On the strong maximum principle for some piecewise linear finite element approximative problems of non-positive type

Vitoriano Ruas

IRIA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105 78150 Le Chesnay
France
Tél. 954 90 20

Rapports de Recherche

N° 43

ON THE STRONG MAXIMUM PRINCIPLE FOR SOME PIECEWISE LINEAR FINITE ELEMENT APPROXIMATE PROBLEMS OF NON-POSITIVE TYPE

Vitoriano RUAS

Novembre 1980

ERRATA for the Rapport de Recherche N° 43

Page 8 , line 3 : ... + $C^2(D^- + CD^+) + \ldots$

Page 19, line 10 : replace the 2nd denominator by 96 ;

line 16 : $m_d \leq M \|a\|_\infty \quad S/6$

line 21 : replace the whole line by

$$(17) \qquad S \leq \frac{3 \sin^2\theta \operatorname{tg} \theta}{\|a\|_\infty (4M+1)}$$

Page 20, line 3 : ... like in Figure 6 ...

line 16 : ... where $\alpha_1 + \alpha_2 \leq \pi + 2\varepsilon$ ...

ON THE STRONG MAXIMUM PRINCIPLE FOR SOME PIECEWISE LINEAR FINITE ELEMENT APPROXIMATE PROBLEMS OF NON-POSITIVE TYPE

by

Vitoriano RUAS

RESUME

On considère d'abord l'approximation par éléments finis linéaires par morceaux du problème modèle $-\Delta u = f$ dans un domaine bi-dimensionnel avec des conditions aux limites non-homogènes.

On sait que si la somme de deux angles opposés à tout côté de la triangulation utilisée ne dépasse pas π, alors le problème approché satisfait à un principe du maximum fort. Dans ce travail on démontre la validité d'un tel résultat pour une classe plus générale de triangulations présentant de telles paires d'angles obtus.

La technique utilisée est ensuite appliquée à d'autres problèmes associés de type non-positif.

SUMMARY

We first consider the piecewise linear finite element approximation of the model problem $-\Delta u = f$ in a two-dimensional domain, with non-homogeneous boundary conditions. It is well known that, if no pair of angles opposite to a given edge of the triangulation add to more than π, then the corresponding discrete problem satisfies a maximum principle. In this paper we prove the validity of such a result for a wider class of triangulations where such pairs of obtuse angles are admissible.

The results are extended to other related problems.

1. INTRODUCTION.

The strong maximum principle has proven to be a powerful tool in deriving pointwise convergent results for approximate solutions of partial differential equations. Furthermore, for physical reasons a discrete problem is often required to satisfy such a principle, whenever the continuous problem does.

As far as the finite element method is concerned, it appears that the results available in this connection provide less than it is actually to be expected from corresponding approximate problems. In this paper we intend to illustrate this fact by giving conditions for the strong maximum principle to hold, for piecewise linear finite element approximations of a wide class of problems, that are more general than those commonly admitted so far.

Let us first state the strong maximum principle we are to consider for the continuous problem, as given in [6].

A linear second order partial differential operator L defined on a space of suitably smooth functions, which are in turn defined on a bounded and connected open set Ω of $\mathbb{R}^2$ ($\mathbb{R}^N$) with boundary $\partial\Omega$, is said to satisfy the maximum principle (in its strong form) if

$$Lu(x) \geq 0 \quad \forall x \in \Omega \quad \text{and} \quad u(x) \geq 0 \quad \forall x \in \partial\Omega$$

imply that

$$u(x) \geq 0 \quad \forall x \in \Omega \ .$$

In particular, if function u is prescribed on the boundary of Ω, say, $u=g$ on $\partial\Omega$, then whenever $Lu(x) \leq 0 \quad \forall x \in \Omega$ we have :

$$(1) \qquad \max_{x \in \overline{\Omega}} u(x) \leq \max \{0, \max_{x \in \partial\Omega} g(x)\} \ .$$

If additionally L satisfies the condition : $Lw \equiv 0$ whenever w is a constant valued function, then we can say that if $Lu(x) \leq 0 \quad \forall x \in \Omega$, the following maximum principle holds :

$$(2) \qquad \max_{x \in \overline{\Omega}} u(x) \leq \max_{x \in \partial\Omega} g(x) \ .$$

As an important particular case, we have that of the operator $aI-\Delta$, where Δ is the Laplacian, I is the identity operator and a is an essentially bounded real non-negative valued function. We may say that the following problem (or its solution u)

$$(\mathcal{P}_a)\begin{cases} au-\Delta u = f \text{ in } \Omega \text{ with } a \not\equiv 0 \\ u=g \text{ on } \partial\Omega \end{cases}$$

and problem

$$(\mathcal{P}_o)\begin{cases} -\Delta u = f \text{ in } \Omega \\ u=g \text{ on } \partial\Omega \end{cases}$$

satisfy respectively the maximum principles (1) and (2), provided that f is a non-positive given function.

Now, in the case of piecewise linear finite element approximations of problems $(\mathcal{P}_a)$ and $(\mathcal{P}_o)$, it was proved by Ciarlet & Raviart [3] that analogous strong maximum principles will hold, provided that the triangulation of Ω is of acute type for the latter, and of strictly acute type together with some suitable boundedness assumptions on the mesh size for the former. Here the expression "triangulation of acute type" means that the maximal angle of all the triangles is less than or equal to $\pi/2$, and less than $\pi/2$ in the strict case.

As remarked in [7, page 78], such conditions can be weakened to the following one(s) :

For every pair (α_1,α_2) of angles opposite to a common edge of a given pair of adjacent triangles of the partition (ref. Figure 1), we have :

$$(3_o) \qquad \alpha_1+\alpha_2 \leq \pi \quad (\text{resp. } (3_a)\ \alpha_1+\alpha_2 < \pi)$$

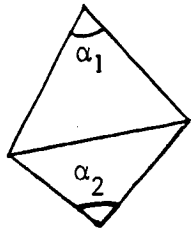


Figure 1

A typical case where (3_o) is fulfilled.

Remark : Whenever the edge under consideration lies on the boundary of the domain, the associated coefficient is zero, irrespective of the opposite angle.

In both cases the argument is the same : the so-generated discrete problem is of non-negative type [3]. This means that the finite element approximate problem leads to the solution of an $\bar{n}\times\bar{n}$ system of linear equations of the form :

$$\bar{A}\bar{u} = \bar{b} \tag{4}$$

where matrix $\bar{A}$, besides being non singular, satisfies :

$$\text{(5)} \quad \begin{cases} \text{a)} & a_{ii} > 0 \quad i=1,2,\ldots,\bar{n} \\ \text{b)} & \sum_{j=1}^{\bar{n}} a_{ij} \geq 0 \quad i=1,2,\ldots,\bar{n} \\ \text{c)} & a_{ij} \leq 0 \quad \text{for } i\neq j \quad 1 \leq i \leq \bar{n} \ , \ 1 \leq j \leq \bar{n} \ . \end{cases}$$

Indeed, a regular matrix that satisfies (5) is known to have a positive(*) inverse, which is a crucial fact in the assertion of the maximum principle [2].

Now, to be more specific, we will give in detail the structure that we are considering for system (4).

Let $\{\phi_i\}_{i=1}^{n}$ be the set of basis functions of the n-dimensional finite element subspace, for the chosen triangulation of Ω, associated to the set of n degrees of freedom, that is to say, the vector $u = (u_1,u_2,\ldots,u_n)$ of unknown values of the solution at the inner nodal points. Similarly we denote by $\{\phi_i^{\partial}\}$, i = 1,..., m, the boundary basis functions, i.e. those associated with the m nodes lying on $\partial\Omega$, at which the values of u are prescribed to be respectively $g_1,g_2,\ldots,g_m$. We set $\bar{n} = n+m$ and we define :

$$\bar{u} = (u_1,u_2,\ldots,u_n,u_{n+1},\ldots,u_{\bar{n}})$$

where $u_{n+i} = g_i \quad i=1,2,\ldots,m.$

(*) The meaning of the positiveness being that every element of the matrix is non-negative. A regular matrix whose inverse is positive is commonly called a monotone matrix.

Now, letting $a(\cdot,\cdot)$ be the bilinear form associated with the problem under consideration, we define the basic finite element matrix A to be the n×n matrix whose entries are :

$$a_{ij} = a(\phi_i,\phi_j) \qquad 1 \le i,j \le n$$

and an n×m boundary matrix $A^{\partial} = \{a^{\partial}_{ij}\}$ such that :

$$a^{\partial}_{ij} = a(\phi_i,\phi_j) \qquad 1 \le i \le n \quad , \quad 1 \le j \le m \ .$$

We also set $b_i = \int_\Omega f\phi_i$, $i=1,2,\ldots,n$.

We actually have to solve an n×n system of equations of the form

$$Au = b - A^{\partial} g \quad \text{where} \quad g = (g_1,g_2,\ldots,g_m) \ .$$

However it will be convenient to consider an extended form of this system, namely, the $\bar{n}\times\bar{n}$ linear system of equations (4), where $\bar{b} = (b_1,b_2,\ldots,b_n,g_1,\ldots,g_m)$ and $\bar{A}$ is the $\bar{n}\times\bar{n}$ finite element matrix whose structure is given below :

$$(6) \qquad [\bar{A}] = \begin{array}{c} \begin{array}{cc} n & m \end{array} \\ \left[\begin{array}{c|c} A & A^{\partial} \\ \hline 0 & I \end{array}\right] \end{array} \begin{array}{l} n \\ m \end{array}$$

Note that both A and $\bar{A}$ will satisfy (5a) and (5b) for the problems that we are to consider, in particular, problems $(\mathcal{P}_a)$ and $(\mathcal{P}_o)$.

Now, even when the finite element matrix $\bar{A}$ does not have all the properties (5), it's inverse can still be positive (as it often appears to be the case of discrete problems for which direct computations are performed). As a matter of fact, if a matrix associated with a finite element discrete problem of non-positive type can be viewed as a certain perturbation of another matrix that does satisfy (5), we can prove that its inverse is positive and, as a consequence, the validity of the strong maximum principle.

The basic tools for achieving this are, a matrix decomposition theorem by Bramble & Hubbard [1] that they used for deriving similar results for non positive finite difference schemes, together with suitable assumptions on the structure of the finite element mesh, which, as one will see, are far from restrictive.

An outline of the paper is as follows :

In Section 2 we give some theoretical preliminaries that are going to be used in the subsequent sections, including the above mentioned theorem, and a general result due to Ciarlet [2] that was previously used in [3] for their own proofs.

In Section 3 we give our conditions for a matrix associated with piecewise linear finite element discretization to fulfill, so that its inverse exists and is positive, and we apply the results to problem $(\mathcal{P}_o)$.

As a conclusion, in Section 4 the application of the above analysis to $(\mathcal{P}_a)$, to other elliptic problems and to the heat equation is briefly discussed.

2. PRELIMINARIES.

Let us first give Bramble & Hubbard's decomposition theorem. Assume that an n×n matrix A satisfies (5a) and (5b). We also assume that there exists a non-empty set J(A) of rows of A, such that for every $k \in J(A)$ we have :

$$\sum_{j=1}^{n} a_{kj} > 0$$

Now for $i \notin J(A)$, we define a connection in A from i to J(A) to be a finite sequence of non-zero elements of the form $a_{ij_1}\ a_{j_1j_2} \cdots a_{j_sk}$, where $k \in J(A)$.

We finally assume that there exists at least one such a connection in A, for every $i \notin J(A)$. Note that whenever A is a finite element matrix for problems such as those we are to consider, this assumption is trivially fulfilled.

Now let us denote by A_d the diagonal matrix whose diagonal coincides with that of A.

We refer to [1, page 352] for the proof of the following :

Theorem 1 (Bramble & Hubbard) : Let B have unit diagonal, satisfy (5b) and have non-empty J(B). If we can write it as

$$B = I-C-D$$

where

$$(7)\quad \begin{cases} \text{a) } C_d = 0 \\ \text{b) } I-C \text{ satisfies (5b) and (5c)} \\ \text{c) } (I-C)^{-1}D \text{ is positive} \\ \text{d) for each } i \notin J(B) \text{ there exists a connection in C from i to J(B).} \end{cases}$$

then B is monotone (i.e. B^{-1} exists and is positive). ■

As a matter of fact, the matrix B that we are going to deal with, is the normalized finite element matrix given by :

$$B = \bar{A}_d^{-1}\,\bar{A} \ .$$

We clearly see that the first two assumptions on B of Theorem 1 are fulfilled. On the other hand, the simple fact that $\bar{A}$ is a finite element matrix is enough to guarantee a non-empty J(B) (actually, for problem $(\mathcal{P}_a)$, for instance, we will have $J(B) = \{1,2,\ldots,\bar{n}\}$).

As far as the conditions (7) on the splitting of B are concerned, as one will see, only (7c) will be difficult to verify. However, as remarked in [1], for a matrix C that satisfies (7b) we can say that the Neumann expansion :

$$(I-C)^{-1} = I+C+C^2+\ldots$$

converges. Now if we define respectively the positive and negative parts of D to be :

$$D^+ = \{d_{ij}^+\} \quad \text{and} \quad D^- = \{d_{ij}^-\}$$

where for any real number c, c^+ and c^- are given by :

$$c^+ = \max\,(0,c) \quad \text{and} \quad c^- = \min\,(0,c),$$

it suffices to prove that $D^- + CD^+$ is a positive matrix in order to verify (7c).

Indeed, since C and D^+ are positive matrices and $D = D^+ + D^-$, we have :

$$(8)\quad \begin{cases} (I-C)^{-1}D = D+CD + C^2D + C^3D + \ldots \quad \text{which gives} \\ (I-C)^{-1}D = D^+ + (D^- + CD^+) + C(D^- + CD^+) + C^2(C^- + CD^+) + \ldots \end{cases}$$

Let us now turn our attention to the particular aspect of Ciarlet's treatment of the discrete maximum principle which we are interested in. The proof of the following theorem can be found in [2, p. 342]. Here we give it in a somewhat simplified form :

Theorem 2 (Ciarlet) : Assume that for problems $(\mathcal{P}_o)$ and $(\mathcal{P}_a)$ the finite element matrix of system (4) is monotone and that condition (5b) holds. Then whenever $f \leq 0$ the following strong discrete maximum principles apply :

$$(9a)\quad \max_{1\leq i\leq n} u_i \leq \max\{0, \max_{n+1\leq i\leq \bar{n}} u_i\} \quad \text{for problem } (\mathcal{P}_a)$$

$$(9_o)\quad \max_{1\leq i\leq n} u_i \leq \max_{n+1\leq i\leq \bar{n}} u_i \quad \text{for problem } (\mathcal{P}_o). \blacksquare$$

As a summary of this section we would say that, if one can find a suitable splitting of $\bar{A}_d^{-1}\bar{A}$ satisfying (7), then with the help of Theorems 1 and 2 one can conclude the validity of the discrete maximum principles (9_a) and (9_o).

3. THE CASE OF A MODEL DIRICHLET PROBLEM.

We begin by introducing the structure of the triangulation that will be assumed henceforth.

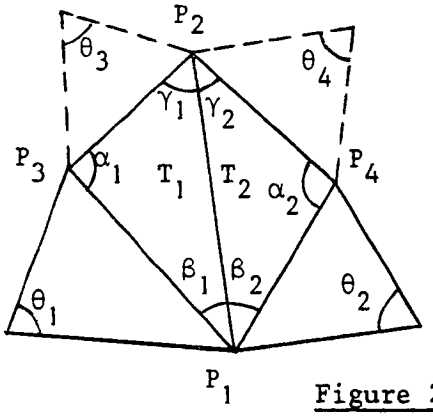
As a model we consider problem $(\mathcal{P}_o)$ and in this case we recall that a strictly positive off-diagonal coefficient will occur whenever condition (3_o) does not hold. Obviously this will be the case where both angles α_1 and α_2 are obtuse, and it is precisely this less favorable case that we are going to consider. For simplicity we denote by P_1 and P_2 the nodes of the edge opposite to such angles α_1 and α_2 where at least one of them, say P_1, does not lie on $\partial\Omega$. In other words, we have $a_{12} > 0$ and whenever $P_2 \notin \partial\Omega$ we have $a_{21} > 0$ as well.

Let us denote by P_3 and P_4 the vertices opposite to segment $\overline{P_1P_2}$ in each neighboring triangle, that are in turn denoted by T_1 and T_2. Let us now refer to Figure 2 and make the following fundamental assumption.

(i) Every (existing) angle opposite to one of the four edges of the quadrilateral $T_1 \cup T_2$ and lying outside the latter, is bounded above by $\pi/2$.

Remark : One or more of such angles may not exist if P_4 or both P_3 and P_4 lie on $\partial\Omega$.

According to assumption (i), one can easily see that the coefficients a_{ij} (or a_{ji}) for i=1,2 and j=3,4 are all strictly negative.



$0 < \theta_i \leq \pi/2$ for i=1,...,J with

J = 4 if $P_2 \notin \partial\Omega$

J = 3 if $P_3 \notin \partial\Omega$ and $P_2, P_4 \in \partial\Omega$

J = 2 if $P_2, P_3, P_4 \in \partial\Omega$.

Figure 2

Local structure of the triangulation around a pair of non acute angles.

Let us also assume for the moment that

(10) $\quad a_{ij} < |a_{ik}|/2$ for i,j=1,2, $i \neq j$ and k=3,4.

As one will see, the above assumption will be trivially satisfied, since we will consider not very big non-acute angles.

Now we represent the part of matrix $B = \bar{A}_d^{-1}\bar{A}$ concerning nodes P_1 and P_2, in a graph structure, as it is easy to visualize it in this way, instead of the associated rows and columns.

In Figure 3 below we indicate the coefficients of the i-th row and j-th column of B, i,j=1,2, that are possibly non-zero. The dotted lines mean non-zero coefficients relating P_i to other neighboring nodes, whose values we do not specify. As a matter of fact, they will be irrelevant to our proof,

since there is no node related to both nodes P_1 and P_2 through a non-zero coefficient other than P_3 and P_4.

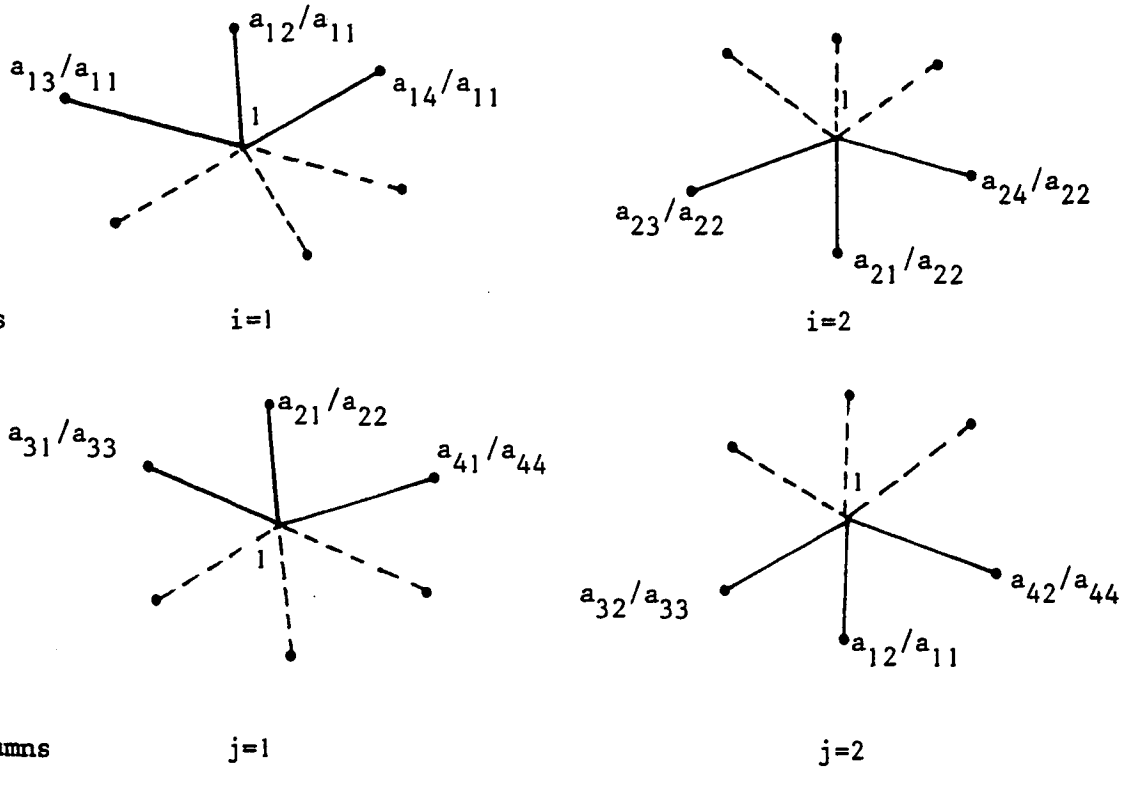


Figure 3

Patterns showing non-zero elements of rows and columns of B.

Now we introduce a splitting of B of the form considered in Theorem 1. In Figure 4 we give the graph structure of relevant two rows and two columns of both C and D which, we conjecture, will be sufficient to illustrate the splitting

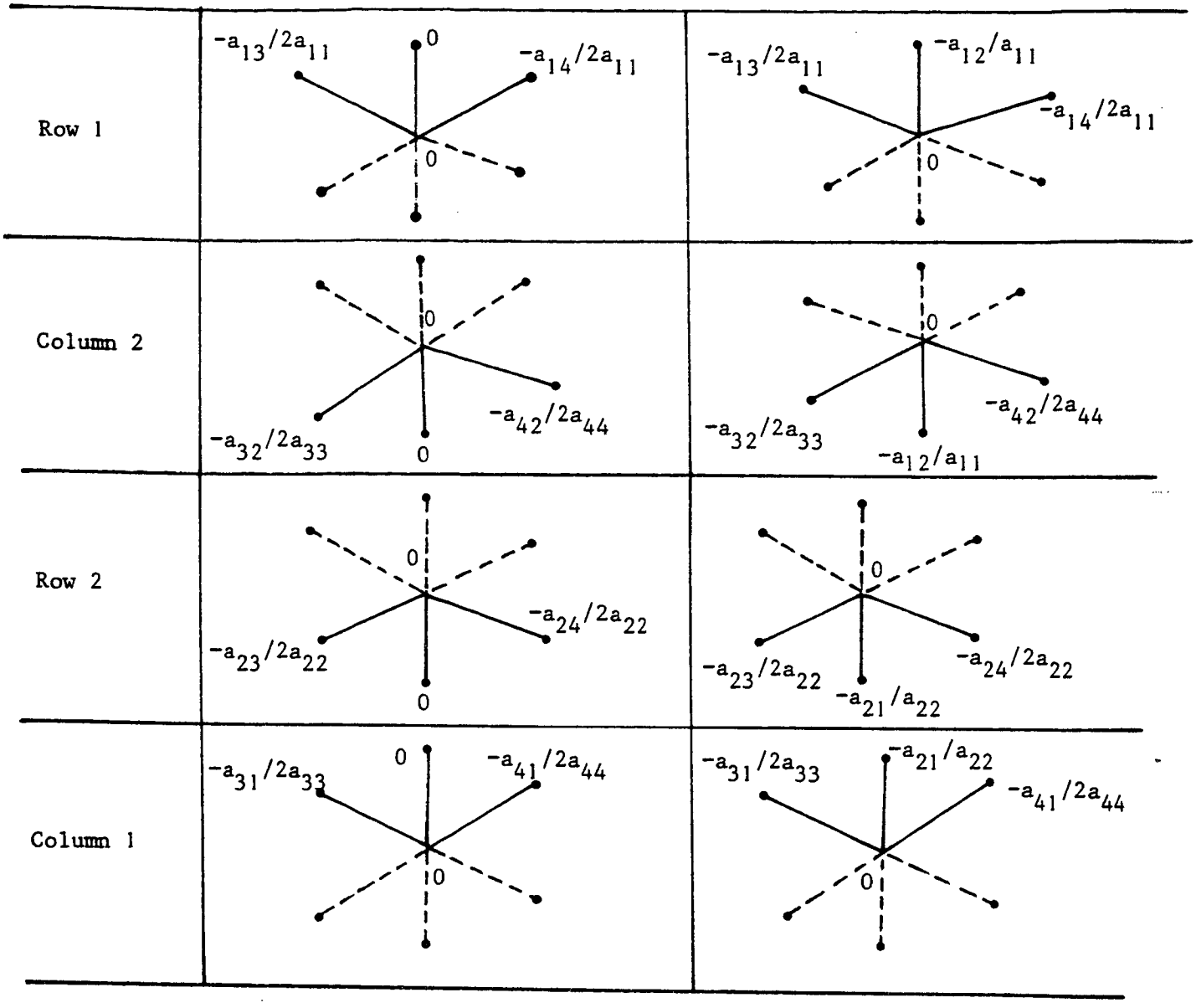


Figure 4

Patterns showing the splitting of B for rows and columns where positive off-diagonal entries appear.

Now, for the sake of simplicity, we will consider at first, the case where none of the nodes P_i, i=1,2,3,4, lies on $\partial\Omega$.

Let us then check that, for the above splitting, all the assumptions (7) are satisfied : this is clearly the case of (7a). As for (7b) we first note that (5c) holds since the a_{14}'s and a_{13}'s are negative. On the other hand we have :

$$1 - \sum_{j=1}^{\bar{n}} c_{ij} = 1 + \sum_{\substack{j \\ j\neq i}} \frac{\bar{a}_{ij}}{2a_{ii}} = 1 + \sum_{\substack{j \\ j\neq i}} \frac{a_{ij}}{a_{ii}} - \sum_{\substack{j \\ j\neq i}} \frac{\bar{a}_{ij}}{2a_{ii}} - \sum_{\substack{j \\ j\neq i}} \frac{a^{+}_{ij}}{a_{ii}}$$

$$\geq \sum_{\substack{j \\ j\neq i}} \frac{\bar{a}_{ij}}{2a_{ii}} - \sum_{\substack{j \\ j\neq i}} \frac{a^{+}_{ij}}{a_{ii}}$$

Now according to assumption (i) and to (10), Σa^{+}_{ij} will be smaller than $-\Sigma \bar{a}_{ij/2}$ (the sums being taken over all the columns of $\bar{A}$ but the i-th, for every row i, i=1,2,...,$\bar{n}$). Thus we have

$$\sum_{j=1}^{\bar{n}} c_{ij} \leq 1$$

that is to say, (5b) holds as well.

Now we note that C can be viewed as a matrix having non-zero off-diagonal elements at the same entries as those of a finite element matrix associated to a triangulation constructed in such a way that its graph structure is the same as the one under consideration but with all the angles opposite to edges of type $\overline{P_1P_2}$ being right angles(*). This allows us to conclude that (7d) is also satisfied.

Now all that is left to do is proving that (7c) holds. But taking (8) into account, it suffices to prove that the negative terms of D are overcome by the corresponding (positive) terms of CD^{+}.

If we examine the negative term d_{12}, for example, we easily see that $(CD^{+})_{12}$ is given by the sum of products of terms of ordered pairs of C and D, belonging to columns and rows associated to neighbors of both P_1 and P_2, i.e., P_3 and P_4 only.

More precisely, for d_{12} we must verify that :

$$(11) \qquad -\frac{a_{13}}{2a_{11}} \cdot \frac{-a_{32}}{2a_{33}} + \frac{-a_{14}}{2a_{11}} \cdot \frac{-a_{42}}{2a_{44}} - \frac{a_{12}}{a_{11}} \geq 0$$

whereas entirely analogous inequalities should hold for d_{21} or any other term alike.

(*) This means that we are eventually considering a different domain. However only the relative position of the nodes matters for the conclusion that follows.

Now we recall Figure 3 and note that a_{12} is formed by two positive contributions, namely, those from triangles T_1 and T_2, whose values are respectively $-(\text{cotg } \alpha_1)/2$ and $-(\text{cotg } \alpha_2)/2$.

On the other hand, according to the assumption (i) we have :

$$-a_{13} \geq \frac{1}{2} \text{ cotg } \gamma_1 \qquad -a_{32} \geq \frac{1}{2} \text{ cotg } \beta_1$$

$$-a_{14} \geq \frac{1}{2} \text{ cotg } \gamma_2 \qquad -a_{42} \geq \frac{1}{2} \text{ cotg } \beta_2$$

Therefore (11) will hold provided that the terms associated to each triangle T_1 and T_2 are positive, that is to say :

$$\frac{1}{16} \frac{\text{cotg } \gamma_r \text{ cotg } \beta_r}{a_{11} a_{r+2,r+2}} + \frac{\text{cotg } \alpha_r}{2a_{11}} \geq 0 \qquad r=1,2$$

Setting $\alpha_r = \pi/2 + \varepsilon_r$, we actually want to prove that

$$\frac{\text{cotg } \gamma_r \text{ cotg } \beta_r}{8a_{r+2,r+2}} - \text{tg } \varepsilon_r \geq 0 \qquad r=1,2$$

For simplicity we drop the indices r and we set $a_d = a_{r+2,r+2}$ (a_d carries the meaning of any diagonal term of $\bar{A}$). Now noting that

$$\beta+\gamma = \frac{\pi}{2} - \varepsilon$$

it suffices to have :

$$\text{cotg } \varepsilon \min_{x+y=\frac{\pi}{2}-\varepsilon} \text{cotg } x \text{ cotg } y \geq 8a_d$$

Since the minimum above is attained for $x=y=\pi/4 - \varepsilon/2$, this means that we want ε to satisfy :

$$(12) \qquad \frac{(1+\text{tg } \varepsilon/2)^3}{\text{tg } \varepsilon/2(1-\text{tg } \varepsilon/2)} \geq 16a_d$$

Now we draw the first general conclusion making the following assumption :

(ii) The triangulation $\mathcal{T}_h$ under consideration belongs to a regular family of triangulations, in the sense that there exists an angle θ such that, denoting by θ_i^T the angles of triangle T we have :

$$\theta_i^T \geq \theta > 0 \qquad i=1,2,3 \qquad \forall T \in \mathcal{T}_h \ .$$

Here the parameter h represents the mesh size as usual. In the case of regular family of partitions it can be defined to be

$$h = \max_{\substack{T \in \mathcal{T}_h \\ i \in \{1,2,3\}}} h_i^T$$

where h_i^T denotes the height of triangle T associated with angle θ_i^T. Indeed we have :

$$a_d = \frac{1}{2} \sum_{s=1}^{\mu} (\operatorname{cotg} \beta_s + \operatorname{cotg} \gamma_s)$$

where the (β_s, γ_s)'s are pairs of angles belonging to a triangle of the set of μ triangles having P_{r+2} as a vertex, but whose own vertices are nodes other than P_{r+2}, r=1,2. This gives :

$$a_d \leq M \operatorname{cotg} \theta$$

where M is the maximal number of triangles around a vertex of $\mathcal{T}_h$.

It is therefore obvious that, if ε is chosen to be sufficiently small, inequality (12) will hold together with assumption (10), namely if ε is such that :

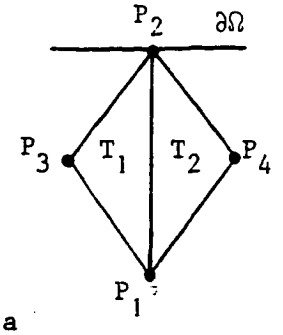
$$(13) \qquad \frac{(1+\operatorname{tg} \varepsilon/2)^3}{\operatorname{tg} \varepsilon/2 (1-\operatorname{tg} \varepsilon/2)} \geq 16 \, M \operatorname{cotg} \theta .^{(*)}$$

What remains to be done is to consider the case where at least one of the vertices P_2, P_3 and P_4 lies on $\partial\Omega$. Here we should distinguish the following situations :

---

* Strictly speaking, we checked that (13) implies (10).

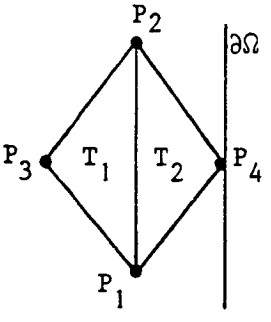
1st case : $P_2 \in \partial\Omega$, $P_3 \notin \partial\Omega$ and $P_4 \notin \partial\Omega$.



5a

This case can be treated as before. Indeed we only need to satisfy an inequality of type (11) for a_{12} (and not for a_{21}) which does not involve any term of the form a_{2j}, $j \neq 2$, that now vanish.

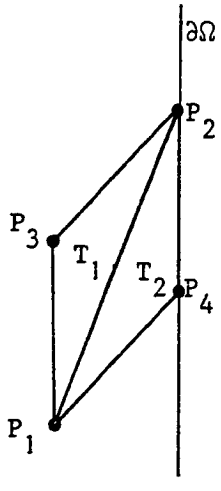
2nd case : $P_2 \notin \partial\Omega$ but $P_4 \in \partial\Omega$



5b

In this case the condition of positiveness of B^{-1} will be more stringent. Indeed, in inequality (11) we can only count on positive contributions from the first term, since now $a_{4j} = 0 \quad \forall\, j \neq 4$. This means that we should overcome contributions to a_{12} (or a_{21}) from both T_1 and T_2 using only terms computed over T_1. Therefore this is a kind of situation to be avoided in practice.

3rd case : $P_2 \in \partial\Omega$ and $P_4 \in \partial\Omega$ but $P_3 \notin \partial\Omega$



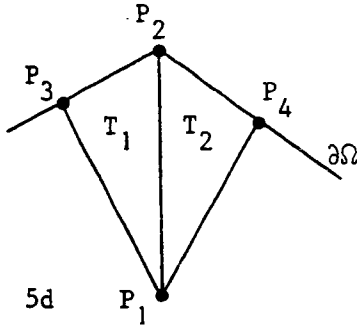
5c

The same difficulties of the previous case will arise here. However, no additional restrictions on the maximal value of ε will be introduced if the function g is constant. Indeed, in this case the positive off-diagonal term a_{12} belonging to A^{∂} is ficticious for it can be replaced by $a'_{12} = 0$, because

$$a_{14}g + a_{12}g = a'_{14}g + a'_{12}g$$

with $a'_{14} = a_{14} + a_{12} \leq 0$, if ε is small enough to satisfy (10). (As we noted before if ε satisfies (13), ε satisfies (10) as well).

4th case : P_2, P_3, $P_4 \in \partial\Omega$



5d

Figure 5

Boundary nodes

This case should be obviously excluded for there is no way to overcome a_{12} in (11), because now $a_{32}=a_{42}=0$. However if g is constant the same remark of the previous case applies, provided ε is small enough to satisfy (10). Indeed in this case we can replace a_{1k} by $a'_{1k} = a_{1k}+a_{12}/2 \geq 0$, k=3,4 and a_{12} by $a'_{12} = 0$.

As it should be pointed out, the bounds for the non-acute angles obtained with the use of (13) are rather severe with respect to what is to be expected in practice. Actually, in particular cases these can be significantly refined.

Let us give a concrete example : take $\theta = 35°$ and M=6 ; using (13) we get a maximal angular increment ε of about 1° ! Now we consider the particular case where the domain is a diamond of unit edge whose non acute angles measure $\pi/2+\varepsilon$. If we triangulate the domain as indicated in Figure 6, i.e. taking a uniform mesh consisting of equal isosceles triangles, we will have a triangulation that fulfills assumption (i). Now if we assume homogeneous

N = number of subdivisions of each edge. $\quad h = \cos \varepsilon/N$

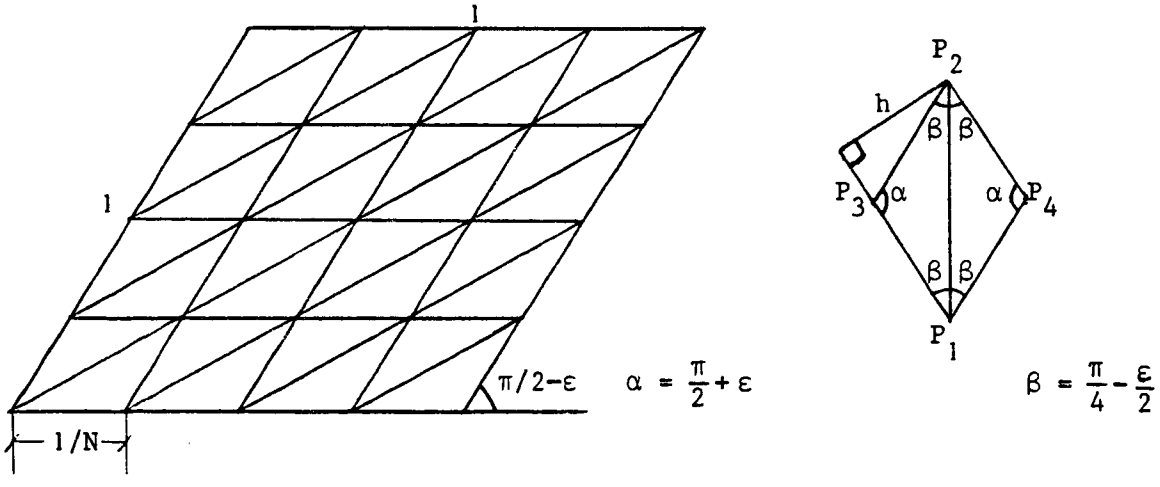


Figure 6

A triangulation of non-acute type for a unit diamond.

boundary conditions for instance, we can disregard boundary phenomena and determine ε so as to satisfy (11) where

$$a_{13} = a_{14} = a_{23} = a_{24} = - \operatorname{cotg}(\pi/4 - \varepsilon/2)$$

$$a_{11} = a_{33} = a_{44} = 4 \operatorname{cotg}(\pi/4-\varepsilon/2) - 2\operatorname{tg} \varepsilon$$

$$a_{12} = \operatorname{tg} \varepsilon$$

Computations lead to the bound $\varepsilon_{Max} \tilde{=} 8°$. Notice that here also we have M=6 and that this maximal value corresponds to $\theta = 41°$, i.e. a value just a little bit greater than the one we considered in the more general example.

Now it would be interesting to check how good the bounds for the increment ε derived in this work actually are, compared to maximal values obtained by direct computation of the inverse matrix. We do this for the same model problem on a diamond considered above that is partitionned like in Figure 6.

Taking homogeneous boundary conditions again, we obtained for N=10 positive inverses up to $\varepsilon = 45°$. Those computations due to E. Fernandez Cara [4] confirm that there is indeed a limit for the non-acute angles beyond which the inverse matrix is no longer positive. Notice however that for this test-problem, the bounds that we obtained by our method are still far from optimal.

4. GENERALIZATIONS TO OTHER RELATED PROBLEMS.

Let us now briefly discuss how the technique developped in the previous section applies to other related problems whose discretization leads to problems of non-positive type.

Let us first consider problem $(\mathcal{P}_a)$. It will be convenient to introduce the following notation :

$$m_{ij} = \int_\Omega a\phi_i\phi_j$$

$$s_{ij} = \int_\Omega \nabla\phi_i\nabla\phi_j$$

so that $a_{ij} = m_{ij}+s_{ij}$.

For simplicity we assume that the triangulation is so that $\alpha_1+\alpha_2 \leq \pi$ for every pair of angles α_1 and α_2 located as in Figure 1.

In this case non-negative off-diagonal terms may arise when

$$-\frac{\operatorname{cotg}\alpha_1 + \operatorname{cotg}\alpha_2}{2} + \frac{\|a\|_\infty}{12}\,\text{area}\,(T_1 \cup T_2) \geq 0$$

where $\|\cdot\|_\infty$ denotes the $L^\infty(\Omega)$-norm.

More specifically we take the least favorable case where $\alpha_1+\alpha_2 = \pi$ in which we will have a positive off-diagonal coefficient of the form

$$a_{12} = m_{12} = \int_{T_1 \cup T_2} a\phi_1\phi_2 \leq \frac{\|a\|_\infty}{12}\,[\,\text{area}(T_1)+\text{area}(T_2)\,] \tag{14}$$

Without specifying proper bounds at the moment, we assume that the measure of the elements are small enough, so that the coefficients of form a_{ik} (or a_{ki}) are negative for i=1,2 and k=3,4. We also maintain assumption (i). Note that, whereas the measure of the elements are irrelevant for the magnitude of s_{ij}, m_{ij} is directly depending on it (e.g. see (14) above).

Let us now apply Theorem 1 for a splitting of $B = \bar{A}_d^{-1}\bar{A}$ of the form

$$B = I-C-D$$

constructed in such a way that all the coefficients of the form a_{12}/a_{ii} (resp. a_{21}/a_{ii}) , i=1,2, a_{12} being given by (14), are assigned to D.

As for any negative off-diagonal element b_{ij} of B, as before we assign $b_{ij}/2$ to each matrix C and D.

Let us now check the conditions under which the assumptions of Theorem 1 are fulfilled.

That $\sum_{j=1}^{\bar{n}} b_{ij} > 0$ for $i=\overline{1,n}$ is obviously the case of problem $(\mathcal{P}_a)$, which means that there is no row of B that does not belong to J(B), as pointed out before.

According to our assumptions, $|s_{ij}| > |m_{ij}|$ whenever $s_{ij} < 0$, so that we have $c_{ij} \geq 0$, and hence 5c) is satisfied by C.

On the other hand we have for i=1,2,...,n :

$$1 - \sum_{j=1}^{\bar{n}} c_{ij} = 1 + \sum_{\substack{j=1 \\ j\neq i}}^{\bar{n}} \frac{(s_{ij}+m_{ij})^-}{2(s_{ii}+m_{ii})} \geq 1 + \sum_{\substack{j=1 \\ j\neq i}}^{\bar{n}} \frac{s_{ij}}{2(s_{ii}+m_{ii})} \geq$$

$$1 - \frac{s_{ii}}{2(s_{ii}+m_{ii})} > 0$$

and

$$1 - \sum_{j=1}^{\bar{n}} c_{ij} = 1 > 0 \quad \text{for } i=n+1,\ldots,\bar{n}, \text{ so that C satisfies 5b).}$$

7d) is satisfied for the same reason as in the case of problem $(\mathcal{P}_o)$.

Again, all that is left to do is proving that 7c) holds, for which it suffices to prove that $D^- + CD^+ \geq 0$.

Using the notation of the previous analysis and assuming that none of the P_i's i=1,2,3,4 lie on $\partial\Omega$, for $d_{12} < 0$ (resp. $d_{21} < 0$) we must verify that :

$$\frac{s_{13}+m_{13}}{2(s_{11}+m_{11})} \cdot \frac{s_{32}+m_{32}}{2(s_{33}+m_{33})} + \frac{s_{14}+m_{14}}{2(s_{11}+m_{11})} \cdot \frac{s_{42}+m_{42}}{2(s_{44}+m_{44})} - \frac{m_{12}}{s_{11}+m_{11}} \geq 0$$

The contribution to m_{12} from triangle T_r is bounded above by $\|a\|_\infty$ area $(T_r)/12$. Therefore it suffices to have :

$$(15) \qquad \frac{\operatorname{cotg} \beta_r \operatorname{cotg} \gamma_r}{16} - \frac{\|a\|_\infty S}{24} (\operatorname{cotg} \beta_r + \operatorname{cotg} \gamma_r) - \frac{\|a\|_\infty S}{12} a_{r+2,r+2} \geq 0$$

where

$$(16) \qquad S = \max_{T \in \mathcal{C}_h} \text{area } (T)$$

Denoting by s_d and m_d generic terms of form s_{ii} and m_{ii}, respectively, we have :

$$s_d \leq M \operatorname{cotg} \theta$$

$$m_d \leq MS/6$$

Taking into account the most unfavorable cases where $\alpha_r = \theta$ for the first term and $\beta_r = \gamma_r = \theta$ for the second term in inequality (15), we can say that, for homogeneous boundary conditions, the discrete maximum principle will hold in the case under study if :

$$(17) \qquad S \leq \frac{\operatorname{tg} \theta}{(M+1)c(\operatorname{tg}^2 \theta + \operatorname{cotg}^2 \theta)} \quad \text{where } c = \operatorname{Max}\left\{ \frac{4\|a\|_\infty}{3}, \frac{1}{6} \right\}$$

Remark : Actually (17) also implies that $|s_{ij}| > m_{ij}$ whenever $s_{ij} < 0$.

As a concrete example we take $g \equiv 0$ for a problem with $a \equiv 1$ defined on a unit square triangulated like in Figure 6 ($\varepsilon=0$). We have $\theta = 45°$, M=6 and thus from (17) we conclude that the maximum principle already holds for meshes as coarse as the one obtained with h=1/3, i.e., with N=3!

It is interesting to note that the bound given by (17) for the mesh size applies in particular to triangulations of acute type.

Now with suitable modifications for the case where $P_2 \in \partial\Omega$, analogous to the case of $(\mathcal{P}_o)$, (17) can be simply stated as follows :

If the mesh size h is sufficiently small the strong discrete maximum principle will hold for problem $(\mathcal{P}_a)$ provided the triangulation $\mathcal{C}_h$, besides belonging to a regular family of partitions, satisfies assumption (i) for all angles α_1 and α_2 of Figure 1, such that $\alpha_1+\alpha_2 \leq \pi$ in connection with every positive off-diagonal coefficient a_{12}.

Of course with suitable modifications we could as well include the case where $\alpha_1+\alpha_2 \leq 2\varepsilon$ for a sufficiently small ε, as we did for problem $(\mathcal{P}_o)$.

With similar arguments one could draw the same conclusion for the case where positive off-diagonal contributions to the discretization matrix of a second order problem depends on a positive power of the mesh size h, provided that the triangulation is of the type considered for problem $(\mathcal{P}_a)$ (i.e. with $\alpha_1+\alpha_2 \leq \pi$ together with assumption (i)).

A typical case would be the following problem :

$$\begin{cases} -\Delta u + \sum_{i=1}^{2} b_i \frac{\partial u_i}{\partial x_i} + au = f \text{ in } \Omega \qquad a, b_1, b_2 \in L^\infty(\Omega) \\ u = g \text{ on } \partial\Omega \end{cases}$$

We would like to conclude with a remark on the strong discrete maximum principle for approximations of the heat equation below :

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } \Omega \times]0,T[\\ u = u_o \qquad \text{on } \Omega \times \{0\} \\ u = g \qquad \text{on } \partial\Omega \times]0,T[. \end{cases}$$

Fujii [5] proved the validity of a maximum principle for a space discretization with piecewise linear finite elements constructed upon a triangulation of strictly acute type and time discretization with standard finite difference schemes.

Although for the sake of conciseness we preferred not to derive here proper bounds for this problem, we can say that without much difficulty one can prove the validity of Fujii's results under the weaker assumptions on the triangulation made here, together with similar boundedness assumptions on the ratios $\Delta t/(h_i^T)^2$, where Δt is the time-mesh size and the h_i^T's represent the heights of a triangle T, i=1,2,3.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor R. GLOWINSKI for his continuous support and encouragement and also to Professor H.B. KELLER of the California Institute of Technology for his valuable advices.

REFERENCES

[1] BRAMBLE J.H. and HUBBARD B.E., New monotone type approximations of elliptic problems, Mathematics of Computation, 18, 349-367 (1964).

[2] CIARLET Ph. G., Discrete maximum principle for finite difference operators, Aequationes Mathematica, 4, 338-352 (1970).

[3] CIARLET Ph. G. and RAVIART P.A., Maximum principle and uniform convergence for the finite element method, Computer Methods in Applied Mechanics and Engineering, 2, 17-31 (1973).

[4] FERNANDEZ CARA E., Personal communication, INRIA, France (1980).

[5] FUJII H., Some remarks on finite element analysis of time-dependent field problems, in Theory and Practice in Finite Element Structural Analysis, University of Tokyo Press, Tokyo, 1973, 91-106.

[6] STAMPACCHIA G., Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus, Ann. Inst. Fourrier (Grenoble), 15, 189-258 (1968).

[7] STRANG G. and FIX G.J., An analysis of the finite element method, Prentice Hall, Englewood Cliffs, N.J. (1973).