DUAL EXTREMUM PRINCIPLES FOR THE HEAT EQUATION SOLVED BY FINITE ELEMENT METHODS I

Vera Lucia da Rocha Lopes
and
José Vitório Zago

RELATÓRIO TÉCNICO Nº 09/89

Abstract: In this work we develop a numerical method for approximate solutions of the heat equation, based on the dual extremum principles of Noble and Sewell. We use the finite element method for discretization with cubic $B-$spline functions as basis in x and piecewise linear functions as basis in t. We exhibit a Hilbert space Y, a bilinear form S associated to it and we verify all the conditions of Lax–Milgram's lemma with which we get proof of existence and uniqueness of solution of the formulation we present. We prove also a convergence theorem and make an analysis of the numerical results obtained.

Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Ciência da Computação
IMECC – UNICAMP
Caixa Postal 6065
13.081 – Campinas, SP
BRASIL

Abril – 1989

Dual Extremum Principles for the Heat Equation Solved by Finite Element Methods I

Vera Lucia da Rocha Lopes
and
José Vitório Zago
Universidade Estadual de Campinas
São Paulo, Brazil

Abstract: In this work we develop a numerical method for approximate solutions of the heat equation, based on the dual extremum principles of Noble and Sewell. We use the finite element method for discretization with cubic B–spline functions as basis in x and piecewise linear functions as basis in t. We exhibit a Hilbert space Y, a bilinear form S associated to it and we verify all the conditions of Lax–Milgram's lemma with which we get proof of existence and uniqueness of solution of the formulation we present. We prove also a convergence theorem and make an analysis of the numerical results obtained.

1. Introduction

We develop a method for the approximation to the solution of the mixed problem:

$$(1)\qquad \begin{cases} \frac{\partial u}{\partial t} + b(t)\frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial u}{\partial x}\right] + q(x,t) \\ u(0,t) = u(1,t) = 0, \quad a(x,t) > 0 \qquad \text{for} \quad (x,t) \in [0,1] \times [0,T] \\ u(x,0) = u_0(x) \end{cases}$$

for the heat equation, using the dual extremum principles stablished by Nobel and Sewell [15] for problems with a generalized Hamiltonian structure:

$$(2) \qquad \begin{cases} P^* v = \frac{\partial X}{\partial w} \\ P w = \frac{\partial X}{\partial v} \end{cases}$$

where P and P^* are closed adjoint linear operators, $X(v,w)$ is a functional convex in v, and concave in w and v and w belong to Hilbert spaces, with inner products $(\cdot\,,\cdot)$ and $\langle\cdot\,,\cdot\rangle$.

The theory of partial differential equations assures that, if $a(x,t)$, $b(t)$ and $q(x,t)$ are functions, say, smooth enough, then problem (1) is well–posed.

The dual extremum principles by Noble and Sewell can be stated in the following

Theorem: If $X(v,w)$ is convex in v and concave in w, then any solutions $(\hat{v},\hat{w})$ of (2)

(i) solves the following problems:

a)
$$\begin{cases} \min & J_\alpha(v,w) = J_\alpha(\hat{v},\hat{w}) \\ \text{subject to} & P^* v = \frac{\partial X}{\partial w} \end{cases}$$

where $J_\alpha(v,w) = (w, \frac{\partial X}{\partial w}) - X(v,w)$;

b)
$$\begin{cases} \max & K_\beta(v,w) = K_\beta(\hat{v},\hat{w}) \\ \text{subject to} & P w = \frac{\partial X}{\partial v} \end{cases}$$

where $K_\beta(v,w) = \langle v, \frac{\partial X}{\partial v}\rangle - X(v,w)$

(ii) is such that $J_\alpha(\hat{v},\hat{w}) = K_\beta(\hat{v},\hat{w})$.

Zago [19] has built all the necessary machinery to put equation (1) in the form (2) introducing the adjoint of equation (1):

$$\begin{cases} -\frac{\partial v}{\partial t} - b(t)\frac{\partial v}{\partial x} = \frac{\partial}{\partial x}[a(x,t)\frac{\partial v}{\partial x}] + r(x,t) \\ v(0,t) = v(1,t) = 0, \quad a(x,t) > 0 \qquad \text{for} \quad (x,t) \in [0,1] \times [0,T] \\ v(x,T) = v_0(x) \end{cases}$$

where $v_0(x)$ and $r(x,t)$ may be any square integrable functions, and defining

$$w_1(x,t) = 0.5\,[u(x,t) + v(x,t)]; \quad w_2(x,t) = 0.5\,[u(x,t) - v(x,t)]$$

$$q_1(x,t) = 0.5\,[q(x,t) + r(x,t)]; \quad q_2(x,t) = 0.5\,[u(x,t) - r(x,t)].$$

After several algebraic manipulations he has gotten the formulation of the dual extremum principles for the heat equation:

a) the minimum principle:

$$J_\alpha(w_1,w_2) = 0.5\int_0^1\int_0^T a(x,t)\left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right]dt\,dx$$

$$-\int_0^1\int_0^T q_1(x,t)w_1(x,t)dt\,dx$$

$$+0.5\int_0^1 [w_1^2(x,T) + w_2^2(x,0) - 2v_0(x)w_1(x,T)]dx$$

subject to:

$$\begin{cases} \frac{\partial w_1}{\partial t} + b(t)\frac{\partial w_1}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial w_2}{\partial x}\right] + q_2(x,t) & (V_\alpha 1) \\ w_1(x,0) + w_2(x,0) = u_0(x) & (V_\alpha 2) \\ w_1(1,t) = w_1(0,t) = 0 & (V_\alpha 3) \end{cases}$$

b) the maximum principle:

$$K_\beta(w_1,w_2) = -0.5\int_0^1\int_0^T a(x,t)\left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right]dt\,dx$$

$$+\int_0^1\int_0^T q_2(x,t)w_2(x,t)dt\,dx$$

$$-0.5\int_0^1 [w_1^2(x,T) + w_2^2(x,0) - 2u_0(x)w_2(x,0)]dx$$

subject to:

$$\begin{cases} \frac{\partial w_2}{\partial t} + b(t)\frac{\partial w_2}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial w_1}{\partial x}\right] + q_1(x,t) & (V_\beta 1) \\ w_1(x,T) - w_2(x,T) = v_0(x) & (V_\beta 2) \\ w_2(1,t) = w_2(x,0) = 0 & (V_\beta 3) \end{cases}$$

Our method is based in this formulation. We observe that the mixed problem (1) is now a boundary problem both in x and in t. This fact has motivated us to use finite elements in both variables.

Since $v_0(x)$ and $r(x,t)$ can be any square integrable functions we are going to take $v_0(x) \equiv 0$ and $r(x,t) \equiv 0$. We choose our basis functions satisfying conditions V_3 (α and β) trivially and substitute conditions V_2 (α and β) in the expressions for J_α and K_β.

In section 2 we state the theorems that show the equivalence between our formulation and equation (1). We also exhibit the space Y in which we work and prove it is a Hilbert space. We define a bilinear form $S : Y \times Y \to I\!R$ such that finding max $K_\beta(w_1, w_2)$ is equivalent to finding $\min\{\frac{1}{2}S(\overline{x}, \overline{x}) - \langle b, \overline{x}\rangle\}$ where $\overline{x} = (w_1, w_2) \in Y$ and $b \in Y$ has the initial conditions of the problem.

We show that S satisfies the hypotheses of Lax–Milgram's lemma and this gives us existence and uniqueness of solution of our formulation. Furthermore we get a bound on the solution in terms of the initial conditions $||\overline{u}|| \leq \frac{1}{c}\,||b||$.

Closing the section we prove a convergence theorem for approximated solutions.

In section 3 we present a finite element method in which we take the basis functions satisfying condition $V_\beta 1$. We work with B–spline cubics in x and piecewise linear functions in t. By equating derivatives to zero we transform the problem of finding max $K_\beta(w_1, w_2)$ in solving a linear system $Rx = b$. Convergence is proved.

In section 4 we analyse the numerical results and the problems we had when solving the linear system above and suggest a new way of constructing a finite element method by imposing condition $V_\beta 1$ to be satisfied in a weak form.

2. Theoretical Background

We present here the results for the maximum principle only. The modifications for the minimum principle are straightforward.

Taking $v_0(x) \equiv 0$, $r(x,t) \equiv 0$ and introducing condition $V_\beta 2$ into the functional we have to find max $K_\beta(w_1, w_2)$, where now

$$
\left[\begin{array}{l}
K_\beta(w_1, w_2) = \; -0.5 \int_0^1 \int_0^T a(x,t)\left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right] dt\, dx \\
\qquad\qquad + 0.5 \int_0^1 \int_0^T q(x,t) w_2(x,t) dt\, dx \\
\qquad\qquad - 0.5 \int_0^1 [w_2^2(x,T) + w_2^2(x,0) - 2u_0(x) w_2(x,0)] dx
\end{array}\right. \tag{3}
$$

subject to

$$\text{(3a)} \qquad \begin{cases} \frac{\partial w_2}{\partial t} + b(t)\frac{\partial w_2}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial w_1}{\partial x}\right] + 0.5q(x,t) \\ \\ w_2(1,t) = w_2(0,t) = 0. \end{cases}$$

Theorem 2.1: If $w_1(x,t)$ and $w_2(x,t)$ are smooth enough functions satisfying

$$\begin{cases} \frac{\partial w_2}{\partial t} + b(t)\frac{\partial w_2}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial w_1}{\partial x}\right] + 0.5q(x,t) \\ \\ w_2(1,t) = w_2(0,t) = 0 \\ \\ w_1(1,t) = w_1(0,t) = 0 \end{cases}$$

and (w_1, w_2) maximizes the functional K_β with the restrictions in (3), then $u(x,t) = w_1(x,t) + w_2(x,t)$ solves equation (1).

Proof. To prove this theorem we have just to use the fact that (w_1, w_2) being the maximum of K_β with the restrictions in (3) is equivalent to the derivative of K_β in the direction (φ_1, φ_2) be equal to zero for all $(\varphi_1(x,t),\ \varphi_2(x,t))$ such that

$$\begin{cases} \frac{\partial \varphi_2}{\partial t} + b(t)\frac{\partial \varphi_2}{\partial x} = \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial \varphi_1}{\partial x}\right] + 0.5q(x,t) \\ \\ \varphi_2(1,t) = \varphi_2(0,t) = 0 \\ \\ \varphi_1(1,t) = \varphi_1(0,t) = 0 \end{cases}$$

This occurs iff $u(x,t) = w_1(x,t) + w_2(x,t)$ satisfies equation (1).

To prove existence and uniqueness of solution of our formulation we use Lax–Milgram's lemma which can be formulated in the following way:

Lax–Milgram Lemma: [2], [3]

"Let $\overline{V}$ be a real Hilbert space with inner product $\langle\ ,\ \rangle_{\overline{V}}$ and norm $||u||_{\overline{V}} = \langle u,\ u\rangle_{\overline{V}}^{\frac{1}{2}},\ h \in \overline{V}$.

Let $\overline{a} : \overline{V} \times \overline{V} \to I\!R$ be a bilinear form with the following properties:

(i) there exists a constant β such that

$$|\overline{a}(u,v)| \leq \beta ||u||_{\overline{V}} ||v||_{\overline{V}}, \quad u,\ v \in \overline{V}$$

(ii) there exists a constant $\rho > 0$ such that

$$\overline{a}(u,u) \geq \rho ||u||_{\overline{V}}^2, \quad u \in \overline{V}.$$

Let $G : \overline{V} \to I\!R$ be a linear mapping with the following property:

(iii) there exists a constant δ such that

$$|G(u)| \leq \delta ||u||_{\overline{V}}, \qquad u \in \overline{V}.$$

Then there exists a unique element $\hat{u} \in \overline{V}$ such that

$$\overline{a}(\hat{u}, u) = G(u) \forall\, u \in \overline{V} \quad \text{and} \quad ||\hat{u}||_V \geq \delta/\rho.$$

Further, if $\overline{a}(\,\cdot\,,\,\cdot\,)$ is symmetric, then $\hat{u}$ is the unique minimizer of

$$f(u) = \frac{1}{2}\overline{a}(u,u) - G(u), \quad u \in \overline{V}.$$

The first thing to do is to exhibit the Hilbert space which we call Y.

At a first glance it seems that a natural choice of Y should be a subspace of $H^1 \times H^1$. If this was the case, if we could find a constant $\rho > 0$ such that $\overline{a}(\overline{u}, \overline{u}) \geq \rho ||\overline{u}||^2_{\overline{1}}$. $u \in Y$ then the lemma would imply in the Y–norm of the solution $\overline{u}$ being bounded by the L_2–norm of the initial condition or, more especifically the lemma would imply in

$$\int_0^1 \int_0^T \left[\frac{\partial \overline{u}}{\partial t}(x,t)\right]^2 dt\, dx$$

being bounded by the L_2–norm of the initial condition. However, if we take

$$u(x,0) = \sum_{n=1}^{\infty} a_n \text{sen}\, n\pi x,$$

it is not difficult to verify that such an estimate cannot hold.

First we consider $H^1_x \times H^1_x$ where

$$H^1_x = \{p(x,t) \in L_2(Q) \quad \text{such that} \quad \frac{\partial p}{\partial x}(x,t) \in L_2(Q)\}$$

and $Q = [0,\,1] \times [0,\,T]$. (For $p \in L_2(Q)$, $\frac{\partial p}{\partial x}$ exists if there exists $g \in L_2(Q)$ such that

$$\int_0^1 \int_0^T g(x,t)\psi(x,t) dt\, dx = -\int_0^1 \int_0^T p(x,t)\frac{\partial \psi}{\partial x} dt\, dx$$

for any $\psi \in C^\infty(Q)$, ψ of compact support).

We work with the restriction $V_\beta 1$ in the form

$$\int_0^1 \int_0^T \frac{\partial w_2}{\partial t}\, \varphi(x,t) dt\, dx + \int_0^1 \int_0^T b(t)\frac{\partial w_2}{\partial x}\, \varphi(x,t) dt\, dx =$$

$$= \int_0^1 \int_0^T \frac{\partial}{\partial x}\left[a(x,t)\frac{\partial w_2}{\partial x}\right]\varphi(x,t)\, dx + 0.5 \int_0^1 \int_0^T q(x,t)\varphi(x,t) dt\, dx$$

for any $\varphi \in V$ where $V = \{\varphi \in C^\infty(Q)$ such that $\varphi(0,t) = \varphi(1,t) = 0\}$ which implies in

$$(4) \quad \left[\begin{array}{l} \int_0^1 \int_0^T w_2(x,T)\varphi(x,T)dx - \int_0^1 w_2(x,0)\varphi(x,0)dx \\ - \int_0^1 \int_0^T w_2(x,t)\frac{\partial\varphi}{\partial t}(x,t)dt\,dx + \int_0^1 \int_0^T b(t)\frac{\partial w_2}{\partial x}\varphi(x,t)dt\,dx = \\ = - \int_0^1 \int_0^T a(x,t)\frac{\partial w_1}{\partial x}\,\frac{\partial\varphi}{\partial x}(x,t)dt\,dx + 0.5\int_0^1 \int_0^T q(x,t)\varphi(x,t)dt\,dx \end{array}\right.$$

for any $\varphi \in V$.

We observe that if w_1 and w_2 are smooth functions, then (4) be verified for all $\varphi \in V$ is equivalent to condition $V_\beta 1$.

We now define the space Y as $Y = \overline{W} \subset H_x^1 \times H_x^1$ where $W = \{f = (f_1, f_2) \in H_x^1 \times H_x^1$ that are smooth functions which equal to zero at $x = 0$ and $x = 1$ and satisfy (4) for any $\varphi \in V\}$ and

$$\|f\|_Y = \int_0^1 \int_0^T \left[\left(\frac{\partial f_1}{\partial x}\right)^2 + \left(\frac{\partial f_2}{\partial x}\right)^2 + f_1^2 + f_2^2\right] dt\,dx.$$

This set Y is not a subspace of $H_x^1 \times H_x^1$, then we work with the space Y where $q(x,t) \equiv 0$ and at the end of the section we show how we get the bounds for the case of the definition above.

Now, for $\overline{x} = (w_1, w_2)$ and $y = (\tilde{w}_1, \tilde{w}_2)$, $\overline{x}$, $y \in Y$

$$\langle \overline{x}, y\rangle_Y = \int_0^1 \int_0^T \left[\frac{\partial w_1}{\partial x}\,\frac{\partial \tilde{w}_2}{\partial x} + w_1\tilde{w}_1 + \frac{\partial w_2}{\partial x}\,\frac{\partial \tilde{w}_2}{\partial x} + w_2\tilde{w}_2\right] dt\,dx.$$

Working with the maximum principle we note that the conditions $V_o 3$ are natural conditions which means that we obtain the same result whether we impose $w_1(1,t) = w_1(0,t) = 0$, or not.

We observe that, for $(w_1, w_2) \in Y$, $w_2(x,T)$ and $w_2(x,0)$ in (4) do make sense as this theorem shows:

Theorem 2.2: For any $\varphi \in V$, the function which takes $(g_1, g_2) \in W$ in

$$\int_0^1 g_2(x,\cdot)\varphi(x,\cdot)dx \in C([0,1], \mathbb{R})$$

has a continuous extension to Y.

Proof: Let $(g_1, g_2) \in Y$ be the limit of the sequence (g_1^n, g_2^n) where for any n, $(g_1^n, g_2^n) \in W$.

So, for any $n \in N$ and any $\varphi \in V$,

$$\int_0^1 \int_s^t g_2^n(x,r)\frac{\partial \varphi}{\partial r}(x,r)dr\,dx = \int_0^1 g_2^n(x,t)\varphi(x,t)dx$$

$$\int_0^1 g_2^n(x,s)\varphi(x,s)dx - \int_0^1 \int_s^t \frac{\partial g_2^n}{\partial r}(x,r)\varphi(x,r)dr\,dx.$$

Using (4) and integration by parts in x we find that

$$(5) \quad \left[\begin{array}{l} \int_0^1 g_2^n(x,t)\varphi(x,t)dx - \int_0^1 g_2^n(x,s)\varphi(x,s)dx = \\ = \int_0^1 \int_s^t g_2^n(x,r)\frac{\partial \varphi}{\partial r}(x,r)dr\,dx - \int_0^1 \int_s^t a(x,t)\frac{\partial g_1^n}{\partial x}(x,r)\frac{\partial \varphi}{\partial x}(x,r)dr\,dx \\ \quad - \int_0^1 \int_s^t b(t)\frac{\partial g_2^n}{\partial x}(x,r)\varphi(x,r)dr\,dx. \end{array} \right.$$

So, if (g_1^n, g_2^n) converges to (g_1, g_2) in $H_x^1 \times H_x^1$ we see that the sequence $\int_0^1 g_2^n(x, \cdot)\varphi(x, \cdot)dx$ is equicontinuous and this proves the theorem. □

We need also the result below which we dont prove here:

Lemma 2.1: For any $x \in [0,1]$, the function which takes $u \in C^\infty(Q)$ in $u(x, \cdot) \in L_2[0,T]$ has a continuous extension to a function from H_x^1 in $L_2[0,T]$.

Theorem 2.3: Y is a Hilbert space.

Proof: Let us take a sequence $\{(w_1^n, w_2^n)\}$ in Y that converges to (w_1, w_2) in $H_x^1 \times H_x^1$. Then, for each $n \in N$ we have

$$(6) \quad \left[\begin{array}{l} \int_0^L w_2^n(x,T)\varphi(x,T)dx - \int_0^1 w_2^n(x,0)\varphi(x,0)dx \\ - \int_0^1 \int_0^T w_2^n(x,t)\frac{\partial \varphi}{\partial x}(x,t)dt\,dx + \int_0^1 \int_0^T b(t)\frac{\partial w_2^n}{\partial x}(x,t)\varphi(x,t)dt\,dx \\ \quad = - \int_0^1 \int_0^T a(x,t)\frac{\partial w_1^n}{\partial x}(x,t)\frac{\partial \varphi}{\partial x}(x,t)dt\,dx \quad \text{for any} \quad \varphi \in V. \end{array} \right.$$

Taking $g_2^n(x,y) = w_2^n(x,y) - w_2(x,y)$, $t = T$ and $s = 0$ in (5) we prove that

$$\int_0^1 w_2^n(x,T)\varphi(x,T)dx \quad \text{converges to} \quad \int_0^1 w_2(x,T)\varphi(x,T)dx$$

and

$$\int_0^1 w_2^n(x,0)\varphi(x,0)dx \quad \text{converges to} \quad \int_0^1 w_2(x,0)\varphi(x,0)dx.$$

With similar arguments we prove the other convergences and so we have the proof that (4) is valid for (w_1, w_2).

Further, the same arguments again and Lemma 2.1 show that $w_1(0,t) = w_1(1,t) = w_2(0,t) = w_2(1,t) = 0$.

So $(w_1, w_2) \in Y$ and Y is closed in $H_x^1 \times H_x^1$ which proves the theorem. □

We now define a bilinear form which we call S:

Definition 2.1: Let $\overline{x} = (w_1, w_2) \in Y$ and $y = (\tilde{w}_1, \tilde{w}_2) \in T \cdot S : Y \times Y \to \mathbb{R}$ is defined by

$$S(\overline{x}, y) = \int_0^1 \int_0^T a(x,t)\left[\frac{\partial w_1}{\partial x}\frac{\partial \tilde{w}_1}{\partial x} + \frac{\partial w_2}{\partial x}\frac{\partial \tilde{w}_2}{\partial x}\right]dt\,dx +$$

$$+ \int_0^1 [w_2(x,T)\tilde{w}_2(x,T) + w_2(x,0)\tilde{w}_2(x,0)]dx.$$

S is obviously bilinear, symmetric and positive definite since $S(\overline{x}, \overline{x}) \geq 0$ and equality holds iff $\overline{x} = 0$.

Further

$$K_\beta(w_1, w_2) = -\frac{1}{2}S(\overline{x}, \overline{x}) + \int_0^1 u_0(x)w_2(x,0)dx.$$

Lemma 2.2: There exists a constant $c > 0$ such that $S(\overline{x}, \overline{x}) \geq c\|\overline{x}\|_Y^2$ for any $\overline{x} = (w_1, w_2) \in Y$.

Proof. We have that

$$S(\overline{x}, \overline{x}) \geq \int_0^1 \int_0^T a(x,t)\left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right]dt\,dx$$

$$\geq \alpha \int_0^1 \int_0^T \left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right]dt\,dx$$

(we are supposing $A \geq a(x,t) \geq \alpha > 0$ in Q).

Using now the Poincaré inequality [1] we have that

$$S(\overline{x}, \overline{x}) \geq 0.5\alpha \int_0^1 \int_0^T \left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2 + w_1^2 + w_2^2\right]dt\,dx = 0.5\alpha\|\overline{x}\|_Y^2.$$

So $c = 0.5\alpha > 0$ which proves the lemma.

Lemma 2.3: There exists a constant k such that $|S(\overline{x}, y)| \leq k \ ||\overline{x}||_Y ||y||_Y$.

Proof: Since S is symmetric and $S(\overline{x}, \overline{x}) \geq 0$ for any $\overline{x} \in Y$ it suffices to show for $\overline{x} = y$.

Let $z = (w_1, w_2)$ and let us find a bound to $S(z, z)$ for any $t \in [0, T]$,

$$w_2^2(x,t) = w_2^2(x,s) - 2\int_t^s w_2(x,r)\,\frac{\partial w_2}{\partial r}(x,r)dr.$$

Using $V_\beta 1$,

$$w_2^2(x,t) = w_2^2(x,s) - 2\int_t^s w_2(x,r)\frac{\partial}{\partial x}\Big[a(x,r)\frac{\partial w_1}{\partial x}\Big]dr + 2\int_t^s b(t)\frac{\partial w_2}{\partial x}dr.$$

So,

$$\begin{aligned} T\int_0^1 w_2^2(x,t)dx &= \int_0^1\int_0^T w_2^2(x,s)dx\,ds \\ &- 2\int_0^1\int_0^T\int_t^s w_2(x,r)\frac{\partial}{\partial x}\Big[a(x,r)\frac{\partial w_1}{\partial x}\Big]dx\,dr\,ds \\ &+ 2\int_0^1\int_0^t\int_t^s b(r)\frac{\partial w_2}{\partial x}dx\,dr\,ds. \end{aligned}$$

Using integration by parts in the second parcel and the fact that the third parcel is zero since $w_2(0,t) = w_2(1,t) = 0$ we have

$$\begin{aligned} T\int_0^1 w_2^2(x,t)dx &= \int_0^1\int_0^T w_2^2(x,s)dx\,ds \\ &+2\int_0^1\int_0^T\int_t^s a(x,r)\frac{\partial w_1}{\partial x}(x,r)\frac{\partial w_2}{\partial x}(x,r)dx\,dr\,ds \\ &\leq \int_0^1\int_0^T w_2^2(x,s)dx\,ds + 2\Big\{\int_0^1\int_0^T\int_t^s\Big[a(x,r)\frac{\partial w_1}{\partial x}(x,r)dx\,dr\,ds\Big\}^{1/2}\cdot \\ &\cdot\Big\{\int_0^1\int_0^T\int_t^s\Big[\frac{\partial w_2}{\partial x}(x,r)\Big\}^2 dx\,dr\,ds\Big\}^{1/2}. \end{aligned}$$

Using the fact that $a(x,t) \leq A$ in Q we get

$$\int_0^1 w_2^2(x,t)dx \le \frac{1}{T}\int_0^1 \int_0^T w_2^2(x,s)dx\,ds +$$

$$+ 2A\Big(\int_0^1\int_0^T \Big(\frac{\partial w_1}{\partial x}\Big)^2 dx\,dr\Big)^2\Big(\int_0^1\int_0^T \Big(\frac{\partial w_2}{\partial x}\Big)^2 dx\,dr\Big)^{1/2}$$

$$\le \frac{1}{T}\int_0^1\int_0^T w_2^2(x,s)dx\,ds + A\Big\{\int_0^1\int_0^T\Big[\frac{\partial w_1}{\partial x}^2(x,r) + \frac{\partial w_2}{\partial x}^2(x,r)\Big]^2 dx\,dr\,ds\Big\}$$

for any $t \in [0,T]$. **Though**

$$S(z,z) = \int_0^1\int_0^T a(x,t)\Big[\Big(\frac{\partial w_1}{\partial x}\Big)^2 + \Big(\frac{\partial w_2}{\partial x}\Big)^2\Big]dt\,dx +$$

$$+ \int_0^1 w_2^2(x,T)dx + \int_0^1 w_2^2(x,0)dx$$

$$\le 3A\int_0^1\int_0^T\Big[\Big(\frac{\partial w_1}{\partial x}\Big)^2 + \Big(\frac{\partial w_2}{\partial x}\Big)^2\Big]dt\,dx + \frac{2}{T}\int_0^1\int_0^T w_2^2(x,s)dz\,dx$$

$$\le k||z||_Y^2 \quad \text{where} \quad k = \max\{3A,\ \frac{2}{T}\}.$$

We observe now that, since $S(\overline{x},y)$ is continuous and $\int_0^1 u_0(x)w_2(x,0)dx$ is linear and continuous, using Riesz' theorem [3] we have $S(\overline{x},y) = \langle \overline{x}, Ry\rangle$ and $\int_0^1 u_0(x)w_2(x,0)dx = \langle b, \overline{x}\rangle$.

So, $K_\beta(w_1,w_2) = K_\beta(\overline{x}) = -\frac{1}{2}\,S(\overline{x},\overline{x}) + \langle b, \overline{x}\rangle = -\frac{1}{2}\langle \overline{x}, R\overline{x}\rangle + \langle b, \overline{x}\rangle$ and then, finding $\max K_\beta(\overline{x})$ is equivalent to finding $\min f(\overline{x})$ where

$$f(\overline{x}) = \frac{1}{2}\,\langle \overline{x}, R\overline{x}\rangle - \langle b, \overline{x}\rangle = \frac{1}{2}\,S(\overline{x},\overline{x}).$$

With Theorem 2.3 and lemmas 2.2 and 2.3 we can use Lax–Milgram's lemma.

Let $G : Y \to \mathbb{R}$ be linear and bounded. Then, by Riesz' theorem there exists $z \in Y$ such that $G(y) = \langle y, z\rangle$ for any $y \in Y$ and $||G|| = ||z||_Y$. On the other hand Lax–Milgram's lemma assures that there exists just one $\widehat{u} \in Y$ such that $S(\widehat{u},y) = G(y)$ for any $y \in Y$ and $||\widehat{u}||_Y \le \frac{1}{c}\,||G||$.

So $\langle y,z\rangle = G(y) = S(\widehat{u},y) = \langle y, R\widehat{u}\rangle$ for $y \in Y$.

Lax–Milgram's lemma also says that this $\widehat{u}$ is the only minimizer of f, since S is symmetric. So $\widehat{u}$ is the only solution of $\max K_\beta$ we used here and the only solution of $R\overline{x} = b$.

It can be proved that $w_1(x,t)$ and $w_2(x,t)$ are functions smooth enough so that $\widehat{u}$ solves the heat equation (1).

We observe that $||\tilde{u}||_Y \leq \frac{1}{c}\ ||G|| = \frac{1}{c}\ ||b||$ gives a bound to the solution in terms of the initial condition of problem (1).

The Non–Homogeneous Case

We work now with the case where $q(x,t) \neq 0$. Let $\tilde{w}_1(x,t)$ be the solution of

$$\begin{cases} \frac{\partial}{\partial x}[a(x,t)\frac{\partial \tilde{w}_1}{\partial x}] + 0.5\, q(x,t) = 0, \quad a(x,t) > 0 \quad \text{in} \quad Q \\ \\ \tilde{w}_1(0,t) = \tilde{w}_1(1,t) = 0. \end{cases}$$

It is easy to see that

$$\tilde{w}_1(x,t) = \frac{0.5 \int_0^1 \int_0^y \frac{q(s,t)}{a(y,t)}}{\int_0^1 \frac{dr}{a(r,b)}} \int_0^x \frac{dr}{a(r,t)} - 0.5 \int_0^y \frac{1}{a(y,t)} \Big(\int_0^x q(s,t) ds \Big) dy$$

Considering $\hat{w}_1 = w_1 - \tilde{w}_1$ we see that Lax–Milgram's lemma can be applied to max $K_\beta(\hat{w}_1, w_2)$ so that $||(\hat{w}_1, w_2||_Y \leq \frac{1}{c}\ ||b||$ where now

$$\langle b, \overline{x} \rangle = \int_0^1 \int_0^T \Big[\frac{\partial w_1}{\partial x} \cdot \frac{\partial \tilde{w}_1}{\partial x} \Big] dt\, dx + \int_0^1 u_0(x) w_2(x,0) dx.$$

The Convergence

The finite element method uses approximations to the solution $\hat{u}$ of max K_β in finite dimensional subspaces of Y, Y_n with dimension n. We call these approximations in Y_n, u_n and we want to know under what conditions $\{u_n\}$ converges to $\hat{u}$ when n goes to infinity.

Theorem 2.4: If $\{Y_n\}$ is a sequence of subspaces of Y such that $B = \cup_n Y_n$ is dense in Y, then $\{u_n\} \to \hat{u}$.

Proof: We work with $P(w_1, w_2) = -K_\beta(w_1, w_2)$ and $\max K_\beta(w_1, w_2) = -\min P(w_1, w_2)$ where $P(\overline{x}) = \frac{1}{2}\langle \overline{x}, R\overline{x} \rangle - \langle b, \overline{x} \rangle$.

So $\hat{u} = \min P(\overline{x})$ in Y and $u_n = \min P(\overline{x})$ in Y_n. It is not difficult to prove that for any $v \in Y$,

$$P(v) - P(\hat{u}) \geq \frac{1}{2} c ||v - \hat{u}||_Y^2. \tag{7}$$

Also, since B is dense in Y, for any $\varepsilon > 0$ there exists $v_n \in Y_n$ such that $||\hat{u} - v_n||_Y < \varepsilon$.

So we find that

$$||P(v_n) - P(\hat{u})|| \leq \frac{1}{2}\,|S(\hat{u},\hat{u}) - S(v_n, v_n)| + ||b||_Y\,||\hat{u} - v_n||_Y$$

$$\leq \left[\frac{k}{2}\,||\hat{u} + v_n||_Y + ||b||_Y\right]\,||\hat{u} - v_n||_Y.$$

But $||v_n||_Y \leq ||\hat{u}||_Y + ||\hat{u} - v_n||_Y \leq ||\hat{u}||_Y + \varepsilon$.

So $||\hat{u} + v_n||_Y \leq 2\,||\hat{u}||_Y + \varepsilon \leq 2\,\frac{1}{c}\,||b||_Y + \varepsilon$ by Lax–Milgram's lemma.

This implies in $||P(v_n) - P(\hat{u})|| \leq M\,||\hat{u} - v_n||_Y < M\,\varepsilon$ where $M = (\frac{k}{c} + 1)||b||_Y + k\varepsilon$.

Since u_n is the minimum of $P(x)$ in Y_n, $P(u_n) \leq P(v_n)$ and since $\hat{u}$ is the minimum of $P(x)$ in Y, $P(\hat{u}) \leq P(u_n)$ and then $P(u_n) - P(\hat{u}) < M\,\varepsilon$.

So, using (7) for u_n and $\hat{u}$, we have $K = \frac{1}{2}\,c\,||u_n - \hat{u}||_Y^2 \leq P(u_n) - P(\hat{u}) < M\,\varepsilon$ which proves the convergence and gives the error estimate:

$$||u_n - \hat{u}||_Y \leq \sqrt{\frac{M\varepsilon}{K}}.$$

Finally we observe that there are other proofs of existence and uniqueness of solution for equation (1). We need the one we gave here because this method is the constructive method we use for numerical approximation. Furthermore the convergence of the method is associated to the bounds we found here.

3. A Finite Element Method.

We present a finite element method to approximate the solution of the particular heat equation:

$$(8)\qquad \begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} \\ u(0,t) = u(1,t) = 0 \quad (x,t) \in Q = [0,1] \times [0,T] \\ u(x,0) = u_0(x) \end{cases}$$

For equation (8) we have to find max $K_\beta(w_1, w_2)$ where

$$(9)\qquad \begin{cases} K_\beta(w_1,w_2) \;=\; -0.5\int_0^1\int_0^T\left[\left(\frac{\partial w_1}{\partial x}\right)^2 + \left(\frac{\partial w_2}{\partial x}\right)^2\right]dt\,dx \\ \qquad\qquad -0.5\int_0^1[w_2^2(x,T) + w_2^2(x,0) - 2u_0(x)w_2(x,0)]dx \end{cases}$$

subject to:

$$\text{(9a)} \qquad \begin{cases} \dfrac{\partial w_2}{\partial t} = \dfrac{\partial^2 w_1}{\partial x^2} & (V_\beta 1) \\ \\ w_2(0,t) = w_2(1,t) = 0 & (V_\beta 3) \end{cases}$$

Since, for condition $V_\beta 1$ we need second derivative in x, we take for basis functions, cubic B–splines in x, $\psi_i(x)$ and piecewise linear functions in t, $\phi_j(t)$.

We choose

$$w_1(x,t) = \sum_{i=1}^{v}\sum_{j=1}^{M} c_{ij}\phi_j'(t)\psi_i(x) \quad \text{with} \quad \psi_i(0) = \psi_i(1) = 0 \quad \text{for every} \quad i.$$

Using $V_\beta 1$ and integrating in t,

$$w_2(x,t) = \sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\phi_j(t)\psi_i''(x).$$

Taking the expressions of $w_1(x,t)$ and $w_2(x,t)$ in (9) we have now to maximise $K_\beta(c_{ij})$ where

$$\begin{aligned} K_\beta(c_{ij}) = \; & - \; 0.5 \int_0^1 \int_0^T \left[\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\phi_j'(t)\psi_i'(x)\right]^2 dt\,dx \\ & - \; 0.5 \int_0^1 \int_0^T \left[\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\phi_j(t)\psi_i'''(x)\right]^2 dt\,dx \\ & - \; 0.5 \int_0^1 \left[\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\phi_j(T)\psi_i''(x)\right]^2 dx \\ & - \; 0.5 \int_0^1 \left[\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\phi_j(0)\psi_i''(x)\right]^2 dx \\ & + \; \int_0^1 u_0(x)\left[\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\phi_j(0)\psi_i''(x)\right] dx \qquad \text{with no restrictions.} \end{aligned}$$

We observe that M being the number of discretization points we have in t, $\Delta t = 1/(M-1)$. Let us call NPX the number of discretization points in x, so

$\Delta x = 1/(NPX-1)$. Note that we have non null contribuitions of the B-splines $\psi_0(x)$, $\psi_1(x), \ldots, \psi_{NPX}(x)$, $\psi_{NPX+1}(x)$, $NPX+2$ functions though. But $\psi_0(0) \neq 0$, $\psi_1(0) \neq 0$, $\psi_2(0) \neq 0$ and $\psi_{NPX-1}(1) \neq 0$, $\psi_{NPX}(1) \neq 0$, $\psi_{NPX+1}(1) \neq 0$ so we work with the $N = NPX - 2$ functions:

$$\tilde{\psi}_1(x) = \psi_2(x) - \psi_0(x)$$

$$\tilde{\psi}_2(x) = \psi_3(x)$$

$$\vdots$$

$$\tilde{\psi}_{N-1}(x) = \psi_{NPX-2}(x)$$

$$\tilde{\psi}_N(x) = \psi_{NPX}(x) - \psi_{NPX+1}(x)$$

which, for simplicity we call again $\psi_1, \psi_2, \ldots, \psi_N$.

Now

$$x_1 = \Delta x,\ x_2 = 2\Delta x, \ldots, x_N = N\Delta x$$

and

$$\psi_i(0) = \psi_i(1) = \psi_i''(0) = \psi_i''(1) = 0, \qquad 1 \leq i \leq N.$$

We observe that the B-spline functions $S_3(z)$ defined in $[-2, 2]$ are

$$S_3(z) = \begin{cases} 0, & z \leq -2 \\ (z^3 + 6z^2 + 12z + 8)/6, & -2 \leq z \leq -1 \\ (-3z^3 - 6z^2 + 4)/6, & -1 \leq z \leq 0 \\ (3z^3 - 6z^2 + 4)/6, & 0 \leq z \leq 1 \\ (-z^3 - 6z^2 - 12z + 8)/6, & 1 \leq z \leq 2 \\ 0, & z \geq 2 \end{cases}$$

and so, to work in $[x_{i-2}, x_{i+2}]$ with $x_i = i\Delta x$, $1 \leq i \leq N$, we have

$$\psi_i(x) = S_3(x/\Delta x - i)$$

$$\psi_i'(x) = (1/\Delta x) S_3'(x/\Delta x - i)$$

$$\psi_i''(x) = (1/\Delta x^2) S_3''(x/\Delta x - i)$$

$$\psi_i'''(x) = (1/\Delta x^3) S_3'''(x/\Delta x - i).$$

And the piecewise linear functions in $[t_{j-1}, t_{j+1}]$ are

$$\phi_j(t) = \begin{cases} 0, & t < t_{j-1} \\ (t - t_{j-1})/\Delta t, & t_{j-1} \le t \le t_j \\ (t_{j+1} - t)/\Delta t, & t_j \le t \le t_{j+1} \\ 0, & t \ge t_{j+1} \end{cases}$$

To find max $K_\beta(c_{ij})$ we equate the derivative of $K_\beta(c_{ij})$ with respect to $c_{\ell m}$ to zero for every ℓ, m, $1 \le \ell \le M$, $1 \le m \le N$ which turns out to be solving a linear system of equations with $M\,N$ unknowns and $M\,N$ equations:

$$-\int_0^1 \int_0^T \Big[\sum_{i=1}^N \sum_{j=1}^M c_{ij} \phi_j'(t) \psi_i'(x)\Big] \phi_\ell'(t) \psi_m'(x) dt\, dx$$

$$-\int_0^1 \int_0^T \Big[\sum_{i=1}^N \sum_{j=1}^M c_{ij} \phi_j(t) \psi_i'''(x)\Big] \phi_\ell(t) \psi_m'''(x) dt\, dx$$

$$-\int_0^1 \Big[\sum_{i=1}^N \sum_{j=1}^M c_{ij} \phi_j(T) \psi_i''(x)\Big] \phi_\ell(T) \psi_m''(x) dx$$

$$-\int_0^1 \Big[\sum_{i=1}^N \sum_{j=1}^M c_{ij} \phi_j(0) \psi_i''(x)\Big] \phi_\ell(0) \psi_m''(x) dx$$

$$+\int_0^1 u_0(x) \phi_\ell(0) \psi_m''(x) dx = 0, \quad 1 \le \ell \le N, \quad 1 \le m \le N.$$

We now substitute the expressions of $\phi_r(t)$ and $\psi_s(x)$ and since it is not difficult we calculate analitically all the integrals involved and finally we obtain the system $Rx = v$ where

$$R = \begin{bmatrix} A & B & 0 & 0 & \cdots & 0 & 0 \\ B & A_1 & B & 0 & \cdots & 0 & 0 \\ 0 & B & A_1 & B & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & A_1 & B \\ 0 & 0 & 0 & 0 & \cdots & B & A \end{bmatrix}$$

with

$$A = (a_{mi}) = \left[\frac{1}{\Delta t}\int_0^1 \psi_i'(x)\psi_m'(x)dx + \right.$$

$$\left. + \frac{\Delta t}{3}\int_0^1 \psi_i'''(x)\psi_m'''(x)dx + \int_0^1 \psi_i''(x)\psi_m''(x)dx\right]_{N\times N}$$

$$B = (b_{mi}) = \left[\left(-\frac{1}{\Delta t}\right)\int_0^1 \psi_i'(x)\psi_m'(x)dx + \frac{\Delta t}{6}\int_0^1 \psi_i'''(x)\psi_m'''(x)dx\right]_{N\times N}$$

$$A_1 = (a_{1_{mi}}) = \left[\frac{2}{\Delta t}\int_0^1 \psi_i'(x)\psi_m'(x)dx + 4\,\frac{\Delta t}{6}\int_0^1 \psi_i'''(x)\psi_m'''(x)dx\right]_{N\times N}$$

and

$$x = (C_1\, C_2 \cdots C_M)^t_{NM\times 1}, \quad \text{with} \quad C_j = (C_{ij} \cdot C_{2j} \cdots C_{Nj})^t_{N\times 1}$$

$$v = (v_1\, v_2 \cdots v_M)^t_{NM\times 1}, \quad \text{with} \quad v_1 = \left(\int_0^1 u_0(x)\psi_1''(x)dx \cdots \int_0^1 u_0(x)\psi_N''(x)dx\right)^t_{N\times 1}$$

and, for $2 \le j \le M$, v is the null vector with N components.

Finally we get the approximation to $\hat{u}(x_\ell, t_k)$ which we call now

$$u(x_\ell, t_k) = w_2(x_\ell, t_k) + w_1(x_\ell, t_k) =$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\phi_j(t_k)\psi_i''(x_\ell) + \sum_{i=j}^{N}\sum_{j=1}^{M} c_{ij}\phi_j'(t_k)\psi_i(x_\ell)$$

for each ℓ and each k.

The Convergence

Let $n = NM$. To use Theorem 2.4 to show that $\{u_n\} \to \hat{u}$ we have only to prove

Theorem 3.1: Let $D = UY_n$; D is dense in Y.

Proof: We have

$$D = \text{span}\,[\phi_j'(t)\psi_i(x),\ \phi_j(t)\psi_i''(x)]$$

and we show that the orthogonal to D in Y is the null space.

Let $(g_1, g_2) \in D^\perp$. So, $(g_1, g_2) \in Y$ and, for any $(p_1, p_2) \in D$,

$\langle (g_1, g_2),\ (p_1, p_2) \rangle_Y = 0$.

Let $p(t)$ and $\psi(x)$ be regular functions such that $\psi(0) = \psi(1) = \psi''(0) = \psi''(1) = 0$ and let $p_1(x,t) = p'(t)\psi(x)$ and $p_2(x,t) = p(t)\psi''(x)$. Then p_1 and $p_2 \in V$ and

$$\frac{\partial p_2}{\partial t} = \frac{\partial^2 p_1}{\partial x^2}.$$

Further $(p_1, p_2) \in D$ so,

$$\int_0^1 \int_0^T \frac{\partial g_1}{\partial x} p'(t)\psi'(x)dt\,dx + \int_0^1 \int_0^T \frac{\partial g_2}{\partial x} p(t)\psi'''(x)dt\,dx = 0. \tag{10}$$

Using (4) for (g_1, g_2) with $\varphi(x,t) = p_1(x,t)$ and (10), we find that

$$\left[\begin{array}{l} \int_0^1 \int_0^T g_2(x,t)[p(t)\psi^{iv}(x) - p''(t)\psi(x)]dt\,dx = \\ \\ \quad = \int_0^1 g_2(x,T)p'(T)\psi(x)dx - \int_0^1 g_2(x,0)p'(0)\psi(x)dx \end{array} \right. \tag{11}$$

Let us take now $\psi(x) = \operatorname{sen} n\pi x$ and $p(t) = e^{n^2\pi^2 t} + e^{-n^2\pi^2 t}$.

Substituting $\psi(x)$ and $p(t)$ in (11) and observing that $p'(0) = 0$ and $p'(T) \neq 0$ we have that, for any $n \in N$, $\int_0^1 g_2(x,T)\operatorname{sen} n\pi x\,dx = 0$ and since $g_2(0,T) = g_2(1,T) = 0$, this implies in $g_2(x,T) = 0$.

With similar arguments we prove that $g_2(x,0) = 0$.

So, (11) becomes

$$\int_0^1 \int_0^T g_2(x,t)[p(t)\psi^{iv}(x) - p''(t)\psi(x)]dt\,dx = 0$$

for any $\varphi(x,t) = p(t)\psi(x)$ in V.

Taking again $\psi(x) = \operatorname{sen} n\pi x$ and now $p(t) = \operatorname{sen} r\pi t$, we find that, for any $k \in N$ and $n \in N$,

$$\int_0^1 \int_0^T g_2(x,t)\operatorname{sen} n\pi x \operatorname{sen} k\pi t\,dt\,dx = 0$$

and, since $g_2(0,t) = g_2(1,t) = 0$ and $g_2(x,0) = g_2(x,T) = 0$, we have $g_2(x,t) \equiv 0$.

Using this in (10) we have

$$\int_0^1 \int_0^T \frac{\partial g_1}{\partial x} p'(t)\psi'(x)dt\,dx = 0$$

for any $p(t)$ and $\varphi(x)$ such that both $p_1(x,t)$ and $p_2(x,t)$ belong to V. Similarly we prove that $g_1(x,t)$ is also identically zero.

4. Numerical Results and Conclusions

To test this method we took for initial condition, $u_0(x) = \operatorname{sen} \pi x$ and so, the exact solution of equation (8) is $u(x,t) = \operatorname{sen} n\pi x e^{-\pi^2 t}$. We could then make a pointwise measure of the error.

We worked in FORTRAN and used a digital VAX computer and solved the linear system $Rx = v$ in four different ways:

(i) Using the program TBLOCK part of which was developed by P. Pulino, which uses the particular structure of matrix R (tridiagonal by blocks) and the fact that each block has band seven. For small values of $n(= NM)$ we had no good results and divergence occured for bigger n ($n = 561$ the maximum error is 0.12×10^6). We worked with double precision but we used no pivotal strategy.

(ii) Using the routines FO1BRF and FO4AXF from $MA - 28$ [5] from NAG'S library. FO1BRF makes the factorization of R in L.U using the sparcity of R. It uses a pivotal strategy which keeps the sparcity of R and controll the loss of precision by rounding at the same time. FO4AXF solves linear sparse systems with just one right hand side vector, using factorization made by FO1BRF. This also diverged.

(iii) Using FO4ATF [18], (i), also from NAG'S library which works with full matrices and solves $Rx = v$ by Crout's method with partial pivoting strategy to decompose R in L.U and also makes iterative refinement of the solution. We got convergence but not as we expected, that is, order Δt in t and $(\Delta x)^2$ in x.

(iv) We obtained praticaly the same results as in (iii) using the Conjugate Gradient method where we stopped the iterative process after n iterations and the norm of the gradient were still 10^{-1}.

The problem here is that the matrix R is very ill conditioned. With the routine FO2AAF [18], (ii) from NAG's library, which gives us all the eigenvalues of a real symmetric matrix by Householder reduction and QL algorithm, we got, for instance, for $n = 119$, $\operatorname{cond}(R) \simeq 10,400$ and for $n = 209$, $\operatorname{cond}(R) \simeq 6 \times 10^7$.

According to Forsythe and Moler [6] a measure of the relative error in the solution of a linear system of equation caused by a relative error in the constant vector or in the matrix entrances is given by $\frac{||\Delta u||}{||u||} \leq \rho \cdot \operatorname{cond}(R) \cdot \beta^{-t}$ where ρ is a quantity no bigger than the basis β in which we work and β^{-t} is the machine epsilon which, for us is about 5×10^{-8}. So, for $n = 119$ we have $(||\Delta u||/||u||) \leq 5 \times 10^{-3}$ and for $n = 209$, $(||\Delta u||/||u||) \leq 30$.

Since we work with cubic B–splines in x and, in some elements of R it appears

$$\int_0^1 \psi_r'(x)\psi_s'(x)dx \quad \text{and in other} \quad \int_0^1 \psi_r'''(x)\psi_s'''(x)dx,$$

with the change of variables we have, in some elements the factor Δx^{-1} and in other elements, the factor Δx^{-5}. This occurs because we impose $V_\beta 1$ to be satisfied exactly and so we need the third derivative of the B–splines.

In a next paper [13] we work with condition $V_\beta 1$ in the weak form

$$\langle \frac{\partial w_2}{\partial t}, \varphi \rangle = \langle \frac{\partial^2 w_1}{\partial x^2}, \varphi \rangle$$

for any φ belonging to the vector space of trial functions. With this we can work with piecewise linear functions both in x and in t and our results are much better.

References

[1] ADAMS, R. A., Sobolev Spaces, Academic Press, New York, 1975.

[2] AXELSSON, O. and BARKER, V. A., Finite Element Solution of Boundary Value Problems – theory and computation, Computer Science and Applied Mathematics, Academic Press, Inc., 1984.

[3] BACHMAN, G. and NARICI, L., Functional Analysis, Academic Press International Edition, 1966.

[4] CIARLET, P. G., SCHULTZ, M. H. and VARGA, R. S., Numerical Methods of High Order Accuracy for Non–Linear Boundary Value Problems I – one dimensional problems, Num. Math. 9, 394-430, 1966.

[5] DUFF, I. S., MA–28, A Set of FORTRAN Subroutines for Sparse Unsymmetric Linear Equations, A.E.R.E. Report R. 8730. MMSO, London, 1977.

[6] FORSYTHE, G. E., MALCOLM, M. A. and MOLER, C. B., Computer Methods for Mathematical Computations, Prentice Hall, 1977.

[7] GOLUB, G. H. and VAN LOAN, C. F., Matrix Computations, The John Hopkins University Press, 1985.

[8] HERRERA, I., A General Formulation of Variational Principles, Instituto de Ingenieria, UNAM, E10 Mexico, DF, 1974.

[9] HERRERA, I. and SEWELL, W. J., Dual Extremum Principles for Non–Negative Unsymmetric Operators, J. Inst. Maths. Applics, 21, 95-115, 1978.

[10] HESTENES, M. R. and STIEFEL, E., Methods of Conjugate Gradients for Solving Linear Systems, J. Res. Nat. Bur. Stand. 49. 409-436. 1952.

[11] HESTENES, M. R., Conjugate Direction Methods in Optimization, Springer–Verlag, Berlin, 1980.

[12] ISAACSON, E. and KELLER, H. B., Analysis of Numerical Methods, John Wiley & Sons, 1966.

[13] LOPES, VERA L. R. and ZAGO, J. V., Dual Extremum Principles for the Heat Equation Solved by Finite Element Methods II, to appear.

[14] NOBLE, B., Variational Finite Element Methods for Initial Value Problems, The Mathematics of The Finite Element Method and Applications, J. R. Whiteman Academic Press, New York, 1973.

[15] NOBLE, B. and SEWELL, M. J., On Dual Extremum Principles in Applied Mathematics, J. Inst. Maths. and its Applics. Vol. 9, number 2, April 1972.

[16] PRENTER, P. M., Splines and Variational Methods, John Wily & Sons, Inc., 1975.

[17] STRANG, G. and FIX, G. J., An Analysis of The Finite Element Method, Prentice–Hall, Inc., Englewood Cliffs, N. J., 1973.

[18] WILKINSON, J. H. and REINSCH, C., Handbook for Automatic Computation, Vol. II, Linear Algebra, Springer–Verlag, 1971, (i) pg. 93-110; (ii) pg. 212-216 and 227-240.

[19] ZAGO, J. V., Approximate Solution of Generalized Hamiltonian Equations with Applications, Ph.D. Thesis, Wisconsin, Madison, 1976.

RELATÓRIOS TÉCNICOS — 1989

01/89 — Uniform Approximation of Continuous Functions With Values in [0, 1] — *João B. Prolla.*

02/89 — On Some Nonlinear Iterative Relaxation Methods in Remote Sensing — *A. R. De Pierro.*

03/89 — A Parallel Iterative Method for Convex Programming with Quadratic Objective — *Alfredo N. Iusem* and *Alvaro R. De Pierro.*

04/89 — Fifth Force, Sixth Force, and all that: a Theoretical (Classical) Comment — *Erasmo Recami* and *Vilson Tonin-Zanchin.*

05/89 — An Application of Singer's Theorem to Homogeneous Polynomials — *Raymundo Alencar.*

06/89 — Summhammer's Experimental Test of the Non-Ergodic Interpretation of Quantum Mechanics — *Vincent Buonomano.*

07/89 — Privileged Reference Frames in General Relativity — *Waldyr A. Rodrigues Jr.* and *Mirian E. F. Scanavini.*

08/89 — On the Numerical Solution of Bound Constrained Optimization Problems — *Ana Friedlander* and *José Mario Martínez.*