On a Finite Element Method for Solving the Neutron Transport Equation

P. LASAINT AND P. A. RAVIART

Introduction.

Let Ω be a convex open set in the (x,y)-plane with boundary Γ. Denote by $\underline{n} = (n_x, n_y)$ the outward unit vector normal to Γ.

Let Q be the unit disk in the (μ,ν)-plane. We consider the following problem: Find a function $u = u(x,y,\mu,\nu)$ such that

$$\mu\frac{\partial u}{\partial x} + \nu\frac{\partial u}{\partial y} + \sigma u = f \quad \text{in } \Omega \times Q, \tag{1.1}$$

$$u(x,y,\mu,\nu) = 0 \quad \text{if } (x,y) \in \Gamma, \ (\mu n_x + \nu n_y)(x,y) < 0. \tag{1.2}$$

Equation (1.1) is the neutron transport equation: The function $u(x,y,\mu,\nu)$ represents the flux of neutrons at the point (x,y) in the angular direction (μ,ν), σ is the nuclear cross section, and f stands for the scattering, the fission and the inhomogeneous source terms. The boundary condition (1.2) simply means that no neutrons are entering the system from outside.

In this paper, we shall be only concerned with the spatial discretization of problem (1.1), (1.2). Thus, we shall assume that the angular direction (μ,ν) is fixed and we shall consider the reduced problem: Given a function f defined over Ω, find a function u defined over Ω such that

$$\underline{\underline{m}} \cdot \operatorname{grad} u + \sigma u = f \quad \text{in } \Omega ,$$

(1.3)

$$u = 0 \quad \text{on } \Gamma_- ,$$

where $\underline{\underline{m}} = (\mu, \nu)$ and

(1.4) $$\Gamma_- = \{(x,y) \in \Gamma \mid \underline{\underline{m}} \cdot \underline{\underline{n}}(x,y) < 0\} .$$

This paper will be devoted to the numerical approximation of problem (1.3) by a finite element method using triangular or quadrilateral elements which has been recently introduced by Reed and Hill [17] and which appears to be very effective in practice. Other finite element methods for solving the neutron transport equation have been introduced by several authors (cf. for instance [10], [14], [15], [16]). We refer to [12] for a mathematical discussion of some of them.

An outline of the paper is as follows. In §2, we study a discontinuous Galerkin method for ordinary differential equations using polynomials of degree k. This Galerkin method is shown to be strongly A-stable and of order $2k+1$. In §3, we introduce the finite element method as a generalization of the discontinuous Galerkin method of §2. We prove the existence and uniqueness of the approximate solution and we give an algorithm for computing this approximate solution. In §4, we derive general error bounds in the L_2-norm. Finally, we give in §5 a superconvergence result.

Note that problem (1.3) is a simple but important example of a first-order hyperbolic problem. In fact, the finite element method studied in this paper provides an effective way for numerically solving such problems. For other finite element methods for solving first order systems of partial differential equations, we refer to [11], [13].

For the sake of simplicity, we have confined ourselves to polygonal domains Ω. It is probably an easy matter to handle general curved domains by using curved isoparametric elements and the analysis given in [5], [6].

2. A Discontinuous Galerkin Method for Ordinary Differential Equations.

We begin by studying the numerical solution of the ordinary differential equation

$$u'(x) = f(x, u(x)), \quad x \geq x_0 ,$$
$$u(x_0) = u_0 , \tag{2.1}$$

on a finite interval $[x_0, x_0+a]$ by a discontinuous Galerkin method. For continuous Galerkin methods and related collocation methods, we refer for instance to Axelsson [1], de Boor and Swartz [2], Hulme [9].

Let $x_n = x_0 + nh$, $0 \leq n \leq N$ $(Nh = a)$ be a uniform mesh for the sake of simplicity. Then we may approximate u by a function u_h which, on each subinterval $[x_n, x_{n+1}]$, reduces to a polynomial of degree $\leq k$. We require that u_h satisfies on each subinterval $[x_n, x_{n+1}]$, $0 \leq n \leq N-1$:

$$(u_h(x_{n+}) - u_h(x_{n-}))v(x_n) + \int_{x_n}^{x_{n+1}} \{u_h'(x) - f(x, u_h(x))\}v(x)dx = 0 \tag{2.2}$$
$$\text{for all } v \in P_k$$

with the initial condition

$$u_h(x_{0-}) = u_0 , \tag{2.3}$$

where P_k denotes the space of all polynomials of degree $\leq k$. Notice that the function u_h is in general discontinuous at the mesh points x_n.

To obtain a computational form of (2.2)-(2.3), we replace the integral in (2.2) by an interpolatory quadrature formula

$$\int_{x_n}^{x_{n+1}} \varphi(x)dx = h \sum_{i=1}^{k+1} b_i \varphi(x_{n,i}) + O(h^{p+1}) , \tag{2.4}$$

$$x_{n,i} = x_n + \xi_i h , \quad 1 \le i \le k+1 , \quad \xi_1 = 0 , \tag{2.5}$$

where b_i and ξ_i are the weights and abscissae for $[0,1]$. Notice that $k+1 \le p \le 2k+1$. Then (2.2) becomes

$$\begin{aligned} &(u_h(x_{n+}) - u_h(x_{n-}))v(x_n) \\ &+ h \sum_{i=1}^{k+1} b_i \{u_h'(x_{n,i}) - f(x_{n,i}, u_h(x_{n,i}))\} v(x_{n,i}) = 0 \\ &\qquad\qquad \text{for all } v \in P_k . \end{aligned} \tag{2.6}$$

Let us now show that the discrete Galerkin method (2.3), (2.6) is equivalent to some implicit Runge-Kutta method. We define

$$\begin{aligned} &u_n = u_h(x_{n-}), \\ &u_{n,1} = u_h(x_{n+}) = u_h(x_{n,1}) , \\ &u_{n,i} = u_h(x_{n,i}) , \quad 2 \le i \le k+1 . \end{aligned} \tag{2.7}$$

We introduce the Lagrange interpolation coefficients

$$\ell_i(x) = \prod_{\substack{j=2 \\ j \ne i}}^{k+1} \frac{x - \xi_j}{\xi_i - \xi_j} , \quad 2 \le i \le k+1 . \tag{2.8}$$

Lemma 1.

The discrete Galerkin method (2.3), (2.6) is equivalent to the following implicit Runge-Kutta method

$$\begin{aligned} u_{n,i} &= u_n + h \sum_{j=1}^{k+1} a_{ij} f(x_{n,j}, u_{n,j}), \quad 1 \le i \le k+1 , \\ u_{n+1} &= u_n + h \sum_{j=1}^{k+1} b_j f(x_{n,j}, u_{n,j}) , \end{aligned} \tag{2.9}$$

where

$$\begin{aligned} &a_{i1} = b_1 , \quad 1 \le i \le k+1 , \\ &a_{ij} = \int_0^{\xi_i} \ell_j(x)dx - b_1 \ell_j(\xi_1), \quad 1 \le i \le k+1, \ 2 \le j \le k+1. \end{aligned} \tag{2.10}$$

Proof.

Let us introduce the basis $\{v_i\}_{1\le i\le k+1}$ for the space P_k defined by

$$v_i(x_{n,j}) = \delta_{ij} , \quad 1 \le i,j \le k+1 .$$

By replacing successively in (2.6) v by v_i, we find that an equivalent form of (2.6) is given by

$$u_h(x_{n+}) - u_h(x_{n-}) + hb_1[u_h'(x_{n,1}) - f(x_{n,1}, u_h(x_{n,1}))] = 0$$

$$u_h'(x_{n,i}) - f(x_{n,i}, u_h(x_{n,i})) = 0 , \quad 2 \le i \le k+1 . \tag{2.11}$$

In the subinterval $[x_n, x_{n+1}]$, we have $u_h' \in P_{k-1}$ so that

$$u_h'(x) = \sum_{j=2}^{k+1} \ell_j\left(\frac{x-x_n}{h}\right) u_h'(x_{n,j})$$

and by (2.11)

$$u_h'(x) = \sum_{j=2}^{k+1} \ell_j\left(\frac{x-x_n}{h}\right) f(x_{n,j}, u_h(x_{n,j})) . \tag{2.12}$$

Taking $x = x_n = x_{n,1}$ in (2.12), substituting this expression into the 1st equation (2.11) and using (2.7), we obtain

$$u_{n,1} = u_n + h\, b_1\{f(x_{n,1}, u_{n,1}) - \sum_{j=2}^{k+1} \ell_j(\xi_1) f(x_{n,j}, u_{n,j})\}. \tag{2.13}$$

On the other hand, we may write for $2 \le i \le k+1$

$$u_h(x_{n,i}) = u_h(x_{n,1}) + \int_{x_{n,1}}^{x_{n,i}} u_h'(x)dx$$

and by (2.7), (2.12), (2.13)

$$u_{n,i} = u_n + h\{b_1 f(x_{n,1}, u_{n,1}) + \sum_{j=2}^{k+1} [\int_0^{\xi_i} \ell_j(x)dx - b_1 \ell_j(\xi_1)] f(x_{n,j}, u_{n,j})\} . \tag{2.14}$$

Similarly, we have

$$u_h(x_{n+1}-) = u_h(x_{n,1}) + \int_{x_n}^{x_{n+1}} u_h'(x)dx$$

and then

$$u_{n+1} = u_n + h\{b_1 f(x_{n,1}, u_{n,1}) + \sum_{j=2}^{k+1} [\int_0^1 \ell_j(x)dx - b_1 \ell_j(\xi_1)] f(x_{n,j}, u_{n,j})\} .$$

By noticing that

$$\int_0^1 \ell_j(x)dx = \sum_{i=1}^{k+1} b_i \ell_j(\xi_i) = b_1 \ell_j(\xi_1) + b_j ,$$

we get

$$u_{n+1} = u_n + h \sum_{j=1}^{k+1} b_j f(x_{n,j}, u_{n,j}) . \tag{2.15}$$

The equations (2.13)-(2.15) are identical to the equations (2.9), (2.10). We then have proved that the discrete Galerkin method leads to the one-step method (2.9), (2.10). Conversely, the Runge-Kutta method (2.9), (2.10) can be clearly viewed as a discrete Galerkin method. ∎

Theorem 1.

The discrete Galerkin method (2.3), (2.6) is a one-step method of order p .

Proof.

Following Butcher [3], Crouzeix [7], we know that the conditions

$$\sum_{j=1}^{k+1} b_j \xi_j^{\ell} = \frac{1}{\ell+1} , \quad 0 \le \ell \le p-1 , \tag{2.16}$$

$$\sum_{j=1}^{k+1} a_{ij} \xi_j^{\ell} = \frac{\xi_i^{\ell+1}}{\ell+1} , \quad 0 \le \ell \le k-1 , \quad 1 \le i \le k+1 \tag{2.17}$$

$$(2.18)\quad \sum_{i=1}^{k+1} b_i a_{ij} \xi_i^{\ell} = \frac{1}{\ell+1} b_j (1-\xi_j^{\ell+1}),\ k+\ell \le p-1,\ 1 \le j \le k+1,$$

are sufficient for the Runge-Kutta method (2.9) to be of order p. Let us show that these conditions hold in the present case.

First, conditions (2.16) simply mean that the interpolatory quadrature formula (2.4) is exact for all polynomials of degree $\le p-1$.

Next, consider conditions (2.17). Using (2.8), we may write

$$x^{\ell} = \sum_{j=2}^{k+1} \ell_j(x) \xi_j^{\ell}, \qquad 0 \le \ell \le k-1,$$

so that

$$\xi_1^{\ell} = \sum_{j=2}^{k+1} \ell_j(\xi_1) \xi_j^{\ell}, \qquad 0 \le \ell \le k-1,$$

$$\frac{\xi_i^{\ell+1}}{\ell+1} = \sum_{j=2}^{k+1} \left(\int_0^{\xi_i} \ell_j(x)dx \right) \xi_j^{\ell}, \quad 0 \le \ell \le k-1,\ 1 \le i \le k+1.$$

Using (2.10), we have

$$\sum_{j=1}^{k+1} a_{ij} \xi_j^{\ell} = b_1 \left(\xi_1^{\ell} - \sum_{j=2}^{k+1} \ell_j(\xi_1) \xi_j^{\ell} \right) + \sum_{j=2}^{k+1} \left(\int_0^{\xi_i} \ell_j(x)dx \right) \xi_j^{\ell}$$

and by the previous relations

$$\sum_{j=1}^{k+1} a_{ij} \xi_j^{\ell} = \frac{\xi_i^{\ell+1}}{\ell+1}, \quad 0 \le \ell \le k-1, \quad 1 \le i \le k+1.$$

Finally, let us show that conditions (2.18) hold. We begin by noticing that

$$(2.19)\quad \sum_{i=1}^{k+1} b_i a_{il} \xi_i^{\ell} = b_1 \sum_{i=1}^{k+1} b_i \xi_i^{\ell} = \frac{b_1}{\ell+1}, \qquad 0 \le \ell \le p-1.$$

On the other hand, following Crouzeix [7], we may write for any continuous function φ

$$\int_0^1 x^{\ell}\Big(\int_0^x \varphi(y)dy\Big)dx = \frac{1}{\ell+1}\int_0^1 (1-x^{\ell+1})\varphi(x)dx \ . \tag{2.20}$$

Taking $\varphi \in P_{k-1}$, we obtain for $k+\ell \leq p-1$

$$\int_0^1 x^{\ell}\Big(\int_0^x \varphi(y)dy\Big)dx = \sum_{i=1}^{k+1} b_i \xi_i^{\ell} \int_0^{\xi_i} \varphi(y)dy$$

$$= \sum_{i=1}^{k+1} b_i \xi_i^{\ell} \sum_{j=2}^{k+1} \Big(\int_0^{\xi_i} \ell_j(y)dy\Big)\varphi(\xi_j)$$

and by (2.10)

$$\int_0^1 x^{\ell}\Big(\int_0^x \varphi(y)dy\Big)dx = \sum_{i,j=1}^{k+1} b_i a_{ij} \xi_i^{\ell} \varphi(\xi_j), \quad \varphi \in P_{k-1}, \quad k+\ell \leq p-1 \ . \tag{2.21}$$

Similarly, we get

$$\int_0^1 (1-x^{\ell+1})\varphi(x)dx = \sum_{j=1}^{k+1} b_j (1-\xi_j^{\ell+1})\varphi(\xi_j), \quad \varphi \in P_{k-1}, \quad k+\ell \leq p-1 \ . \tag{2.22}$$

Hence, combining (2.19)-(2.22), we have for all $\varphi \in P_{k-1}$ and for $k + \ell \leq p-1$

$$\sum_{j=1}^{k+1} \Big[\sum_{i=1}^{k+1} b_i a_{ij} \xi_i^{\ell} - \frac{1}{\ell+1} b_j (1 - \xi_j^{\ell+1}) \Big] \varphi(\xi_j) = 0 \ .$$

This implies

$$\sum_{i=1}^{k+1} b_i a_{ij} \xi_i^{\ell} = \frac{1}{\ell+1} b_j (1-\xi_j^{\ell+1}), \quad k+\ell \leq p-1, \quad 2 \leq j \leq k+1.$$

■

In order to investigate the stability properties of the one-step method (2.9), we consider the differential equation

$$u' = \lambda u \tag{2.23}$$

where λ is a complex constant with $Re(\lambda) < 0$.

Lemma 2.

Applied to the differential equation (2.23), the one-step method (2.9), (2.10) gives

$$u_{n+1} = R(\lambda h)\, u_n \tag{2.24}$$

where $R(z) = \frac{P(z)}{Q(z)}$ is the quotient of two polynomials $P(z)$ and $Q(z)$ of degree $\leq k$ and $\leq k+1$, respectively.

Proof.

Applied to (2.23), the one-step method (2.9) becomes

$$u_{n,i} = u_n + \lambda h \sum_{j=1}^{k+1} a_{ij}\, u_{n,j} , \quad 1 \leq i \leq k+1 \tag{2.25}$$

$$u_{n+1} = u_n + \lambda h \sum_{j=1}^{k+1} b_j\, u_{n,j} . \tag{2.26}$$

Using obvious notations, we may write equations (2.25) in the form

$$(I - \lambda h\, [a_{ij}])[u_{n,i}] = u_n[1]$$

where the identity matrix I and $[a_{ij}]$ are $(k+1)\times(k+1)$-matrices. Since $a_{i1} = b_1$, $1 \leq i \leq k+1$, we get from Cramer's rule

$$u_{n,i} = \frac{P_i(\lambda h)}{Q(\lambda h)}\, u_n , \quad 1 \leq i \leq k+1$$

where $P_1(z)$ is a polynomial of degree k whose leading coefficient is $b_1^{-1} \det [a_{ij}]$, $P_i(z)$, $2 \leq i \leq k+1$, are polynomials of degree $\leq k-1$ and where $Q(z)$ is a polynomial of degree $k+1$ whose leading coefficient is $\det[a_{ij}]$.

Using (2.26), we obtain

$$u_{n+1} = \frac{P(\lambda h)}{Q(\lambda h)} u_n$$

where

$$P(z) = Q(z) - z \sum_{j=1}^{k+1} P_j(z) .$$

Clearly, in $P(z)$, the coefficient of z^{k+1} vanishes. The lemma is then proved. ∎

Let us now recall the following definition: A one-step method is strongly A-stable if

$$\begin{aligned} &|R(z)| < 1 \qquad \text{for } Re(z) < 0 , \\ &|R(z)| \to 0 \qquad \text{as } Re(z) \to -\infty. \end{aligned} \tag{2.27}$$

Theorem 2.

The Galerkin method (2.2), (2.3) is a strongly A-stable one-step method of order $2k+1$.

Proof.

Consider first the discrete Galerkin method (2.3), (2.6) associated with the Gauss-Radau abscissae ξ_i, $1 \leq i \leq k+1$ ($\xi_1 = 0$). Then, we have $p = 2k+1$ in (2.4). By Theorem 1, this discrete Galerkin method is a one-step method of order $2k+1$ so that

$$R(z) = \exp(z) + O(z^{2k+2}) .$$

Moreover, by Lemma 2, $R(z)$ is the quotient of two polynomials $P(z)$ and $Q(z)$ of degree $\leq k$ and $\leq k+1$, respectively. Then, necessarily, $R(z)$ is the subdiagonal $(k+1,k)$ Padé rational approximation of $\exp(z)$. Using a result of Axelsson [1], we know that such a Padé approximation satisfies conditions (2.27). Hence, the discrete Galerkin method (2.3), (2.6) associated with the Gauss-Radau abscissae is

a strongly A-stable one-step method of order $2k+1$.

Now, it is a simple but lengthy matter to prove that the Galerkin method (2.2), (2.3) and the Gauss-Radau discrete Galerkin method (2.3), (2.6) are one-step methods of the same order $2k+1$. Moreover, these two methods coincide when applied to the differential equation (2.23). This completes the proof of the theorem. ∎

3. A Finite Element Method for the Neutron Transport Equation.

Consider now our neutron transport problem (1.3). First, we need some notations. Let us denote by $L_2(\Omega)$ the space of real-valued functions v which are square integrable over Ω. We provide $L_2(\Omega)$ with the usual norm

$$\|v\|_{0,\Omega} = \left(\int_{\Omega} |v(x)|^2 dx\right)^{\frac{1}{2}} . \tag{3.1}$$

Given any integer $m \geq 0$, let

$$H^m(\Omega) = \{v \in L_2(\Omega) \mid \partial^\alpha v \in L_2(\Omega),\ |\alpha| \leq m\} \tag{3.2}$$

be the usual Sobolev space provided with the norm

$$\|v\|_{m,\Omega} = \left(\sum_{|\alpha| \leq m} \|\partial^\alpha v\|^2_{0,\Omega}\right)^{\frac{1}{2}} \tag{3.3}$$

In (3.2), (3.3), $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}^2$ is a multi-index, $|\alpha| = \alpha_1 + \alpha_2$, and

$$\partial^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial}{\partial x_2}\right)^{\alpha_2} .$$

We shall also use the following semi-norm

$$|v|_{m,\Omega} = \left(\sum_{|\alpha| = m} \|\partial^\alpha v\|^2_{0,\Omega}\right)^{\frac{1}{2}} \tag{3.4}$$

Let us introduce the operator

$$A = \underline{\underline{m}} \cdot \operatorname{grad} + \sigma = \mu \frac{\partial}{\partial x} + \nu \frac{\partial}{\partial y} + \sigma \tag{3.5}$$

and the space

$$D(A) = \{v \in L_2(\Omega) \mid \underline{\underline{m}} \cdot \operatorname{grad} v \in L_2(\Omega)\}. \tag{3.6}$$

Then, as a consequence of [8], we have the following result.

Theorem 3.

Assume that $\sigma \in L_\infty(\Omega)$ and $f \in L_2(\Omega)$. Then, problem (1.3) has a unique strong solution $u \in D(A)$.

With the substitution

$$u = \exp(\lambda(\frac{x}{\mu} + \frac{y}{\nu}))w,$$

equation (1.3) becomes

$$\underline{\underline{m}} \cdot \operatorname{grad} w + (\sigma + \lambda)w = \exp(-\lambda(\frac{x}{\mu} + \frac{y}{\nu}))f .$$

Thus, by eventually changing $\sigma(x,y)$ into $\sigma(x,y) + \lambda$, we can restrict ourselves to the case where σ is positive. More precisely, we shall assume in the sequel that

$$M \geq \sigma(x,y) \geq \alpha > 0 \quad \text{a.e. in } \Omega . \tag{3.7}$$

Let us now generalize the one-dimensional discontinuous Galerkin method of §2 to our two-dimensional neutron transport problem. For the sake of simplicity, we shall assume in the following that $\bar{\Omega}$ is a polygon. In order to approximate problem (1.3), we first construct a triangulation $\mathfrak{T}_h$ of $\bar{\Omega}$ with triangles and convex quadrilaterals K with diameters $\leq h$. With any $K \in \mathfrak{T}_h$, we associate the finite-dimensional space P_K of real-valued functions defined on K such that

$$P_K \subset H^1(K) . \tag{3.8}$$

We then consider the finite-dimensional space

$$(3.9)\qquad V_h = \{v \mid v \in L_2(\Omega),\ v|_K \in P_K \ \text{ for all } K \in \mathfrak{T}_h\}.$$

It is worthwhile to notice that in general a function $v \in V_h$ does not satisfy any continuity requirement at the interelement boundaries.

Let $K \in \mathfrak{T}_h$ and let ∂K be the boundary of K. We set

$$(3.10)\qquad \begin{aligned} \partial_- K &= \{(x,y) \in \partial K \mid \underline{\underline{m}} \cdot \underline{\underline{n}}(x,y) < 0\}, \\ \partial_+ K &= \{(x,y) \in \partial K \mid \underline{\underline{m}} \cdot \underline{\underline{n}}(x,y) > 0\}, \end{aligned}$$

where $\underline{\underline{n}} = (n_x, n_y)$ is the outward unit vector normal to the boundary ∂K.

Then, the finite element approximation of problem (1.3) that we shall consider here can be stated as follows. Find a function $u_h \in V_h$ such that for all $K \in \mathfrak{T}_h$

$$(3.11)\qquad -\int_{\partial_- K} \underline{\underline{m}} \cdot \underline{\underline{n}}\,(u_h - \xi_h)v\,ds + \int_K (Au_h - f)v\,dxdy = 0$$

for all $v \in P_K$

where

$$(3.12)\qquad \xi_h = \begin{cases} 0 & \text{on } \partial_- K \cap \Gamma_-, \\ \text{outward trace of } u_h & \text{on } \partial_- K \setminus (\partial_- K \cap \Gamma_-). \end{cases}$$

This method is clearly a direct generalization of the discontinuous Galerkin method (2.2), (2.3).

Before proving existence and uniqueness of the solution $u_h \in V_h$, we shall show that there exists an ordering of the elements of $\mathfrak{T}_h$ well suited for numerically solving equations (3.11), (3.12).

Lemma 3.

There exists an ordering $K_1, K_2, \ldots, K_I$ of the elements of $\mathfrak{T}_h$ such that, for all $i = 1, \ldots, I$, each side of

$\partial_- K_i$ is either a subset of Γ_- or a subset of $\partial_+ K_j$ for some $j < i$.

Proof.

Let us introduce first some notations. We shall say that K is a boundary element if at least one side of ∂K is a subset of Γ, and that K is a semi-boundary element if one and only one vertex of K belongs to Γ. Let us consider Γ_- and let us number clockwise the corresponding

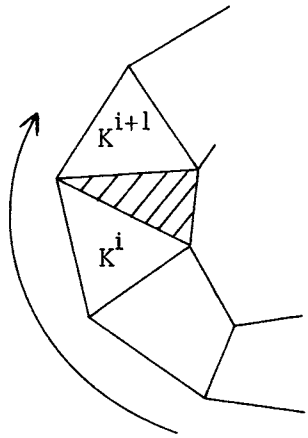

Figure 1

boundary elements $K^1, K^2, \ldots, K^s$. Two consecutive boundary elements K^i and K^{i+1} can have a common side or not. In the latter case (cf. Figure 1), there exists at least one semi-boundary element located between K^i and K^{i+1}. Then, we shall say that a side of K^i (resp. K^{i+1}) is semi-common with K^{i+1} (resp. K^i) if it is a subset of the union of the semi-boundary elements located between K^i and K^{i+1}.

Next, we show that there exists at least one boundary element K such that $\partial_- K \subset \Gamma_-$. To prove this, let us assume on the contrary that $\partial_- K^i \not\subset \Gamma_-$ for all $i = 1, \ldots, s$ and let us show that this assumption leads to a contradiction. Consider the first boundary element K^1 and use the notation of Figure 2.

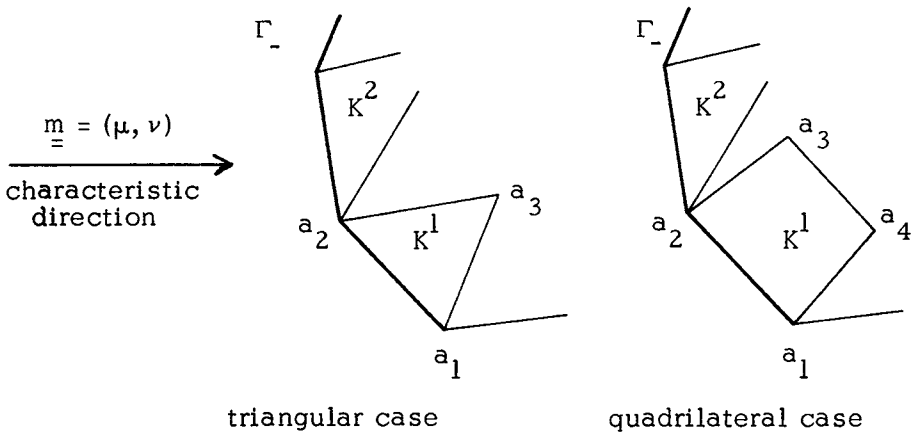

Figure 2

In the triangular case (resp. in the quadrilateral case), the side $[a_1, a_3]$ (resp. $[a_1, a_4]$) of K^1 is a subset of $\partial_+ K^1$. Otherwise, K^1 would not be the first boundary element of Γ_-. Then, the side $[a_2, a_3]$ of K^1 which is common or semi-common with K^2 belongs to $\partial_- K^1$. Otherwise, we should get $\partial_- K^1 = [a_1, a_2] \subset \Gamma_-$ which is excluded. Therefore, the side of K^2 which is common or semi-common with K^1 belongs to $\partial_+ K^2$. More generally, we get for every $i = 1, \ldots, s-1$ the following property: the side of K^i which is common or semi-common with K^{i+1} is a subset of $\partial_- K^i$ and therefore the side of K^{i+1} which is common or semi-common with K^i is a subset of $\partial_+ K^{i+1}$. Now consider the last boundary element K^s and use the notations of

Figure 3. In the triangular case (resp. in the quadrilateral case), the side $[a_1, a_3]$ (resp. $[a_1, a_4]$) of K^s is a subset of $\partial_+ K^s$. Moreover, the side $[a_2, a_3]$ is a subset of $\partial_+ K^s$. Otherwise, K^s would not be the last boundary element of Γ_-. Thus, we get $\partial_- K^s = [a_1, a_2] \subset \Gamma_-$ which has been excluded. The existence of a boundary element K such that $\partial_- K \subset \Gamma_-$ is then proved.

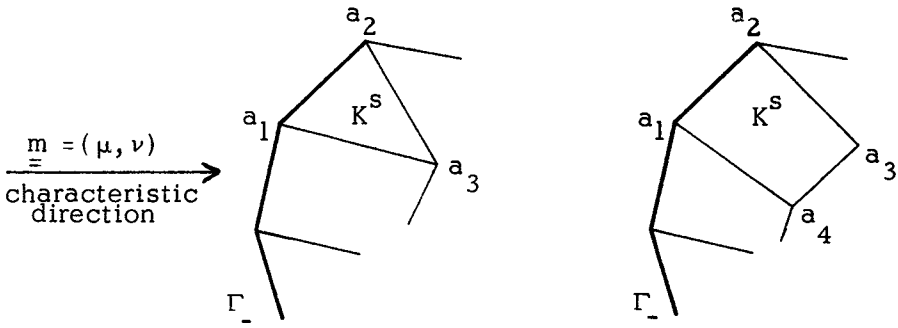

Figure 3

Now, choose for K_1 a boundary element of Γ_- such that $\partial_- K_1 \subset \Gamma_-$ and define $\Omega_1 = \Omega \setminus \Omega \cap K_1$, $\Gamma_{1-} = \partial_- \Omega_1$. Note that each side of Γ_{1-} is either a subset of Γ_- or a subset of $\partial_+ K_1$. By the previous argument, there exists a boundary element K_2 of Γ_{1-} such that $\partial_- K_2 \subset \Gamma_{1-}$, etc. Repeating this process, we take into account all the elements of $\mathfrak{T}_h$, and we obtain an ordering $K_1, K_2, \ldots, K_I$ of the elements of $\mathfrak{T}_h$ such that the desired property holds. ∎

This proof suggests an ordering algorithm for the element of $\mathfrak{T}_h$ which is effectively used in practice.

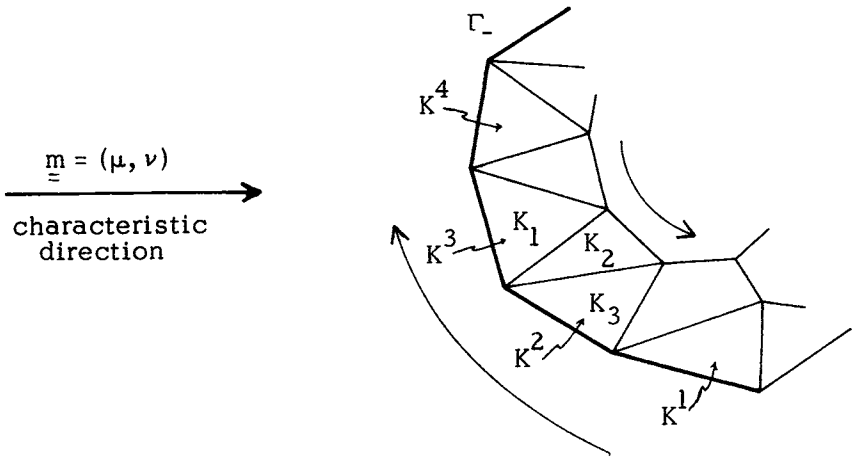

Figure 4

Consider the sequence $K^1, K^2, \ldots, K^s$ of boundary elements of Γ_-. For K_1 we choose the first element K of this sequence which satisfies $\partial_- K \subset \Gamma_-$. Let K^p be this element ($p = 3$ in Figure 4). From $K_1 = K^p$, we then number counterclockwise $K_1, K_2, \ldots, K_r$ the boundary and semi-boundary elements located between K^p and K^1 which satisfy the following condition: for all $i = 1, \ldots, r$, each side of $\partial_- K_i$ is either a subset of Γ_- or a subset of $\partial_+ K_j$ for some $j < i$ ($r = 3$ in Figure 4). Next, we replace the set Ω by $\Omega \backslash \Omega \cap (\bigcup_{i=1}^{r} K_i)$ and we repeat the process, etc.

We are now able to prove

Theorem 4.

Assume that $f \in L_2(\Omega)$ and that condition (3.7) holds. Then, there exists a unique function $u_h \in V_h$ which satisfies equations (3.11) and (3.12) for all $K \in \mathfrak{I}_h$.

Proof.

Clearly, the finite element method (3.11), (3.12) is equivalent to an $N \times N$ linear system of equations with $N = \dim V_h$. Then, it is sufficient to prove the uniqueness of the solution u_h. Thus, let us assume that $f = 0$ and let us show that necessarily $u_h = 0$. Let $K_1, K_2, \ldots, K_I$ be an ordering of the elements $K \in \mathcal{T}_h$ such that the condition of Lemma 3 holds. If $u_h = 0$ in $K_1 \cup K_2 \cup \ldots \cup K_{i-1}$, then $\xi_h = 0$ on $\partial_- K_i$ and equation (3.11) becomes, in $K = K_i$,

$$-\int_{\partial_- K} \underline{\underline{m}} \cdot \underline{\underline{n}}\, u_h v\, ds + \int_{K_i} (Au_h) v\, dxdy = 0$$

for all $v \in P_{K_i}$

Taking $v = u_h$ and using Green's formula

$$\int_{K_i} (\underline{\underline{m}} \cdot \operatorname{grad} u_h) u_h\, dxdy = \tfrac{1}{2} \int_{\partial K_i} \underline{\underline{m}} \cdot \underline{\underline{n}}\, u_h^2 ds ,$$

we get

$$\int_{\partial_+ K_i} \underline{\underline{m}} \cdot \underline{\underline{n}}\, u_h^2 ds - \int_{\partial_- K_i} \underline{\underline{m}} \cdot \underline{\underline{n}}\, u_h^2\, ds + \int_{K_i} \sigma u_h^2\, dxdy = 0 .$$

Using (3.7) and (3.10), we obtain $u_h = 0$ in K_i. Therefore, using an inductive argument, we get $u_h = 0$ in Ω. ∎

In practice, the computation of the approximate solution $u_h \in V_h$ goes along the following lines:

(i) Find an ordering $K_1, K_2, \ldots, K_I$ of the elements $K \in \mathcal{T}_h$ which satisfies the condition of Lemma 3, for instance by using the previous algorithm;

(ii) Compute successively u_h in $K_1, K_2, \ldots, K_I$. The computation of u_h in each K_i has a local character and involves the numerical solution of a $d_i \times d_i$ linear system where $d_i = \dim P_{K_i}$.

In other words, by using an ordering of $\mathfrak{T}_h$ such that the condition of Lemma 3 holds, the $N \times N$ matrix of the approximate problem becomes block triangular and the ith diagonal block is a $d_i \times d_i$ matrix associated with the ith element K_i.

Note that, in many practical problems, the geometry of Ω and the triangulation $\mathfrak{T}_h$ are so simple that step (i) becomes obvious.

4. General Error Bounds.

Let us now derive some estimates for the error $u_h - u$ when the solution u of problem (1.3) is smooth enough. We begin with

Lemma 4.

For any $K \in \mathfrak{T}_h$, any $v \in P_K$ and any function $\eta \in L_2(\partial_- K)$ we have the estimate

$$\begin{aligned} &\tfrac{1}{2}\int_{\partial_+ K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u_h - v)^2 ds + \tfrac{1}{2}\int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(\xi_h - \eta)^2 ds \\ &\quad - \tfrac{1}{2}\int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,((u_h - v) - (\xi_h - \eta))^2 ds + \int_K \sigma (u_h - v)^2 dx dy \\ &\quad = \int_{\partial_+ K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u - v)(u_h - v) ds + \int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u - \eta)(u_h - v) ds \\ &\qquad + \int_K (u - v) A^*(u_h - v) dx dy \end{aligned} \tag{4.1}$$

where A^* is the formal adjoint of the operator A, i.e.,

$$A^* = -\underline{\underline{m}} \cdot \operatorname{grad} + \sigma . \tag{4.2}$$

Proof.

Given $v \in P_K$ and $\eta \in L_2(\partial_- K)$, we set:

$$w = u_h - v \in P_K, \qquad \zeta = \xi_h - \eta . \tag{4.3}$$

Consider the expression

$$X_h = -\int_{\partial_- K} \underline{m}\cdot\underline{n}\ (w-\zeta)w ds + \int_K (Aw)w\, dxdy . \tag{4.4}$$

First, using Green's formula, we obtain

$$X_h = \tfrac{1}{2}\int_{\partial K} \underline{m}\cdot\underline{n}\ w^2 ds - \int_{\partial_- K} \underline{m}\cdot\underline{n}\ (w-\zeta)w\, ds + \int_K \sigma w^2 dx\, dy .$$

Since

$$(w-\zeta)w = \tfrac{1}{2}\left(w^2 - \zeta^2 + (w-\zeta)^2\right),$$

we get

$$X_h = \tfrac{1}{2}\int_{\partial_+ K} \underline{m}\cdot\underline{n}\ w^2 ds + \tfrac{1}{2}\int_{\partial_- K} \underline{m}\cdot\underline{n}\ \zeta^2 ds - \tfrac{1}{2}\int_{\partial_- K} \underline{m}\cdot\underline{n}\ (w-\zeta)^2 ds + \int_K \sigma w^2 dx\, dy \tag{4.5}$$

On the other hand, using (3.11), we obtain

$$X_h = \int_{\partial_- K} \underline{m}\cdot\underline{n}\ (v-\eta)w\, ds + \int_K (f - Av)w\, dxdy$$

and therefore

$$X_h = \int_{\partial_- K} \underline{m}\cdot\underline{n}\ (v-\eta)w\, ds + \int_K A(u-v)w\, dxdy .$$

Since $u \in D(A)$, we may write

$$\int_K A(u-v)w\,dx\,dy = \int_K (u-v)A^* w\,dxdy + \int_{\partial K} \underline{\underline{m}} \cdot \underline{\underline{n}}\,(u-v)w\,ds$$

so that

$$X_h = \int_{\partial_+ K} \underline{\underline{m}} \cdot \underline{\underline{n}}\ (u-v)w\,ds + \int_{\partial_- K} \underline{\underline{m}} \cdot \underline{\underline{n}}\ (u-\eta)w\,ds + \int_K (u-v)A^* w\,dxdy\,. \tag{4.6}$$

By combining (4.3), (4.5) and (4.6), we get the desired estimate. ∎

In order to get explicit error bounds, we need to define more precisely the finite-dimensional spaces P_K. Let K be an element of $\mathcal{T}_h$. If K is a triangle, there exists an affine invertible mapping F_K which maps a reference triangle $\hat{K}$ onto K ($\hat{K}$ is usually chosen as a unit, isosceles, right triangle). If K is a nondegenerate convex quadrilateral, there exists a biaffine invertible mapping F_K which maps the reference element $\hat{K} = [-1,+1]^2$ onto K. Note that this mapping F_K becomes affine when K is a parallelogram.

In both cases, let $\hat{P} \subset H^1(\hat{K})$ be a finite-dimensional space of real-valued functions defined on the reference element $\hat{K}$. We shall always assume in the following that

$$P_K = \{p \mid p = \hat{p} \circ F_K^{-1},\ \hat{p} \in \hat{P}\}\,. \tag{4.7}$$

We shall make constant use of the one-to-one correspondence

$$\hat{v} \mapsto v = \hat{v} \circ F_K^{-1}\,, \qquad v \mapsto \hat{v} = v \circ F_K$$

between the functions $\hat{v}$ defined on $\hat{K}$ and the functions v defined on K.

For any integer $m \geq 0$, let P_m denote the space of all polynomials of degree $\leq m$ in the two variables x, y and let Q_m denote the space of all polynomials of the form

$$p(x,y) = \sum_{i,j=0}^{m} c_{ij}\, x^i\, y^j, \quad c_{ij} \in \mathbb{R}.$$

We shall need

Hypothesis H.1.

There exists an integer $k \geq 0$ such that:

(4.8) $P_k \subset \hat{P}$ if $\hat{K}$ is the reference triangle,

(4.9) $Q_k \subset \hat{P}$ if $\hat{K}$ is the reference quadrilateral $[-1,+1]^2$.

Let us now introduce the following geometric parameters:

$h(K)$ = diameter of K,

(4.10) $\rho(K)$ = sup{diameter of the circles contained in K},

$\theta_i(K)$ $(1 \leq i \leq 4)$ = angles of K if K is a quadrilateral.

Hypothesis H.2.

There exists a constant $\sigma > 1$ independent of h such that

(4.11) $h(K) \leq \sigma\, \rho(K)$ for all $K \in \mathfrak{T}_h$.

Moreover, there exists a constant γ independent of h with $0 < \gamma < 1$ such that

(4.12) $\max_{1 \leq i \leq 4} |\cos \theta_i(K)| \leq \gamma$ for all quadrilateral $K \in \mathfrak{T}_h$.

Given a reference element $\hat{K}$, we define $\hat{\pi}$ to be the orthogonal projection operator on $L_2(\hat{K})$ to $\hat{P}$. For any $K \in \mathfrak{T}_h$, we define $\pi_K \in \mathfrak{L}(L_2(K); P_K)$ by

$$\widehat{\pi_K v} = \hat{\pi}\,\hat{v} \quad \text{for all} \quad v \in L_2(K) . \tag{4.13}$$

Then, for any $v \in L_2(\Omega)$, we define $\pi_h v$ to be the function in V_h such that

$$\pi_h v\big|_K = \pi_K v \quad \text{for all} \quad K \in \mathfrak{T}_h . \tag{4.14}$$

Let us now state some standard results which can be easily proved by using the techniques of Ciarlet and Raviart [4], [5].

Lemma 5.

Assume that Hypothesis H.2 holds. Then, there exists a constant $C > 0$ independent of $K \in \mathfrak{T}_h$ such that for all $p \in P_K$

$$|p|_{1,K} \leq C(h(K))^{-1} \|p\|_{0,K} , \tag{4.15}$$

$$\|p\|_{0,K'} \leq C(h(K))^{-\frac{1}{2}} \|p\|_{0,K} , \tag{4.16}$$

where K' is any side of K and $\|p\|_{0,K'} = (\int_{K'} |p|^2 ds)^{\frac{1}{2}}$.

Lemma 6.

Assume that Hypotheses H.1, H.2 and (4.13) hold. Then, there exists a constant $C > 0$ independent of $K \in \mathfrak{T}_h$ such that for all $v \in H^{k+1}(K)$

$$|v - \pi_K v|_{m,K} \leq C(h(K))^{k+1-m} \|v\|_{k+1,K} , \qquad m = 0, 1, \tag{4.17}$$

$$\|v - \pi_K v\|_{0,K'} \leq C(h(K))^{k+\frac{1}{2}} \|v\|_{k+1,K} \tag{4.18}$$

where K' is any side of K.

Let $K_1, K_2, \ldots, K_I$ be a fixed ordering of the elements of $\mathcal{T}_h$ which satisfies the condition of Lemma 3. For all $i = 1, \ldots, I$, we set

$$\Omega_i = \overset{\circ}{\bigcup_{j=1}^{i}} K_j \tag{4.19}$$

and we define $\partial_+\Omega_i$ and $\partial_-\Omega_i$ in the usual way. Note that $\partial_-\Omega_i \subset \Gamma_-$.

Theorem 5.

Assume that Hypotheses H.1 and H.2 hold. Assume in addition that the solution u of problem (1.3) belongs to $H^{k+1}(\Omega)$. Then, there exists a constant $C > 0$ independent of h such that for all $i = 1, \ldots, I$

$$\|u_h - u\|_{0,\Omega_i} \leq Ch^k \|u\|_{k+1,\Omega_i}, \tag{4.20}$$

$$\left(\int_{\partial_+\Omega_i} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u_h - u)^2 ds\right)^{\frac{1}{2}} \leq Ch^k \|u\|_{k+1,\Omega_i}, \tag{4.21}$$

$$\left(\sum_{j=1}^{i} \int_{\partial_- K_j} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u_h - \xi_h)^2 ds\right)^{\frac{1}{2}} \leq Ch^k \|u\|_{k+1,\Omega_i}. \tag{4.22}$$

Proof.

For any $K \in \mathcal{T}_h$, we define

$$\eta_h = \begin{cases} 0 & \text{on } \partial_- K \cap \Gamma_-, \\ \text{outward trace of } \pi_h u & \text{on } \partial_- K \setminus (\partial_- K \cap \Gamma_-). \end{cases} \tag{4.23}$$

We start from equation (4.1) with $v = \pi_h u$, $\eta = \eta_h$ and we estimate the corresponding right hand side member. First, we have[†]

$$\left|\int_K (u - \pi_h u) A^* (u_h - \pi_h u)\,dx\,dy\right| \leq c_1 \|u - \pi_h u\|_{0,K} \|u_h - \pi_h u\|_{1,K}$$

[†] In the sequel, we shall denote by c_i various constants independent of h.

and by (4.15), (4.17)

$$|\int_K (u-\pi_h u)A^*(u_h-\pi_h u)dxdy| \leq c_2 h^k \|u\|_{k+1,K}\|u_h-\pi_h u\|_{0,K}. \tag{4.24}$$

Next, using (4.16) and (4.18), we obtain

$$|\int_{\partial_+ K} \underline{\underline{m}}\cdot\underline{\underline{n}}\ (u-\pi_h u)(u_h-\pi_h u)ds| \leq c_3 h^k \|u\|_{k+1,K}\|u_h-\pi_h u\|_{0,K}. \tag{4.25}$$

Similarly, we get

$$|\int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\ (u-\eta_h)(u_h-\pi_h u)ds| \leq c_4 h^k \|u\|_{k+1,\mathfrak{K}_K}\|u_h-\pi_h u\|_{0,K} \tag{4.26}$$

where $\mathfrak{K}_K$ is the union of the elements of $\mathfrak{T}_h$ which have a side contained in $\partial_- K$.

Thus, combining (4.1) with $v = \pi_h u$, $\eta = \eta_h$, (4.24), (4.25), (4.26) and using (3.7), we obtain

$$\begin{aligned}&\tfrac{1}{2}\int_{\partial_+ K} \underline{\underline{m}}\cdot\underline{\underline{n}}\ (u_h-\pi_h u)^2 ds \\ &-\tfrac{1}{2}\int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\ ((u_h-\pi_h u)-(\xi_h-\eta_h))^2 ds + \alpha\|u_h-\pi_h u\|^2_{0,K} \\ \leq\ &-\tfrac{1}{2}\int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\ (\xi_h-\eta_h)^2 ds + c_5 h^k \|u\|_{k+1,K\cup\mathfrak{K}_K}\|u_h-\pi_h u\|_{0,K}.\end{aligned}$$

Summing over all the elements K_j, $1 \leq j \leq i$, and using (3.12), (4.23), we get

$$\begin{aligned}&\tfrac{1}{2}\int_{\partial_+ \Omega_i} \underline{\underline{m}}\cdot\underline{\underline{n}}\ (u_h-\pi_h u)^2 ds \\ &-\tfrac{1}{2}\sum_{j=1}^{i}\int_{\partial_- K_j} \underline{\underline{m}}\cdot\underline{\underline{n}}\ ((u_h-\pi_h u)-(\xi_h-\eta_h))^2 ds + \alpha\|u_h-\pi_h u\|^2_{0,\Omega_i} \\ &\leq c_6 h^k \|u\|_{k+1,\Omega_i}\|u_h-\pi_h u\|_{0,\Omega_i}.\end{aligned} \tag{4.27}$$

From (4.17) and (4.27), we deduce

$$\|u_h - u\|_{0,\Omega_i} \leq \|u_h - \pi_h u\|_{0,\Omega_i} + \|\pi_h u - u\|_{0,\Omega_i}$$

$$\leq \frac{c_6}{\alpha} h^k \|u\|_{k+1,\Omega_i} + c_7 h^{k+1} \|u\|_{k+1,\Omega_i}$$

so that (4.20) holds.

Next, we have by (4.27)

$$\left(\int_{\partial_+\Omega_i} \underline{\underline{m}}\cdot\underline{\underline{n}} \; (u_h - \pi_h u)^2 ds\right)^{\frac{1}{2}} \leq c_8 h^k \|u\|_{k+1,\Omega_i}$$

and by (4.18)

$$\left(\int_{\partial_+\Omega_i} \underline{\underline{m}}\cdot\underline{\underline{n}} \; (\pi_h u - u)^2 ds\right)^{\frac{1}{2}} \leq c_9 h^{k+\frac{1}{2}} \|u\|_{k+1,\Omega_i} .$$

This proves inequality (4.21).

Similarly, we have by (4.27)

$$\left(\sum_{j=1}^{i} \int_{\partial_-\Omega_i} \underline{\underline{m}}\cdot\underline{\underline{n}} \; ((u_h - \pi_h u) - (\xi_h - \eta_h))^2 ds\right)^{\frac{1}{2}}$$

$$\leq c_{10} h^k \|u\|_{k+1,\Omega_i}$$

and by (4.18)

$$\left(\sum_{j=1}^{i} \int_{\partial_- K_j} \underline{\underline{m}}\cdot\underline{\underline{n}} \; (\pi_h u - u)^2 ds\right)^{\frac{1}{2}} \leq c_{11} h^{k+\frac{1}{2}} \|u\|_{k+1,\Omega_i} ,$$

$$\left(\sum_{j=1}^{i} \int_{\partial_- K_j} \underline{\underline{m}}\cdot\underline{\underline{n}} \; (\eta_h - u)^2 ds\right)^{\frac{1}{2}} \leq c_{12} h^{k+\frac{1}{2}} \|u\|_{k+1,\Omega_i} .$$

This implies inequality (4.22). ∎

5. A Superconvergence Result.

Let us notice that the error estimates of Theorem 5 are not optimal in the exponent of the parameter h. In fact, numerical calculations have shown that these error bounds could not be improved in general. However, the one-

dimensional results of §2 clearly indicate that better estimates must hold in some special cases. Indeed, we shall prove in this section that the rate of convergence of our finite element method is $O(h^{k+1})$ when all the elements $K \in \mathfrak{T}_h$ are rectangles and when $\hat{P} = Q_k$. In the sequel, we shall confine ourselves to this particular case.

On the interval $[-1,+1]$, let $-1<\theta_1<\theta_2<\ldots<\theta_{k+1}=1$ denote the $(k+1)$ Gauss-Radau quadrature abscissae. In the reference square $\hat{K} = [-1,+1]^2$, we consider the points $\hat{a}_{ij}$ with coordinates (θ_i, θ_j), $1 \le i,j \le k+1$.

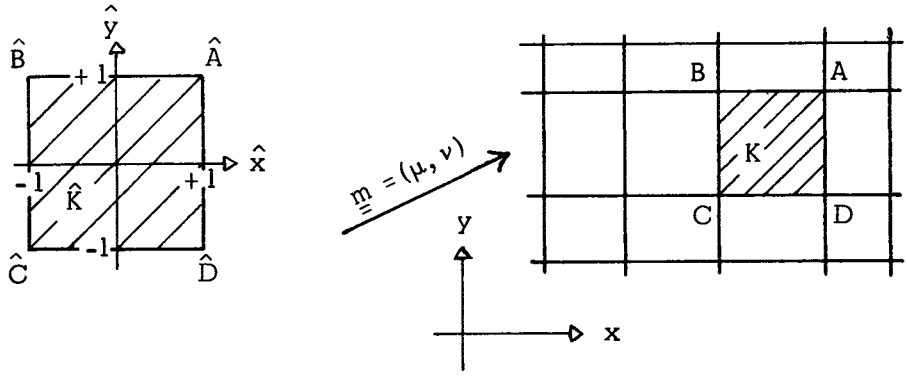

Figure 5

Just for convenience, we shall assume that the sides of the rectangles $K \in \mathfrak{T}_h$ are parallel to the (x,y) axes and that the coefficients μ, ν are > 0. Given a rectangle K with vertices A, B, C, D as in Figure 5, we denote by F_K the affine invertible mapping such that $A = F_K(\hat{A}), \ldots, D = F_K(\hat{D})$. Given a function $v \in C^0(K)$, we define $r_K v$ as the unique polynomial in Q_k which interpolates v at the points $a_{ij} = F_K(\hat{a}_{ij})$, $1 \le i,j \le k+1$. Then, for any $v \in C^0(\Omega)$, we define $r_h v$ to be the function in V_h such that

$$r_h v|_K = r_K v \quad \text{for all } K \in \mathfrak{T}_h . \tag{5.1}$$

We provide $L_\infty(\Omega)$ with the following norm

$$\|v\|_{0,\infty,\Omega} = \sup\{ |v(x)| ; x \in \Omega \}.$$

Given any integer $m \geq 0$, let

$$W_\infty^m(\Omega) = \{ v \in L_\infty(\Omega) \mid \partial^\alpha v \in L_\infty(\Omega), \ |\alpha| \leq m \}$$

be the Sobolev space provided with the norm

$$\|v\|_{m,\infty,\Omega} = \max\{ \|\partial^\alpha v\|_{0,\infty,\Omega}; \ |\alpha| \leq m \} .$$

Using [4], for instance, one can easily prove

Lemma 7.

Assume that Hypothesis H.2 holds. Then there exists a constant $c > 0$ independent of $K \in \mathfrak{T}_h$ such that

$$\|v - r_K v\|_{0,K} \leq c(h(K))^{k+1} \|v\|_{k+1,K} \quad \text{for all} \quad v \in H^{k+1}(K), \tag{5.2}$$

$$\|v - r_K v\|_{0,K'} \leq c(h(K))^{k+3/2} \|v\|_{k+1,\infty,K} \quad \text{for all} \tag{5.3}$$

$v \in W_\infty^{k+1}(K)$ where K' is any side of $\partial_+ K$.

We are now able to prove

Theorem 6.

Assume that all the elements $K \in \mathfrak{T}_h$ are rectangles, that $\hat{P} = Q_k$, and that Hypothesis H.2 holds. Assume, in addition, that the solution u of problem (1.3) belongs to $H^{k+2}(\Omega) \cap W_\infty^{k+1}(\Omega)$. Then, there exists a constant $C > 0$ independent of h such that, for all $i = 1, \ldots, I$,

$$\|u_h - u\|_{0,\Omega_i} \leq C h^{k+1} \|u\|_{k+2,\Omega_i} , \tag{5.4}$$

$$\left(\int_{\partial_+ \Omega_i} \underline{\underline{m}} \cdot \underline{\underline{n}} \ (u_h - u)^2 ds\right)^{\frac{1}{2}}$$

$$\leq C h^{k+1} (\|u\|_{k+2,\Omega_i} + \|u\|_{k+1,\infty,\Omega_i}) . \tag{5.5}$$

Proof.

For any $K \in \mathfrak{T}_h$, we now define

$$\eta_h = \begin{cases} 0 & \text{on } \partial_- K \cap \Gamma_-, \\ \text{outward trace of } r_h u & \text{on } \partial_- K \setminus (\partial_- K \cap \Gamma_-). \end{cases} \tag{5.6}$$

We start from equation (4.1) with $v = r_h u$, $\eta = \eta_h$. The corresponding right hand side may be written in the form

$$X_K(u, u_h - r_h u) = Z_K(u, u_h - r_h u) + \int_K \sigma(u - r_h u)(u_h - r_h u)\,dx\,dy \tag{5.7}$$

where

$$\begin{aligned} Z_K(u,w) &= \int_{\partial_+ K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u - r_h u)w\,ds \\ &\quad + \int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u - \eta_h)w\,ds \\ &\quad - \int_K (u - r_h u)\,\underline{\underline{m}}\cdot\operatorname{grad} w\,dx\,dy\,. \end{aligned} \tag{5.8}$$

We now use the following essential lemma which will be proved later.

Lemma 8.

With the same assumptions as in Theorem 6, there exists a constant $C > 0$ independent of $K \in \mathfrak{T}_h$ such that for all $w \in Q_k$

$$|Z_k(u,w)| \le C(h(K))^{k+1} \|u\|_{k+2,K} \|w\|_{0,K}\,. \tag{5.9}$$

Using (5.2), (5.7) and (5.9), we obtain for all $K \in \mathfrak{T}_h$

$$|X_K(u, u_h - r_h u)| \le c_1 h^{k+1} \|u\|_{k+2,K} \|u_h - r_h u\|_{0,K}\,. \tag{5.10}$$

Thus, combining (4.1) with $v = r_h u$, $\eta = \eta_h$, (5.10) and using (3.7), we get

$$\tfrac{1}{2}\int_{\partial_+ K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u_h - r_h u)^2 ds + \alpha \|u_h - r_h u\|^2_{0,K}$$

$$\leq -\tfrac{1}{2}\int_{\partial_- K} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(\xi_h - \eta_h)^2 ds + c_1 h^{k+1} \|u\|_{k+2,K} \|u_h - r_h u\|_{0,K}.$$

Summing over all the elements K_j, $1 \leq j \leq i$, we obtain

$$\tfrac{1}{2}\int_{\partial_+ \Omega_i} \underline{\underline{m}}\cdot\underline{\underline{n}}\,(u_h - r_h u)^2 ds + \alpha \|u_h - r_h u\|^2_{0,\Omega_i}$$

(5.11)

$$\leq c_1 h^{k+1} \|u\|_{k+2,\Omega_i} \|u_h - r_h u\|_{0,\Omega_i}.$$

Thus, the estimates (5.4) and (5.5) are simple consequences of inequality (5.11) and Lemma 7. ∎

Proof of Lemma 8.

Consider a rectangle $K \in \mathfrak{T}_h$ with vertices A, B, C, D (cf. Figure 5). Let us denote by Δx (resp. Δy) the length of the side AB (resp. BC). We may write

$$Z_K(u,w) = \mu Z_{K,x}(u,w) + \nu Z_{K,y}(u,w) \tag{5.12}$$

with

$$Z_{K,x}(u,w) = \int_D^A (u - r_h u) w\, dy - \int_C^B (u - \eta_h) w\, dy - \int_K (u - r_h u) \frac{\partial w}{\partial x}\, dxdy,$$

$$Z_{K,y}(u,w) = \int_B^A (u - r_h u) w\, dx - \int_C^D (u - \eta_h) w\, dx - \int_K (u - r_h u) \frac{\partial w}{\partial y}\, dxdy.$$

By using the one-to-one correspondence $v \mapsto \hat{v} = v \circ F_K$, we get

$$Z_{K,x}(u,w) = \frac{\Delta y}{2} \hat{Z}_{\hat{x}}(\hat{u},\hat{w}) \tag{5.13}$$

with

$$\hat{Z}_{\hat{x}}(\hat{u},\hat{w}) = \int_{-1}^{+1}(\hat{u}(1,\hat{y})-\tilde{\hat{r}}\hat{u}(1,\hat{y}))\hat{w}(1,\hat{y})d\hat{y}$$

$$-\int_{-1}^{+1}(\hat{u}(-1,\hat{y})-\hat{\eta}(\hat{y}))\,\hat{w}(-1,\hat{y})d\hat{y}$$

$$-\int_{-1}^{+1}\int_{-1}^{+1}(\hat{u}-\hat{r}\hat{u})\,\frac{\partial\hat{w}}{\partial\hat{y}}\,d\hat{x}d\hat{y},$$

where $\hat{r}\hat{u}$ is the polynomial in Q_k which interpolates $\hat{u}$ at the points $\hat{a}_{ij}$, $1 \le i,j \le k+1$, and where $\hat{\eta}$ is the polynomial of degree $\le k$ which interpolates the function $\hat{y} \mapsto \hat{u}(-1,\hat{y})$ at the points θ_i, $1 \le i \le k+1$.

Clearly

$$\hat{Z}_{\hat{x}}(\hat{u},\hat{w}) = 0 \quad \text{for all} \quad \hat{u},\hat{w} \in Q_k .$$

Now, when $\hat{u} = \hat{x}^{k+1}$, we have

$$\hat{u}(1,\hat{y}) = \hat{r}\hat{u}(1,\hat{y}) = 1 , \quad \hat{u}(-1,\hat{y}) = \hat{\eta}(\hat{y}) = (-1)^{k+1} .$$

Moreover, $\hat{r}\hat{u}$ does not depend on $\hat{y}$ and then, for all $\hat{w} \in Q_k$, the function $\hat{x} \mapsto (\hat{u}-\hat{r}\hat{u})(\hat{x})\,\frac{\partial\hat{w}}{\partial\hat{x}}(\hat{x},\hat{y})$ is a polynomial of degree $\le 2k$ which vanishes at the (k+1) Gauss-Radau points θ_i. Therefore,

$$\int_{-1}^{+1}(\hat{u}-\hat{r}\hat{u})\hat{w}\,d\hat{x} = 0 \quad \text{for all} \quad \hat{w} \in Q_k .$$

Thus, when $\hat{u} = \hat{x}^{k+1}$, we get

$$\hat{Z}_{\hat{x}}(\hat{u},\hat{w}) = 0 \qquad \text{for all} \quad \hat{w} \in Q_k .$$

On the other hand, when $\hat{u} = \hat{y}^{k+1}$, $\hat{r}\hat{u}$ is independent of $\hat{x}$ so that we obtain by integration by parts

$$\int_{-1}^{+1}\int_{-1}^{+1}(\hat{u}-\hat{r}\hat{u})\,\frac{\partial\hat{w}}{\partial\hat{x}}\,d\hat{x}d\hat{y} = \int_{-1}^{+1}(\hat{u}(1,\hat{y}) - \hat{r}\hat{u}(1,\hat{y}))\hat{w}(1,\hat{y})d\hat{y}$$

$$-\int_{-1}^{+1}(\hat{u}(-1,\hat{y})-\hat{\eta}(\hat{y}))\hat{w}(-1,\hat{y})d\hat{y}.$$

This gives again

$$\hat{Z}_{\hat{x}}(\hat{u},\hat{w}) = 0 \quad \text{for all } \hat{w} \in Q_k .$$

Therefore, we have proved that

$$\hat{Z}_{\hat{x}}(\hat{u},\hat{w}) = 0 \quad \text{for all } \hat{u} \in P_{k+1} \text{ and all } \hat{w} \in Q_k .$$

Then, for fixed $\hat{w} \in Q_k$, the linear functional $\hat{u} \mapsto \hat{Z}_{\hat{x}}(\hat{u},\hat{w})$ is continuous on $H^{k+2}(\hat{K})$, has norm $\leq c_1 \|\hat{w}\|_{0,\hat{K}}$ and vanishes over P_{k+1}. By the Bramble-Hilbert lemma in the form given in [4, Lemma 6], we get for all $\hat{u} \in H^{k+2}(\hat{K})$ and all $\hat{w} \in Q_k$

$$|\hat{Z}_{\hat{x}}(\hat{u},\hat{w})| \leq c_2 |\hat{u}|_{k+2,\hat{K}} \|\hat{w}\|_{0,\hat{K}} .$$

Going back to the element K by using the correspondence $\hat{v} \mapsto v = \hat{v} \circ F_K^{-1}$ and (5.13), we obtain for all $u \in H^{k+2}(K)$ and all $w \in Q_k$

$$|Z_{K,x}(u,w)| \leq c_3 (h(K))^{k+1} \|u\|_{k+2,K} \|w\|_{0,K} . \tag{5.14}$$

Likewise, we get

$$|Z_{K,y}(u,w)| \leq c_4 (h(K))^{k+1} \|u\|_{k+2,K} \|w\|_{0,K} . \tag{5.15}$$

Then, combining (5.12), (5.14) and (5.15), we obtain the desired inequality (5.9). ∎

Note that the error estimates of Theorem 6 are now optimal in the exponent of the parameter h. However, as the one-dimensional results of §1 suggest, we conjecture that, for any rectangle $K \in \mathfrak{I}_h$, there exist some points of $\partial_+ K$ where even more precise error bounds hold. Unfortunately, we have not been able to prove the existence of such points.

References.

1. Axelsson, O., A class of A-stable methods, B.I.T. 9 (1969), 185-199.

2. de Boor, C., and B. Swartz, Collocation at Gaussian points, SIAM J. Numer. Anal. 10 (1973), 582-606.

3. Butcher, J. C., Implicit Runge-Kutta processes, Math. Comp. 18 (1964), 50-64.

4. Ciarlet, P. G., and P.-A. Raviart, General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods, Arch. Rat. Mech. Anal. 46 (1972), 177-199.

5. Ciarlet, P. G., and P.-A. Raviart, Interpolation theory over curved elements, with applications to finite element methods, Comp. Meth. Appl. Mech. Eng. 1 (1972), 217-249.

6. Ciarlet, P. G., and P.-A. Raviart, The combined effect of curved boundaries and numerical integration in isoparametric finite element methods, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (A.K. Aziz, Ed.), 409-474, Academic Press. New York, 1972.

7. Crouzeix, M., Thesis, to appear.

8. Friedrichs, K. O., Symmetric positive differential equations, Comm. Pure Appl. Math. 11 (1958), 333-418.

9. Hulme, B. L., Discrete Galerkin and related one-step methods for ordinary differential equations, Math. Comp. 26 (1972), 881-891.

10. Kaper, H. G., G. K. Leaf, and A. J. Lindeman, Application of finite element techniques for the numerical solution of the neutron transport and diffusion equations, Proc. Conf. on Transport Theory, 2nd Conf-710107, Los Alamos (1971).

11. Lesaint, P., Finite element methods for symmetric hyperbolic equations, Numer. Math. 21 (1973), 244-255.

12. Lesaint, P., Finite element methods for the transport equation, to appear in RAIRO, Série Mathématiques.

13. Lesaint, P., Thesis, to appear.

14. Lesaint, P., and J. Gérin-Roze, Isoparametric finite element methods for the neutron transport equation, to appear.

15. Miller, W. F. Jr., E. E. Lewis, and E. C. Rossow, The application of phase-space finite elements to the two-dimensional transport equation in x-y geometry, to appear in Nucl. Sci. Eng.

16. Ohnishi, T., Application of finite element solution technique to neutron diffusion and transport equations, Proc. Conf. on New Developments in Reactor Mathematics and Applications, Conf.-710302, Idaho Falls (1971).

17. Reed, W. H., and T. R. Hill, Triangular mesh methods for the neutron transport equation, to appear in Proc. Amer. Nucl. Soc.

FINITE ELEMENTS AND THE NEUTRON TRANSPORT EQUATION

P. Lesaint
Service de Mathématiques Appliquées,
Centre d'Etudes de Limeil, B. P. 27
94190 VILLENEUVE-St. Georges,
France

P. -A. Raviart
Analyse Numerique, Tour 55-65,
Université de Paris VI
4, Place Jussieu,
75230, Paris CEDEX 05,
France