in a different norm.

In the last section, we propose a new procedure which we obtain when we solve system (1.8) by the Calahan A-stable third order method (see, e.g., [4]). The defining equations for the approximate values U^n will be

$$\begin{aligned}(W^{n+1}, \varphi)_0 + bka(W^{n+1}, \varphi) &= -ka(U^n, \varphi), \qquad \forall\ \varphi \in V_h^p,\\ (Z^{n+1}, \varphi)_0 + bka(Z^{n+1}, \varphi) &= -ka(U^n, \varphi) + \beta ka(W^{n+1}, \varphi),\\ U^{n+1} &= U^n + \tfrac{1}{4}(3W^{n+1} + Z^{n+1}),\\ b = \tfrac{1}{2}(1 + \tfrac{1}{3}\sqrt{3}), &\qquad \beta = \tfrac{2}{3}\sqrt{3}.\end{aligned} \tag{1.9}$$

We shall be able to prove that

$$\|u^n - U^n\| \leqq C(h^p + k^3), \qquad 0 \leqq n \leqq T/k.$$

Here

$$\|u\|^2 = \|u\|_0^2 + k\|u\|_1^2. \tag{1.10}$$

Hence, it also holds

$$\|u^n - U^n\|_0 \leqq C(h^p + k^3), \qquad 0 \leqq n \leqq T/k.$$

At this time, we have to assume $k \geqq ch$, $c = \text{const} > 0$. This assumption is no real restriction from the computational point of view because our effort must be to choose k of the same magnitude as h. We shall also show that the new procedure compares favorably with the Crank-Nicolson discretization from the computational point of view.

2. Backward and Crank-Nicolson Discretization.

For simplicity, we assume in the following that

$$a_{ij}(x),\ a(x),\ g(x) \in C^\infty(\bar{\Omega}), \qquad \Gamma \in C^\infty. \tag{2.1}$$

Further, we assume that

$$a_{ij} = a_{ji}, \qquad \sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j \geqq \alpha \sum_{i=1}^N \xi_i^2, \qquad \alpha = \text{const} > 0,\ a(x) \geqq 0. \tag{2.2}$$

We state some facts about the solution $u(x, t)$ of (1.1). It is of the form $u(x, t) = \sum_{i=1}^\infty e^{-\lambda_i t} g_i \psi_i(x)$ where λ_i and $\psi_i(x)$ are (positive) eigenvalues and (orthonormal) eigenfunctions, respectively, of the problem

$$-L\psi = \lambda\psi, \qquad \psi|_\Gamma = 0 \tag{2.3}$$

and g_i are the Fourier coefficients of the function $g(x)$. Ladyženskaja proved (see [6, Section 17]) that if $g \in H^m$ and

$$g|_\Gamma = Lg|_\Gamma = \cdots = L^{[(m-1)/2]}g|_\Gamma = 0 \tag{2.4}$$

then $u(x, t) \in H^m$ for $t \geqq 0$. Subsequently, we will need estimates for $\|u^{n+1} - u^n\|_{p+1}$ and $\|u^{n+1}\|_{p+1}$. We obtain them easily by means of two inequalities by Ladyženskaja (see [6, Section 17]). The first one holds for any series $\sum_{i=1}^\infty g_i\psi_i(x)$:

$$(2.5)\qquad \left\|\sum_{i=1}^{\infty} g_i\psi_i(x)\right\|_m^2 \leqq C\sum_{i=1}^{\infty}(\lambda_i^m + 1)g_i^2.^{*}$$

The other one holds for any $g \in H^m$ satisfying (2.4):

$$(2.6)\qquad \sum_{i=1}^{\infty} \lambda_i^m g_i^2 \leqq C\|g\|_m^2$$

(this is a consequence of (17.6) from [6]).

To find the estimate for the difference $\omega = u^{n+1} - u^n$ (under assumption (2.4)), we calculate the Fourier coefficients ω_i. They are equal to $e^{-kn\lambda_i}(e^{-k\lambda_i} - 1)g_i$. As $|e^{-\tau} - 1| \leqq \tau$ for $\tau \geqq 0$, it follows that $|\omega_i| \leqq k\lambda_i|g_i|$. By means of (2.5) and (2.6), we get

$$\|\omega\|_{m-2}^2 \leqq Ck^2\sum_{i=1}^{\infty}\lambda_i^m g_i^2 \leqq Ck^2\|g\|_m^2.$$

Hence

$$(2.7)\qquad \|u^{n+1} - u^n\|_{m-2} \leqq Ck\|g\|_m.$$

In the same way, we find that

$$(2.8)\qquad \|u^{n+1}\|_m \leqq C\|g\|_m.$$

We now introduce two theorems. Only the second one will be proved because the proof of the first one is analogous. In both cases the initial approximation is chosen as follows:

$$(2.9)\qquad U^0 = \hat{g}(x),$$

where $\hat{g}(x) \in V_h^p$ satisfies (1.3), i.e.,

$$(2.10)\qquad \|g - \hat{g}\|_j \leqq Ch^{p+1-j}\|g\|_{p+1}, \qquad j = 0, 1.$$

THEOREM 1. *Let $g(x)$ satisfy* (2.4) *with $m = \max(p + 3, 4)$. Then, for the approximations U^n determined uniquely by* (2.9) *and* (1.5), *it follows that*

$$(2.11)\qquad \|u^n - U^n\|_{H^1} \leqq C(h^p + k)\,\|g\|_{H^m}, \qquad 0 \leqq n \leqq T/k.$$

THEOREM 2. *Let $g(x)$ satisfy* (2.4) *with $m = \max(p + 3, 6)$. Then, for the approximations U^n determined uniquely by* (2.9) *and* (1.6), *it follows that*

$$(2.12)\qquad \|u^n - U^n\|_{H^1} \leqq C(h^p + k^2)\|g\|_{H^m}, \qquad 0 \leqq n \leqq T/k.$$

Proof of Theorem 2. Let us consider the function $\sigma = u^{n+1} - u^n - \frac{1}{2}kL(u^{n+1} + u^n)$. σ belongs to H^{m-2}. We need to estimate $\|\sigma\|_0$. The Fourier coefficients of σ are

$$\begin{aligned}\sigma_i &= [e^{-(n+1)k\lambda_i} - e^{-nk\lambda_i} + \tfrac{1}{2}k\lambda_i(e^{-(n+1)k\lambda_i} + e^{-nk\lambda_i})]g_i \\ &= (1 + \tfrac{1}{2}k\lambda_i)\left[e^{-k\lambda_i} - \frac{1 - \frac{1}{2}k\lambda_i}{1 + \frac{1}{2}k\lambda_i}\right]e^{-nk\lambda_i}g_i.\end{aligned}$$

By expanding in the Taylor series, we find that for τ sufficiently small

[*] In the sequel, C is a generic constant, not necessarily the same in any two places, which does not depend on h, k, n and g.

$$\left|(1 + \tfrac{1}{2}\tau)\left(e^{-\tau} - \frac{1 - \frac{1}{2}\tau}{1 + \frac{1}{2}\tau}\right)\right| \leqq C\tau^3.$$

As the left-hand side is bounded by $C(1 + \frac{1}{2}\tau)$ for $\tau \geqq 0$, this inequality holds for all $\tau \geqq 0$ (with possibly greater constant C). Therefore, $|\sigma_\imath| \leqq Ck^3\lambda_\imath^3|g_i|$ and $\|\sigma\|_0^2 \leqq Ck^6 \sum_{\imath=1}^{\infty} \lambda_i^6 g_i^2$. By (2.6), it follows that

$$\|\sigma\|_0 \leqq Ck^3\|g\|_m. \tag{2.13}$$

Now, for $\varphi \in V_h^p$, we have

$$\begin{aligned}(\sigma, \varphi)_0 &= (u^{n+1} - u^n, \varphi)_0 - \tfrac{1}{2}k(L(u^{n+1} + u^n), \varphi)_0 \\ &= (u^{n+1} - u^n, \varphi)_0 + \tfrac{1}{2}ka(u^{n+1} + u^n, \varphi).\end{aligned}$$

Denoting $e^n = u^n - U^n$ and subtracting (1.6) from the last equation, we get

$$(e^{n+1} - e^n, \varphi)_0 + \tfrac{1}{2}ka(e^{n+1} + e^n, \varphi) = (\sigma, \varphi)_0 \qquad \forall \varphi \in V_h^p. \tag{2.14}$$

We choose $\varphi = e^{n+1} - e^n - \psi$, $\psi = \omega - \hat{\omega}$, $\omega = u^{n+1} - u^n$. Certainly, φ belongs to V_h^p because it is equal to $-(U^{n+1} - U^n) + \hat{\omega}$. From (2.7) and (1.3), it follows that

$$\|\psi\|_j \leqq Ch^{p+1-j}\|\omega\|_{p+1} \leqq Ckh^{p+1-j}\|g\|_m, \qquad j = 0, 1. \tag{2.15}$$

Substituting the above value of φ in (2.14), we obtain

$$\begin{aligned}&\|e^{n+1} - e^n\|_0^2 + \tfrac{1}{2}ka(e^{n+1} + e^n, e^{n+1} - e^n) \\ &\qquad = (e^{n+1} - e^n, \psi)_0 + \tfrac{1}{2}ka(e^{n+1} + e^n, \psi) + (e^{n+1} - e^n, \sigma)_0 - (\sigma, \psi)_0.\end{aligned}$$

Denoting $a(u, u)$ by $|u|_1^2$, apply the inequality $|ab| \leqq \frac{1}{2}\epsilon a^2 + b^2/2\epsilon$ (with different values of ϵ), the inequality $|a(u, v)| \leqq |u|_1|v|_1$ and the estimates (2.13), (2.15) repeatedly and obtain the result

$$\begin{aligned}&\|e^{n+1} - e^n\|_0^2 + \tfrac{1}{2}k[|e^{n+1}|_1^2 - |e^n|_1^2] \\ &\qquad \leqq \tfrac{1}{4}\|e^{n+1} - e^n\|_0^2 + Ck^2h^{2(p+1)}\|g\|_m^2 + \tfrac{1}{2}k^2[|e^{n+1}|_1^2 + |e^n|_1^2] + Ck^2h^{2p}\|g\|_m^2 \\ &\qquad\quad + \tfrac{1}{4}\|e^{n+1} - e^n\|_0^2 + Ck^6\|g\|_m^2 + Ck^6\|g\|_m^2 + Ck^2h^{2(p+1)}\|g\|_m^2.\end{aligned}$$

Hence

$$\tfrac{1}{2}\|e^{n+1} - e^n\|_0^2 + \tfrac{1}{2}k(1 - k)\,|e^{n+1}|_1^2 \leqq \tfrac{1}{2}k(1 + k)\,|e^n|_1^2 + Ck^2A$$

where $A = (h^{2p} + k^4)\,\|g\|_m^2$, so that it certainly follows that

$$|e^{n+1}|_1^2 \leqq \frac{1 + k}{1 - k}\,|e^n|_1^2 + CkA.$$

Setting $q = (1 + k)/(1 - k)$ and $n = 0, 1, \cdots$ in the last inequality, we find

$$|e^n|_1^2 \leqq q^n|e^0|_1^2 + Ck\,\frac{q^n - 1}{q - 1}\,A. \tag{2.16}$$

As $q > 1$, it follows that $q^n \leqq (1 + k)^{Tk^{-1}}/(1 - k)^{Tk^{-1}}$ for $n \leqq T/k$. Now $(1 + k)^{Tk^{-1}} \to e^T$ and $(1 - k)^{Tk^{-1}} \to e^{-T}$ for $k \to 0$. Therefore q^n is bounded for $n \leqq T/k$: $q^n \leqq C$. Furthermore, we have $0 < 1/(q - 1) < 1/2k$. From (2.16), it follows that

$$|e^n|_1^2 \leqq C|e^0|_1^2 + CA$$

and with respect to (2.10) (notice that $|e^0|_1 = |g - \hat{g}|_1 \leqq C\|g - \hat{g}\|_1$):

$$|e^n|_1^2 \leqq C(h^{2p} + k^4)\|g\|_m^2, \qquad 0 \leqq n \leqq T/k.$$

Taking the square root and realizing that $\|v\|_1 \leqq C|v|_1$ for $v \in H_0^1$, we get the final estimate (2.12).

3. The Calahan Discretization. Let us compare, first, the amount of arithmetic operations which are necessary to carry out the procedures (1.6) and (1.9). As before, $v_i(x)$, $i = 1, \cdots, l$, denote the basis functions of the space V_h^p. Let M be the so-called mass matrix, $M = \{(v_i, v_j)_0\}_{i,j=1}^l$ and K the stiffness matrix, $K = \{a(v_i, v_j)\}_{i,j=1}^l$. If $\mathbf{v} = (v_1, \cdots, v_l)^T$ (the superscript T denotes transposition) and $U^n = (\boldsymbol{\alpha}^n)^T\mathbf{v}$ where $\boldsymbol{\alpha}^n = (\alpha_1^n, \cdots, \alpha_l^n)^T$, then (1.6) leads to the solution of the system $(M + \frac{1}{2}kK)\boldsymbol{\alpha}^{n+1} = (M - \frac{1}{2}kK)\boldsymbol{\alpha}^n$. It can be written in this way:

$$(3.1) \qquad (M + \tfrac{1}{2}kK)\boldsymbol{\vartheta}^n = M\boldsymbol{\alpha}^n, \qquad \boldsymbol{\alpha}^{n+1} = 2\boldsymbol{\vartheta}^n - \boldsymbol{\alpha}^n.$$

Hence the main arithmetic operations necessary to carry out the Crank-Nicolson discretization consist of two parts: (1) We have to compute the matrices M and K and to carry out the forward elimination for the matrix $M + \frac{1}{2}kK$. (2) At every time step we have to multiply the matrix M by a vector and to carry out the back substitution.

The procedure (1.9) leads to the solution of two systems with the same matrix. If we set $W^{n+1} = (\boldsymbol{\beta}^{n+1})^T v$, $Z^{n+1} = (\boldsymbol{\gamma}^{n+1})^T\mathbf{v}$, then these systems are

$$(3.2) \qquad \begin{aligned} (M + bkK)\boldsymbol{\beta}^{n+1} &= -kK\boldsymbol{\alpha}^n, \\ (M + bkK)\boldsymbol{\gamma}^{n+1} &= -kK\boldsymbol{\alpha}^n + \beta kK\boldsymbol{\beta}^{n+1}. \end{aligned}$$

We see that the first part of the main arithmetic operations necessary to carry out the Calahan discretization is the same whereas the second part is only twice as large as in the case of the Crank-Nicolson discretization. This is certainly a favorable result and, as the use of cubic polynomials ($p = 3$) in two-dimensional elliptic problems gives very good numerical results and has other advantages (see [10]), we can expect that the procedure (1.9) will prove itself useful in applications.

We now formulate and prove

THEOREM 3. *Let $g(x)$ satisfy* (2.4) *with* $m = \max(p + 1, 8)$ *and let* $k \geqq ch$, $c =$ const > 0. *Then, for the approximations U^n determined uniquely by* (2.9) *and* (1.9),

$$(3.3) \qquad \|u^n - U^n\| \leqq C(h^p + k^3)\|g\|_m, \qquad 0 \leqq n \leqq T/k.$$

(*The norm* $\|\cdot\|$ *is defined by* (1.10).) *Hence*

$$(3.4) \qquad \|u^n - U^n\|_{L_2} \leqq C(h^p + k^3)\|g\|_m, \qquad 0 \leqq n \leqq T/k.$$

Proof. Multiplying the first and the second equation in (1.9) by $\frac{3}{4}$ and $\frac{1}{4}$, respectively, adding and putting $U^{n+1} - U^n$ for $\frac{1}{4}(3W^{n+1} + Z^{n+1})$, we get

$$(3.5) \qquad \begin{aligned} &(U^{n+1}, \varphi)_0 + bka(U^{n+1}, \varphi) \\ &\qquad = (U^n, \varphi)_0 - (1 - b)ka(U^n, \varphi) + \tfrac{1}{4}\beta ka(W^{n+1}, \varphi) \quad \forall\varphi \in V_h^p. \end{aligned}$$

Let $w^{n+1} \in H_0^1$ be the solution of

$$(3.6) \qquad (w^{n+1}, \varphi)_0 + bka(w^{n+1}, \varphi) = -ka(u^n, \varphi) \quad \forall \varphi \in H_0^1.$$

Consider the function

$$\sigma = u^{n+1} - bkLu^{n+1} - u^n - (1-b)kLu^n + \frac{\beta}{4b}[w^{n+1} - kLu^n].$$

As before, we must estimate $\|\sigma\|_0$. To this end, we compute the Fourier coefficients σ_i of σ. After elementary computations, we get

$$\sigma_i = e^{-nk\lambda_i}(1 + bk\lambda_i)\left\{e^{-k\lambda_i} - \left[1 - \frac{k\lambda_i}{1 + bk\lambda_i} - \frac{\beta}{4}\left(\frac{k\lambda_i}{1 + bk\lambda_i}\right)^2\right]\right\}g_i.$$

By expanding in the Taylor series, we find that, for τ sufficiently small,

$$\left|(1 + b\tau)\left\{e^{-\tau} - \left[1 - \frac{\tau}{1 + b\tau} - \frac{\beta}{4}\left(\frac{\tau}{1 + b\tau}\right)^2\right]\right\}\right| \leqq C\tau^4.$$

By the same argument which we used in the proof of Theorem 2, this inequality is true for all $\tau \geqq 0$. Therefore $|\sigma_i| \leqq Ck^4\lambda_i^4$ and

$$(3.7) \qquad \|\sigma\|_0 \leqq Ck^4 \|g\|_m.$$

With respect to (3.6), we have that, for $\varphi \in V_h^p$,

$$\begin{aligned}(\sigma, \varphi)_0 &= (u^{n+1}, \varphi)_0 + bka(u^{n+1}, \varphi) - (u^n, \varphi)_0 + (1-b)ka(u^n, \varphi) \\ &\quad + \frac{\beta}{4b}[(w^{n+1}, \varphi)_0 + ka(u^n; \varphi)] \\ (3.8) \qquad &= (u^{n+1}, \varphi)_0 + bka(u^{n+1}, \varphi) - (u^n, \varphi)_0 + (1-b)ka(u^n, \varphi) \\ &\quad - \frac{\beta}{4}ka(w^{n+1}, \varphi).\end{aligned}$$

We denote $\eta^{n+1} = w^{n+1} - W^{n+1}$, subtract (3.5) from (3.8) and the first equation (1.9) from (3.6). We get

$$\begin{aligned}(3.9) \qquad &(e^{n+1}, \varphi)_0 + bka(e^{n+1}, \varphi) \\ &= (e^n, \varphi)_0 - (1-b)ka(e^n, \varphi) + \tfrac{1}{4}\beta ka(\eta^{n+1}, \varphi) + (\sigma, \varphi)_0 \quad \forall \varphi \in V_h^p,\end{aligned}$$

$$(3.10) \qquad (\eta^{n+1}, \varphi)_0 + bka(\eta^{n+1}, \varphi) = -ka(e^n, \varphi) \quad \forall \varphi \in V_h^p.$$

First we choose $\varphi = e^{n+1} - \psi$, $\psi = \omega - \hat{\omega}$, $\omega = u^{n+1}$. From (2.8) and (1.3), it follows that

$$(3.11) \qquad \|\psi\|_j \leqq Ch^{p+1-j}\|g\|_m, \qquad j = 0, 1.$$

Substituting the above value of φ in (3.9), we obtain

$$\begin{aligned}\|e^{n+1}\|_0^2 + bk|e^{n+1}|_1^2 &= (e^n, e^{n+1})_0 - (1-b)ka(e^n, e^{n+1}) + \tfrac{1}{4}\beta ka(\eta^{n+1}, e^{n+1}) \\ &\quad + (\sigma, e^{n+1})_0 + (e^{n+1}, \psi)_0 + bka(e^{n+1}, \psi) - (e^n, \psi)_0 \\ &\quad + (1-b)ka(e^n, \psi) - \tfrac{1}{4}\beta ka(\eta^{n+1}, \psi) - (\sigma, \psi)_0.\end{aligned}$$

Making use of the inequality $|ab| \leqq \frac{1}{2}\delta a^2 + b^2/2\delta$ with various values of δ, the inequality $|a(u, v)| \leqq |u|_1 \, |v|_1$ and the estimates (3.7) and (3.11), we get

$$\begin{aligned}||e^{n+1}||_0^2 + bk|e^{n+1}|_1^2 \leqq {} & \tfrac{1}{2}||e^n||_0^2 + \tfrac{1}{2}||e^{n+1}||_0^2 + \frac{1-b}{2}k|e^n|_1^2 + \frac{1-b}{2}k|e^{n+1}|_1^2 \\ & + \tfrac{1}{4}\beta k\left[\tfrac{1}{2}b|\eta^{n+1}|_1^2 + \frac{1}{2b}|e^{n+1}|_1^2\right] \\ & + \tfrac{1}{4}k||e^{n+1}||_0^2 + Ck^7||g||_m^2 \\ & + \tfrac{1}{4}ch||e^{n+1}||_0^2 + Ch^{2p+1}||g||_m^2 + \epsilon k|e^{n+1}|_1^2 + \epsilon^{-1}Ckh^{2p}||g||_m^2 \\ & + \tfrac{1}{2}ch||e^n||_0^2 + Ch^{2p+1}||g||_m^2 + \epsilon k|e^n|_1^2 + \epsilon^{-1}Ckh^{2p}||g||_m^2 \\ & + \epsilon k|\eta^{n+1}|_1^2 + \epsilon^{-1}Ckh^{2p}||g||_m^2 + Ck^8||g||_m^2 + Ch^{2(p+1)}||g||_m^2,\end{aligned}$$

where the positive number ϵ will be chosen later. The preceding inequality can be simplified (by using the assumption $k \geqq ch$) in the following way:

$$\begin{aligned}& \tfrac{1}{2}(1-k)||e^{n+1}||_0^2 + k\left(\frac{3b-1}{2} - \frac{\beta}{8b} - \epsilon\right)|e^{n+1}|_1^2 \\ (3.12) \qquad & \leqq \tfrac{1}{2}(1+k)||e^n||_0^2 + \left(\frac{1-b}{2} + \epsilon\right)k|e^n|_1^2 + k\left(\frac{\beta b}{8} + \epsilon\right)|\eta^{n+1}|_1^2 \\ & \quad + (1+\epsilon^{-1})Ck[h^{2p} + k^6]||g||_m^2.\end{aligned}$$

To derive the final inequality, we must estimate the term $|\eta^{n+1}|_1^2$. At this time, we choose $\varphi = \eta^{n+1} - \psi$, $\psi = \omega - \hat{\omega}$, $\omega = w^{n+1}$. The Fourier coefficients ω_i of the function ω are $\omega_i = -k\lambda_i/(1 + bk\lambda_i)e^{-nk\lambda_i}g_i$. Hence $|\omega_i| \leqq |g_i|/b$ and, by (2.5) and (2.6), $||\omega||_{p+1} \leqq C||g||_m$; therefore

$$||\psi||_i \leqq Ch^{p+1-i}||g||_m, \qquad j = 0, 1.$$

Setting $\varphi = \eta^{n+1}$ in (3.10), we have

$$\begin{aligned}||\eta^{n+1}||_0^2 + bk|\eta^{n+1}|_1^2 & = -ka(e^n, \eta^{n+1}) + (\eta^{n+1}, \psi)_0 + bka(\eta^{n+1}, \psi) + ka(e^n, \psi) \\ & \leqq \tfrac{1}{2}bk|\eta^{n+1}|_1^2 + \frac{1}{2b}k|e^n|_1^2 + \tfrac{1}{2}||\eta^{n+1}||_0^2 + Ch^{2(p+1)}||g||_m^2 \\ & \quad + \epsilon\frac{b}{2}k|\eta^{n+1}|_1^2 + \epsilon^{-1}Ckh^{2p}||g||_m^2 + \frac{\epsilon}{2b}k|e^n|_1^2 \\ & \quad + \epsilon^{-1}Ckh^{2p}||g||_m^2;\end{aligned}$$

hence

$$\tfrac{1}{2}b(1-\epsilon)k|\eta^{n+1}|_1^2 \leqq \frac{1}{2b}k(1+\epsilon)\,|e^n|_1^2 + (1+\epsilon^{-1})Ckh^{2p}||g||_m^2;$$

and finally

$$(3.13) \qquad |\eta^{n+1}|_1^2 \leqq b^{-2}\frac{1+\epsilon}{1-\epsilon}|e^n|_1^2 + (1+\epsilon^{-1})Ch^{2p}||g||_m^2.$$

From (3.12) and (3.13), it follows that

$$\begin{aligned}(3.14) \qquad & \tfrac{1}{2}(1-k)||e^{n+1}||_0^2 + C_1(\epsilon)k|e^{n+1}|_1^2 \\ & \leqq \tfrac{1}{2}(1+k)||e^n||_0^2 + C_2(\epsilon)k|e^n|_1^2 + (1+\epsilon^{-1})Ck(h^{2p} + k^6)||g||_m^2\end{aligned}$$

where

$$C_1(\epsilon) = \frac{3b-1}{2} - \frac{\beta}{8b} - \epsilon, \qquad C_2(\epsilon) = \frac{1-b}{2} + \epsilon + \left(\frac{\beta b}{8} + \epsilon\right) b^{-2}\,\frac{1+\epsilon}{1-\epsilon}.$$

Since

$$C_1(0) = \frac{3b-1}{2} - \frac{\beta}{8b} > 0.43, \qquad C_2(0) = \frac{1-b}{2} + \frac{\beta}{8b} < 0.29,$$

we can choose $\epsilon = \epsilon_0 > 0$ such that $C_1(\epsilon_0) \geqq C_2(\epsilon_0)$. We set $C_1(\epsilon_0) = \frac{1}{2}\gamma$ and obtain, from (3.14),

$$(1-k)\|e^{n+1}\|_0^2 + \gamma k|e^{n+1}|_1^2 \leqq (1+k)\|e^n\|_0^2 + \gamma k|e^n|_1^2 + Ck(h^{2v} + k^6)\|g\|_m^2$$

and further

$$(1-k)[\|e^{n+1}\|_0^2 + \gamma k|e^{n+1}|_1^2] \leqq (1+k)[\|e^n\|_0^2 + \gamma k|e^n|_1^2] + Ck(h^{2v} + k^6)\|g\|_m^2.$$

For the moment, let $\|\cdot\|^2$ be defined not by (1.10) but by $\|\cdot\|^2 = \|\cdot\|_0^2 + \gamma k|\cdot|_1^2$. Then the last inequality can be written in the following way:

$$\|e^{n+1}\|^2 \leqq q\|e^n\|^2 + Ck(h^{2v} + k^6)\|g\|_m^2,$$

where again $q = (1+k)/(1-k)$. Repeating the argument used at the end of the proof of Theorem 2, we get

$$\|e^n\|^2 \leqq C(h^{2v} + k^6)\|g\|_m^2, \qquad 0 \leqq n \leqq T/k.$$

Therefore

$$\|e^n\|_0^2 \leqq C(h^{2v} + k^6)\|g\|_m^2, \qquad |e^n|_1^2 \leqq C(h^{2v} + k^6)\|g\|_m^2$$

and (3.3) follows immediately.

Concluding Remark. The discretizations in time introduced here are derived from A-stable methods for solving ordinary differential equations. The approximation of the function $e^{-\tau}$ used in the first two procedures is that of Padé approximation. This is not the case in the third procedure.

We know that the q-stage implicit Runge-Kutta method of order $2q$ is A-stable (see, e.g., [4]). The approximation of the function $e^{-\tau}$ used in this method is Padé diagonal approximation. One can apply this class of methods to solving Eq. (1.8). For $q = 2$ (see [4, p. 39]), one gets the following scheme:

$$\left.\begin{aligned} &(W^{n+1}, \varphi)_0 + \beta_{11}ka(W^{n+1}, \varphi) + \beta_{12}ka(Z^{n+1}, \varphi) = -ka(U^n, \varphi), \\ &\beta_{21}ka(W^{n+1}, \varphi) + (Z^{n+1}, \varphi)_0 + \beta_{22}ka(Z^{n+1}, \varphi) = -ka(U^n, \varphi), \end{aligned}\right\} \ \forall \varphi \in V_h^v,$$

$$U^{n+1} = U^n + \tfrac{1}{2}(W^{n+1} + Z^{n+1}),$$

$$\beta_{11} = \beta_{22} = \frac{1}{4}, \qquad \beta_{12} = \frac{1}{4} - \frac{\sqrt{3}}{6}, \qquad \beta_{21} = \frac{1}{4} + \frac{\sqrt{3}}{6}.$$

It is questionable whether this procedure, even with its fourth order accuracy, will prove to be useful in applications because of the much greater number of arithmetic operations necessary for carrying it out.

4. Appendix. 1. Elsewhere, we shall derive error bounds for the procedures introduced here which are uniform for $0 \leqq n < \infty$. E.g., for the Crank-Nicolson discretization, it holds without any restriction on the increments k and h that

$$\max_{0 \leqq n < \infty} \|u^n - U^n\|_0 \leqq C[h^{p+1} + k^2] \lg \frac{1}{k} \|g\|_m \tag{4.1}$$

if g satisfies (2.4) with $m = \max(p + 1, 4)$. Such estimates indicate that good results in long time calculations can be expected. Nevertheless, there is one difference between the Calahan** and backward discretization on the one hand and the Crank-Nicolson discretization on the other hand which speaks for preferring the former methods. The exact solution of problem (1.1) has the property that

$$\|u(x, t)\|_0 \leqq e^{-\lambda_1 t}\|g\|_0, \qquad t \geqq 0, \tag{4.2}$$

for any $g \in L_2(\Omega)$. Here, λ_1 is the smallest (positive) eigenvalue of the operator $-Lu$. We want to point out that the Calahan and backward schemes preserve this asymptotic behavior characteristic of (1.1), i.e., for any $U_0 \in L_2(\Omega)$ and for any k not too large, say $k \leqq 1$, it holds that

$$\|U^n\|_0 \leqq e^{-\alpha_0 nk}\|U^0\|_0, \qquad n = 1, 2, \cdots, \quad \alpha_0 = \text{const} > 0. \tag{4.3}$$

The Crank-Nicolson scheme has the same property if k tends to zero fast enough with respect to h, namely, if $k = h^\alpha$ with $\alpha \geqq 1$. On the other hand, if $k = h^\alpha$, $\alpha < 1$, then (4.3) is not true. More exactly, for any sufficiently small positive k, there exists an U^0 such that, for $n = 1, 2, \cdots$, it holds that

$$\|U^n\|_0 \geqq e^{-\Theta(k)nk}\|U^0\|_0, \qquad 0 \leqq \Theta(k) \leqq Ck^{2(1/\alpha - 1)} \to 0. \tag{4.4}$$

2. We assume the following additional properties of the basis $\{v_1, v_2, \cdots, v_l\}$ of the space V_h^p:

$$\text{(a)} \qquad \|\varphi\|_1^2 \leqq Ch^{-2}\|\varphi\|_0^2 \quad \forall \varphi \in V_h^p,$$

$$\text{(b)} \qquad \|v_i\|_1^2/\|v_i\|_0^2 \geqq c_0 h^{-2}, \qquad c_0 = \text{const} > 0.$$

The finite element subspaces used in applications possess these properties.

Let us consider the Calahan discretization and let $U^n = (\alpha^n)^T \mathbf{v}$ as before. To get α^{n+1}, we have to compute β^{n+1} and γ^{n+1} from (3.2); then $\alpha^{n+1} = \alpha^n + \frac{1}{4}(3\beta^{n+1} + \gamma^{n+1})$. There is a recurrence relation between α^{n+1} and α^n. However, in this relation, the matrix $A = M^{-1}K$ appears which is not symmetric. Therefore, we set

$$\mathbf{a}^n = M^{1/2}\alpha^n .$$

By elementary, though not short computations, we get $\mathbf{a}^{n+1} = R(kS)\mathbf{a}^n$ and

$$\mathbf{a}^n = [R(kS)]^n \mathbf{a}^0 \tag{4.5}$$

where

$$R(\tau) = 1 - \frac{\tau}{1 + b\tau} - \frac{\beta}{4}\left(\frac{\tau}{1 + b\tau}\right)^2, \qquad S = M^{-1/2}KM^{-1/2} \tag{4.6}$$

** In the linear homogeneous case, the Calahan discretization is identical with the Makinson scheme of the third order of accuracy (see J. G. Makinson, "Stable high order implicit methods for the numerical solution of systems of differential equations," *Comput. J.*, v. 11, 1968, pp. 305–310).

(b and β are defined in (1.9)). With respect to the meaning of the matrix M, we easily see that

$$(4.7) \qquad ||U^n||_0 = ||\mathrm{a}^n||.$$

From (4.5), we have

$$(4.8) \qquad ||U^n||_0 \leqq ||R(kS)||^n||U^0||_0, \qquad n = 1, 2, \cdots .$$

Now, S is a symmetric positive definite matrix (M and K are positive definite matrices); hence

$$(4.9) \qquad ||R(kS)|| = \max_{\Lambda} |R(k\Lambda)|$$

where the maximum is taken over all eigenvalues Λ of the matrix S. Let us find bounds for Λ.

If $S\mathrm{a} = \Lambda\mathrm{a}$ and $\varphi = (M^{-1/2}\mathrm{a})^T\mathrm{v}$, then $\Lambda = a(\varphi, \varphi)/||\varphi||_0^2$. As $a(\varphi, \varphi) \geqq \lambda_1||\varphi||_0^2$ $\forall \varphi \in H_0^1$, it certainly holds that $\lambda_1 \leqq \Lambda$. On the other hand, from property (a), it follows that $\Lambda \leqq Ch^{-2}$. Therefore,

$$(4.10) \qquad \lambda_1 k \leqq k\Lambda \leqq Ckh^{-2}.$$

The function $R(\tau)$ is decreasing in the interval $\langle 0, \infty)$, $R(1) = 1$ and $R(\infty) = -[\surd 3 - 1] > -1$. With respect to (4.10), we have

$$\max_{\Lambda} |R(k\Lambda)| \leqq \max_{k\lambda_1 \leqq \tau < \infty} |R(\tau)| \leqq \max[|R(k\lambda_1)|, \surd 3 - 1].$$

First, let $\lambda_1 \leqq 1$. Then $k\lambda_1 \leqq 1$. As $R(\tau)$ is decreasing and $R'(0) = -1$, there exists a $\beta_0 > 0$ such that $|R(\tau)| \leqq e^{-\beta_0\tau}$ for $0 \leqq \tau \leqq 1$. Furthermore, $\surd 3 - 1 = e^{-\beta_1} \leqq e^{-\beta_1 k}$ for $k \leqq 1$, $\beta_1 = -lg[\surd 3 - 1]$. If we set $\beta_2 = \max(\lambda_1\beta_0, \beta_1) > 0$, we have $\max_\Lambda|R(k\Lambda)| \leqq e^{-\beta_2 k}$ and, from (4.8) and (4.9), (4.3) follows with $\alpha_0 = \beta_2$. If $\lambda_1 > 1$, we have $\max_\Lambda|R(k\Lambda)| \leqq \max_{k \leqq \tau < \infty}|R(\tau)|$ and we get (4.3) with $\alpha_0 = \max(\beta_0, \beta_1)$.

The backward scheme leads to (4.5) where $R(\tau) = 1/(1 + \tau)$. The same arguments apply as before.

3. The Crank-Nicolson scheme can also be expressed in the form (4.5) where $R(\tau) = (1 - \frac{1}{2}\tau)/(1 + \frac{1}{2}\tau)$. This function is also a Padé rational approximation of the function $e^{-\tau}$ and, again, $R(\tau)$ decreases in $\langle 0, \infty)$ and $|R(\tau)| < 1$ for $\tau > 0$. If $k = h^\alpha$, $\alpha \geqq 1$, then kh^{-2} does not increase too fast, $kh^{-2} \leqq k^{-1}$. Now,

$$||R(kS)|| = \max_{\Lambda} |R(k\Lambda)| \leqq \max_{k\lambda_1 \leqq \tau \leqq Ckh^{-2}} |R(\tau)| \leqq \max_{k\lambda_1 \leqq \tau \leqq k^{-1}} |R(\tau)|.$$

Because, near $\tau = \infty$, $R(\tau)$ behaves in a similar way as near $\tau = 0$, it is easy to prove (4.3) again.

The situation is different if $k = h^\alpha$, $\alpha < 1$. The reason is that the approximation $R(\tau) = (1 - \frac{1}{2}\tau)/(1 + \frac{1}{2}\tau)$ does not possess the property

$$(4.11) \qquad |R(\infty)| < 1.$$

Let us examine this case. The maximum eigenvalue $\Lambda_{\max}$ is bounded from below by c_0h^{-2} which follows from assumption (b) and the fact that

$$\Lambda_{\max} = \max \frac{\mathrm{a}^T S\mathrm{a}}{||\mathrm{a}||^2} = \max \frac{\alpha^T K\alpha}{\alpha^T M\alpha} = \max_{\varphi \in V_h^p} \frac{a(\varphi,\varphi)}{||\varphi||_0^2}.$$

Therefore,

$$k\Lambda_{\max} \geqq c_0 kh^{-2} = c_0 k^{-1-\delta}, \qquad \delta = 2(1/\alpha - 1), \qquad c_0 = \text{const} > 0.$$

As $R(\tau)$ is decreasing in $\langle 0, \infty)$, we have $||R(kS)|| = \max\{|R(k\Lambda_{\min})|, |R(k\Lambda_{\max})|\}$. Further, for τ small, it holds that

$$0 < R(\tau) \leqq 1 - \tfrac{1}{2}\tau, \qquad |R(1/\tau)| \geqq 1 - 4/\tau.$$

Hence, for k small, we get

$$0 \leqq |R(k\Lambda_{\min})| \leqq R(k\lambda_1) \leqq |R(c_0 k^{-1-\delta})| \leqq |R(k\Lambda_{\max})|$$

and

$$||R(kS)|| = |R(k\Lambda_{\max})| \geqq |R(c_0 k^{-1-\delta})| \geqq 1 - 4c_0 k^{1+\delta} = e^{-\Theta(k)k} \tag{4.12}$$

where

$$\Theta(k) = -\frac{1}{k}\, lg(1 - 4c_0 k^{1+\delta}) = 4c_0 k^{\delta} + O(k^{1+2\delta}).$$

Let a^0 be an eigenvector of the matrix $R(kS)$ belonging to the eigenvalue $R(k\Lambda_{\max})$ and $U^0 = (M^{-1/2}\mathrm{a}^0)^T\mathrm{v}$. Then, by (4.5),

$$\mathrm{a}^n = [R(kS)]^n \mathrm{a}^0 = [R(k\Lambda_{\max})]^n \mathrm{a}^0$$

and, by (4.7), (4.12),

$$||U^n||_0 = |R(k\Lambda_{\max})|^n \, ||U^0||_0 \geqq e^{-\Theta(k)nk} \, ||U^0||_0.$$

Computing Center
Technical University
Obránců míru 21
602 00 Brno, Czechoslovakia

1. J. H. BRAMBLE & V. THOMÉE, "Semidiscrete-least squares methods for a parabolic boundary value problem," *Math. Comp.*, v. 26, 1972, pp. 633–648.

2. J. H. BRAMBLE & V. THOMÉE, "Discrete time Galerkin methods for a parabolic boundary value problem," *Ann. Mat. Pura Appl.* (To appear.)

3. J. DOUGLAS, JR. & T. DUPONT, "Galerkin methods for parabolic equations," *SIAM J. Numer. Anal.*, v. 7, 1970, pp. 575–626. MR **43** #2863.

4. C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

5. J. HLAVÁČEK, "On a semi-variational method for parabolic equations. I, II," *Apl. Mat.*, v. 17, 1972, pp. 327–351; ibid., v. 18, 1973, pp. 43–64.

6. O. A. LADYŽENSKAJA, V. A. SOLONNIKOV & N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, "Nauka", Moscow, 1967; English transl., Transl. Math. Monographs, vol. 23, Amer. Math. Soc., Providence, R.I., 1968. MR **39** #3159a,b.

7. W. VISSER, *A Finite Element Method For the Determination of Non-Stationary Temperature Distribution and Thermal Deformations*, Proc. Conf. Matrix Meth. Struct. Mech., Air Force Inst. of Techn., Wright-Patterson A. F. Base, Ohio, 1965.

8. E. L. WILSON & R. E. NICKELL, "Application of finite element method to heat conduction analysis," *Nuclear Eng. Design*, v. 4, 1966, pp. 276–286.

9. M. ZLÁMAL, "Some recent advances in the mathematics of finite elements," in *The Mathematics of Finite Elements and Applications*, edited by J. R. Whiteman, Academic Press, London, 1972, pp. 59–81.

10. M. ZLÁMAL, "The finite element method in domains with curved boundaries," *Int. J. Numer. Meth. Eng.*, v. 5, 1973, pp. 367–373.

11. M. ZLÁMAL, "Curved elements in the finite element method. I," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 229–240.

12. M. ZLÁMAL, "Curved elements in the finite element method. II," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 347–362.