COMPUTATIONAL & APPLIED MATHEMATICS

Volume 31, N. 3, pp. 505–522, 2012
Copyright © 2012 SBMAC
ISSN 0101-8205 / ISSN 1807-0302 (Online)
www.scielo.br/cam

A weighted mass explicit scheme for convection-diffusion equations

VITORIANO RUAS*

Research associate, Departamento de Engenharia Mecânica,
PUC-Rio – Catholic University of Rio de Janeiro, Brazil.
Research collaborator, Institut Jean le Rond d'Alembert,
UPMC – Université Pierre et Marie Curie, Paris 6, France.
Visiting professor, Instituto de Ciências Matemáticas e Computação,
USP – University of São Paulo, São Carlos, Brazil.

E-mails: vitoriano.ruas@pq.cnpq.br / vitoriano.ruas@upmc.fr / vitoriano@icmc.usp.br

Abstract. An explicit scheme based on a weighted mass matrix, for solving time-dependent convection-diffusion problems was recently proposed by the author and collaborators. Convenient bounds for the time step, in terms of both the method's weights and the mesh step size, ensure its stability in space and time, for piecewise linear finite element discretisations in any space dimension. In this work we study some techniques for choosing the weights that guarantee the convergence of the scheme with optimal order in the space-time maximum norm, as both discretisation parameters tend to zero.

Mathematical subject classification: Primary: 65M60; Secondary: 76Rxx.

Key words: Convection-diffusion, convergence, explicit scheme, finite elements, piecewise linear, time-dependent, weighted mass.

1 Introduction

This work deals with an explicit scheme introduced in [6], for the numerical time integration of the convection-diffusion equations, discretised in space by techniques based on variational formulations such as the finite element method.

#CAM-698/12. Received: 22/IX/12. Accepted: 28/IX/12.
*Supported by CNPq grant 307996/2008-5.

In this framework, since the mid-eighties, the most widespread manner to deal with dominant convection has been the use of stabilizing procedures based on the space mesh parameter, among which the streamline upwind Petrov-Galerkin (SUPG) technique introduced by Hughes & Brooks (cf. [2]) is one of the most popular.

The author and collaborators studied in [7] a contribution in this direction, based on a standard Galerkin approach, and a space discretisation of the convection-diffusion equations with piecewise linear finite elements, combined with a non standard explicit forward Euler scheme for the time integration. The main theoretical result in that work, states that the numerical solution is stable in the maximum norm in both space and time (and even convergent with order $h|lnh|$ if the mesh is of the acute type [10]), provided that roughly the time step is bounded by the space mesh parameter h multiplied by a mesh-independent constant, for a high Péclet number. As it should be clarified, the scheme under consideration follows similar principles to the one long exploited by Kawahara and collaborators, for simulating convection dominated phenomena (see e.g. [4], among several other papers published by them before and later on). The originality of our contribution relies on the fact that it not only introduces a reliable scheme for any space dimension, but also exhibits rigorous conditions for it to provide converging sequences of approximations in the sense of the space-time maximum norm.

The main purpose of this work, is to specify procedures for the determination of sets of weights that characterize our explicit scheme. As for the theoretical contribution, we prove that for such choices of the weights the method converges with optimal order in the maximum space-time norm.

An outine of the paper is as follows: In Section 2 we recall the problem to solve together with the type of discretisation corresponding to our explicit scheme; more particularly the weighted manner to deal with the mass matrices on both sides of the discrete equations is described. In Section 3 we further recall the stability results that hold for the method being considered, in the sense of the space and time maximum norm, together with the conditions to be fulfilled by the sequences of meshes and time steps in order to ensure convergence. We proceed in Section 4 by studying in detail some particular choices of the weights associated with the scheme, that satisfy the conditions leading to both stability and convergence. Finally in Section 5 we consider some implementation aspects of the method and give corresponding numerical results.

2 The problem to solve and its discretisation

Let us consider a time-dependent convection-diffusion problem described as follows:

Find a scalar valued function $u(\mathbf{x}, t)$ defined in $\bar{\Omega} \times [0, \infty)$, Ω being a bounded open subset of $\Re^N$ with boundary $\partial\Omega$, $N = 1, 2$ or 3, such that,

$$\begin{cases} u_t + \mathbf{a} \cdot \nabla u - \nu \Delta u = f \ \text{ in } \Omega \times (0, \infty) \\ u = g \ ; \ \text{ on } \ \partial\Omega \times (0, \infty) \\ u = u^0 \ ; \ \text{ in } \ \Omega \ \text{ for } \ \mathrm{t = 0} \end{cases} \tag{1}$$

where u_t represents the first order derivative of u with respect to t, ν is a positive constant and $\mathbf{a}$ is a given solenoidal convective velocity at every time t, assumed to be uniformly bounded in $\Omega \times (0, \infty)$. The data f and g are respectively, a given forcing function belonging to $L^\infty[\Omega \times (0, \infty)]$, and a prescribed value in $L^\infty[\partial\Omega \times (0, \infty)]$. We further assume that $u^0 \in L^\infty(\Omega)$ and that for every $\mathbf{x} \in \Omega$ $g(\mathbf{x}, \cdot)$ is of bounded variation in $(0, \infty)$. In (1) ν represents the inverse of the Péclet number.

Without loss of essential aspects, in all the sequel we assume that Ω is an interval if $N = 1$, a polygon if $N = 2$ or a polyhedron if $N = 3$. In so doing we consider a partition $\mathcal{T}_h$ of Ω into $N-$ simplices, with maximum edge length equal to h. We assume that $\mathcal{T}_h$ satisfies the usual compatibility conditions for finite element meshes, and that it belongs to a quasi-uniform family of partitions. We further define a second mesh parameter $h_{\min}$ as the minimum height of all the elements of $\mathcal{T}_h$ if $N = 2$ or 3 and the minimum length of $K \in \mathcal{T}_h$ if $N = 1$.

Let N_h be the number of nodes of $\mathcal{T}_h$, denoted by P_j, $j = 1, 2, \ldots, N_h$. We assume that these are numbered in such a manner that the first I_h nodes are located in the interior of Ω and the remaining $N_h - I_h$ nodes are located on $\partial\Omega$. Now for every $K \in \mathcal{T}_h$ we denote by $P_1(K)$ the space of polynomials of degree less than or equal to one defined in K. In so doing we introduce the following spaces or manifolds associated with $\mathcal{T}_h$:

$$V_h := \left\{ v \mid \ v \in C^0(\bar{\Omega}) \ \text{and} \ v|_K \in P_1(K), \ \forall K \in \mathcal{T}_h \right\},$$

$$V_h^0 := V_h \cap H_0^1(\Omega),$$

We further introduce for any function ϕ defined in $C^0(\partial\Omega)$ the following manifold of V_h:

$$V_h^{\phi} := \left\{ v \in V_h \mid v(P_j) = \phi(P_j) \ \forall \ \text{vertex} \ P_j \ \text{of} \ \mathcal{T}_h \ \text{on} \ \partial\Omega \right\},$$

Now let u_h^0 be the field of $V_h^{g(\cdot,0)}$ satisfying $u_h^0(P_j) = u^0(P_j)$ for every vertex P_j of $\mathcal{T}_h$, and $\Delta t > 0$ be a given time step. Defining g^n on $\partial\Omega$ by $g^n(\cdot) = g(\cdot, n\Delta t)$, f^n in Ω by $f^n(\cdot) = f(\cdot, n\Delta t)$ and $\mathbf{a}^n$ in Ω by $\mathbf{a}^n(\cdot) = \mathbf{a}(\cdot, n\Delta t)$, for $n = 1, 2, \ldots$, idealistically we wish to determine approximations $u_h^n(\cdot)$ of $u(\cdot, n\Delta t)$ for $n \in \mathbb{N}^*$, by solving the following finite element discrete problem described below, corresponding to a modification of the first order forward Euler scheme.

For n successively equal to $1, 2, \ldots$, we wish to determine $u_h^n \in V_h^{g^n}$ of the form

$$u_h^n = \sum_{j=1}^{N_h} u_j^n \varphi_j,$$

where φ_j is the canonical basis function of V_h associated with the j−th node of $\mathcal{T}_h$, i.e. P_j, $u_j^n \in \Re$ being the value of u_h^n at P_j. We denote by S_j the support of φ_j and by Π_j its measure.

The unknowns u_i^n for $n = 1, 2, \ldots$, are recursively determined by the following expressions:

$$m_{ii}^L u_i^n = \sum_{j=1}^{N_h} \left[m_{ij}^W - \Delta t a_{ij}^{n-1} \right] u_j^{n-1} + \Delta t b_i^n, \ \text{for} \ i = 1, \ldots, I_h, \tag{2}$$

where, setting $\mathbf{a}_i^n := \mathbf{a}^n(P_i)$ and $f_i^n := f^n(P_i)$, the coefficients a_{ij}^n and b_i^n are given by:

$$a_{ij}^n = \int_{\Omega} \left[(\mathbf{a}_i^n \cdot \nabla)\varphi_j \varphi_i + \nu \nabla\varphi_j \cdot \nabla\varphi_i \right], \tag{3}$$

$$b_i^n = \int_{\Omega} f_i^{n-1} \varphi_i.$$

Coefficient m_{ii}^L is the well-known lumped mass diagonal matrix (cf. [9]) given by $\frac{\Pi_i}{N+1}$. The mass matrix coefficients m_{ij}^W on the right hand side of (2) in turn are defined by a weighted quadrature formula described as follows.

Let M_i be the number of nodes different from P_i lying in the closure of S_i, i.e. $\bar{S}_i$, and P_{k_j} be such nodes for $j = 1, 2, ..., M_i$ with $1 \leq k_j \leq N_h$. Let also W_j^i be the measure fractions associated with P_{k_j} given by:

$$W_j^i = \frac{meas(S_i \cap S_{k_j})}{N+1} \tag{4}$$

and ω_j^i be corresponding strictly positive weights satisfying,

$$\sum_{j=1}^{M_i} \omega_j^i W_j^i = \frac{N\Pi_i}{(N+1)(N+2)} \tag{5}$$

Notice that since each N-simplex in S_i appears in exactly N measure fractions W_j^i, we necessarily have:

$$\sum_{j=1}^{M_i} W_j^i = \frac{N\Pi_i}{N+1} \tag{6}$$

Now selecting the nodes P_{k_j} in $\bar{S}_i$ different from P_i, we define,

$$m_{ik_j}^W = \frac{h_{\min}}{\nu + h_{\min}} \omega_j^i W_j^i \quad \text{for } i \neq k_j \tag{7}$$

together with

$$m_{ii}^W = \frac{\nu}{h_{\min} + \nu} m_{ii}^L + \frac{h_{\min}}{h_{\min} + \nu} m_{ii}^C. \tag{8}$$

where m_{ii}^C is the $i - th$ diagonal coefficient of the standard consistent mass matrix ([9]) given by

$$\frac{2\Pi_i}{(N+1)(N+2)}.$$

Naturally enough, by definition, $m_{ij}^W = 0$ if P_j does not lie in $\bar{S}_i$.

Typically we may choose $\omega_j^i = \frac{1}{N+2}$ for every j and for every node P_i, thereby generating a weighted combination of the lumped mass and the consistent mass matrix (cf. [9]) on the right hand side of (2), with weights equal to

$$\frac{\nu}{h_{\min} + \nu} \quad \text{and} \quad \frac{h_{\min}}{h_{\min} + \nu},$$

respectively. However, except for the case of uniform meshes, in principle this is not the choice to make, if one wishes to reach the best results in terms of accuracy.

3 Stability and convergence of the scheme

In this Section we recall the stability and convergence results proven in [7] that hold for the above defined weighted mass scheme. In short they state that, provided Δt is chosen conveniently small with respect to the spatial mesh parameter, the scheme (2) is stable in the sense of the maximum norm. Moreover under a suitable condition on the mesh it converges in both space and time in the same sense, as h and Δt go to zero.

First we have to define the following quantities:

- $A = \sup\limits_{t\in(0,\infty)} \max\limits_{1\leq i\leq I_h} |\mathbf{a}_i(t)|$;
- $\omega = \min\limits_{1\leq i\leq I_h} \min\limits_{1\leq j\leq M_i} \omega_j^i$.

The following theorem proved in [7] states the stability result that holds for scheme (2).

Theorem 1. *If Δt fulfills the condition*

$$\Delta t \leq \frac{\omega h_{\min}^3}{(\nu + h_{\min})[A h_{\min} + (N+1)\nu]}, \tag{9}$$

then the finite element solution sequence $\{u_h^n\}_n$ given by $u_h^n = \sum\limits_{j=1}^{N_h} u_j^n \varphi_j$ generated by (2) *for $n = 1, 2, \ldots$ satisfies the following stability result for every $m \in \mathbb{N}$, whereby $||F||_{0,\infty,D}$ denotes the L^∞−norm of a function F defined in an open set D of $\Re^N$, and $BV[G]$ represents the standard norm of a function $G(t)$ having bounded variation for $t \in (0,\infty)$:*

$$||u_h^m||_{0,\infty,\Omega} \leq ||u^0||_{0,\infty,\Omega} + \max\{\max_{P\in\partial\Omega} BV[g(P,\cdot)], \Delta t \sum_{n=1}^{m} ||f^{n-1}||_{\infty,\Omega}\} \tag{10}$$

The above stability result can be refined as follows, in the particular case where the partition $\mathcal{T}_h$ is of the *acute type* (see e.g. [10]).

Theorem 2. (cf. [7]). *Assume that the partition $\mathcal{T}_h$ is of the acute type (cf. [10]). Then if Δt satisfies the condition*

$$\Delta t \leq \frac{h_{\min}^2}{\nu + h_{\min}} min \left[\frac{\omega}{A}, \frac{\nu(N+2) + 2h_{\min}}{\nu(N+1)(N+2)} \right] \tag{11}$$

the finite element solution sequence $\{u_h^n\}_n$ given by $u_h^n = \sum_{j=1}^{N_h} u_j^n \varphi_j$ generated by (2) *for $n = 1, 2, \ldots$ satisfies the stability condtion* (10).

In [7] we give error estimates for the approximations of the solution of (1) generated by (2) under condition (9). In particular we recall here that, provided the weights ω_j^i are suitably chosen, this scheme provides convergent approximations in the maximum norm, as both h and Δt go to zero, under the assumption (11) of Theorem 2. For this purpose we use Sobolev spaces $W^{m,\infty}(D)$ equipped with the standand norm and seminorm denoted respectively by $|| \cdot ||_{m,\infty,D}$ and $| \cdot |_{m,\infty,D}$, where m is a non negative integer and D is a subset of $\Re^N$ (cf [1]).

As usual a suitable consistency result is needed, which together with the stability results given in this Section leads to convergence. Actually the consistency of our scheme is a consequence of the following lemma:

Lemma 1. (cf. [7]). *Let P_i be a node of $\mathcal{T}_h$, for $i \in \{1, 2, \ldots, I_h\}$, and $\mathbf{l}_j^i$ be the vector leading from P_i to its neighbor P_{k_j}, that is, the j-th node belonging to $\bar{S}_i$, $j = 1, 2, \ldots, M_i$. Then there exists strictly positive weights ω_j^i satisfying* (5) *such that*

$$\sum_{j=1}^{M_i} \omega_j^i W_j^i \mathbf{l}_j^i = \mathbf{0}. \tag{12}$$

Then we can establish the validity of the following convergence result for scheme (2):

Theorem 2. (cf. [7]). *Let the strictly positive weights ω_j^i, $\forall i \in \{1, 2, \ldots, I_h\}$ and $\forall j \in \{1, 2, \ldots, M_i\}$, satisfy* (12)-(5). *Assume that for a given finite time $T > 0$ both the solution u of* (1) *and u_t belong to $W^{2,\infty}(\Omega)$. Assume also that $\forall t \in [0, T]$, $\mathbf{a}(\cdot, t) \in [W^{1,\infty}(\Omega)]^N$ and $f(\cdot, t) \in W^{1,\infty}(\Omega)$, and $(u_t)_t$ belongs to $L^\infty(\Omega)$. Finally let a strictly positive integer k_T be defined as the minimum* *of all integers k such that the quantity $\Delta t := \dfrac{T}{k}$ fulfills the condition* (11). *Let us further assume that, besides belonging to a quasiuniform family, all the partitions $\mathcal{T}_h$ in use are of the acute type, and the quantity ω is bounded below away from zero independently of h. Then there exists a constant C independent of u, h and Δt, such that the following estimate applies:*

$$\begin{aligned}
&\max_{1\leq m\leq k_T} ||u^m - u_h^m||_{0,\infty,\Omega} \leq Ch|ln\ h| \max_{0\leq s\leq T}\{ \\
&||u(\cdot,s)||_{2,\infty,\Omega} + ||u_t(\cdot,s)||_{1,\infty,\Omega} \\
&+h||u_t(\cdot,s)||_{2,\infty,\Omega} + h||(u_t)_t(\cdot,s)||_{0,\infty,\Omega} \\
&+||\mathbf{a}(\cdot,s)||_{1,\infty,\Omega}(||u(\cdot,s)||_{1,\infty,\Omega} + h||u(\cdot,s)||_{2,\infty,\Omega}) \\
&+||\mathbf{a}(\cdot,s)||_{0,\infty,\Omega}||u(\cdot,s)||_{2,\infty,\Omega} + ||f(\cdot,s)||_{1,\infty,\Omega}\}
\end{aligned} \tag{13}$$

4 Consistent choices of the weights

In this section we describe two coherent strategies to determine a set of M_i strictly positive weights, for each mesh inner node P_i, that are proven to be bounded below away from zero, independently of the mesh parameter h.

Let us first consider the one-dimensional case. From (12) and (5) it is trivially seen that the pair of weights (ω_1^i, ω_2^i) associated with inner node P_i is uniquely defined by the equations:

$$\left\{\begin{array}{lcl}
-\omega_1^i(l_1^i)^2 + \omega_2^i(l_2^i)^2 & = & 0 \\
\frac{1}{2}[\omega_1^i l_1^i + \omega_2^i l_2^i] & = & \dfrac{l_1^i + l_2^i}{6},
\end{array}\right. \tag{14}$$

where l_1^i and l_2^i are the lengths of the intervals of $\mathcal{T}_h$ having P_i as the right and the left end, respectively. This yields $\omega_1^i = \dfrac{l_2^i}{3l_1^i}$ and $\omega_2^i = \dfrac{l_1^i}{3l_2^i}$.

Next we switch to the case $N > 1$. In principle for $N = 2$ or $N = 3$ there are infinitely many solutions, except for the case of the least possible value of M_i, i.e. $M_i = N + 1$, in which the solution is necessarily unique. The two constructions described below allow for the unique determination of a set of weights satisfying (12) and (5), and incidentally they apply even to the particular case where this set is unique.

4.1 A first set of weights

Let $E_i^+ := \{\mathbf{e}_j^i\}_{j=1}^{M_i}$ be the set of unit vectors corresponding to the vectors $\mathbf{l}_j^i$, that is, $\mathbf{e}_j^i = \mathbf{l}_j^i / l_j^i$, where l_j^i is the modulus of $\mathbf{l}_j^i$. Let also $E_i^- := \{-\mathbf{e}_j^i\}_{j=1}^{M_i}$, and $E_i = E_i^+ \cup E_i^-$. Setting $K_i = card(E_i)$, we clearly have $K_i \geq M_i$. We number the M_i vectors of $E_i \cap E_i^+$ in the same manner as the vectors of E_i^+.

Let us first consider the case where $E_i^- = E_i^+ = E_i$. Then for every $j_1 \in \{1, ..., M_i\}$ there is necessarily another $j_2 \in \{1, ..., M_i\}$ such that $\mathbf{e}_{j_1}^i + \mathbf{e}_{j_2}^i = \mathbf{0}$. This implies in particular that M_i must be an even number. Hence we may choose the weights in pairs, say $(\omega_{j_1}^i, \omega_{j_2}^i)$, in the same way as in the one-dimensional case (cf. (14)). More specifically, we number the vectors in E_i in such a manner that the first $M_i/2$ ones form a subset of E_i whose vectors do not have any vector opposite to it in this subset, and from $M_i/2 + 1$ up to M_i the vectors in the complementary subset consisting of corresponding opposite vectors. In so doing the weights satisfy:

$$\left\{ \begin{array}{lcl} -\omega_{j_1}^i l_{j_1}^i W_{j_1}^i + \omega_{j_2}^i l_{j_2}^i W_{j_2}^i & = & 0 \\ \omega_{j_1}^i W_{j_1}^i + \omega_{j_2}^i W_{j_2}^i & = & \dfrac{W_{j_1}^i + W_{j_2}^i}{N+2}, \end{array} \right. \tag{15}$$

for all pairs $(j_1, j_2) \in \{1, \ldots, M_i/2\} \times \{M_i/2 + 1, \ldots, M_i\}$, such that $\mathbf{e}_{j_1}^i + \mathbf{e}_{j_2}^i = \mathbf{0}$. Notice that the set of weights determined by solving (15) trivially satisfy both (12) and (5).

Next we assume that $E_i^- \neq E_i^+$ (or equivalently, $E_i \neq E_i^+$). In this case we have $K_i > M_i$, and we number the $K_i - M_i$ vectors in E_i that do not belong to E_i^+ from $M_i + 1$ up to K_i in an arbitrary order, say $\mathbf{e}_k^i$, $k = M_i + 1, \ldots, K_i$. By assumption the intersection of $\bar{S}_i$ with the half straight line with origin at P_i and oriented in the sense and direction of $\mathbf{e}_k^i$ for $k > M_i$, cannot be an edge $P_i P_{k_j}$ with $j \leq M_i$. Letting the segment $P_i Q_k^i$ be such intersection, where Q_k^i is necessarily a point of the boundary of S_i, it follows that it is contained in either a single N-simplex of S_i for $N = 2$ or $N = 3$, or in a common face of exactly two neighboring tetrahedra for $N = 3$. In any case, the vector leading from P_i to Q_k^i, for $k = M_i + 1, \ldots, K_i$, still denoted by $\mathbf{l}_k^i$, is a non trivial convex combination of either exactly N edge vectors $\mathbf{l}_j^i$ with $1 \leq j \leq M_i$ pertaining to the same N-simplex of S_i, or of two such edge vectors pertaining to a common face of two neighboring tetrahedra of S_i. Let J be the number of such edge vectors, i.e., either $J = N$ for $N = 2$ or $N = 3$, or $J = 2$ for $N = 3$ only, and $P_{k_{m_l}}$, for $l = 1, \ldots, J$ with $1 \leq m_l \leq M_i$, be the vertices of S_i whose convex combination yields $\mathbf{l}^i_k$ for $k = M_i + 1, \ldots, K_i$. Let us denote the modulus of $\mathbf{l}^i_k$ by l^i_k, for $k = M_i + 1, \ldots, K_i$ too. In so doing we have:

$$\mathbf{l}^i_k = \sum_{l=1}^{J} \alpha^k_l \mathbf{l}^i_{m_l}, \quad k = M_i + 1, \ldots, K_i, \tag{16}$$

where the α^k_l's for $l = 1, \ldots, J$, are coefficients of a non trivial convex combination, that is, $0 < \alpha^k_l < 1$, $l = 1, \ldots, J$ and $k = M_i + 1, \ldots, K_i$, with $\sum_{l=1}^{J} \alpha^k_l = 1$.

Now we momentarily assign to each node P_{k_j} of S_i different from P_i, and to each point Q^i_j, $j = M_i + 1, \ldots, K_i$, the same measure fraction, say

$$\tilde{W}^i_j := \frac{N\Pi_i}{(N+1)K_i},$$

and the weight

$$\tilde{\omega}^i_j := \frac{K_i}{N+2} \times \left[\sum_{k=1}^{K_i} \frac{l^i_j}{l^i_k} \right]^{-1}.$$

Thanks to the fact that $\sum\limits_{j=1}^{K_i} \mathbf{e}^i_j = \mathbf{0}$ by construction, we necessarily have:

$$\sum_{j=1}^{K_i} \tilde{\omega}^i_j \tilde{W}^i_j \mathbf{l}^i_j = \mathbf{0}, \tag{17}$$

together with

$$\sum_{j=1}^{K_i} \tilde{\omega}^i_j \tilde{W}^i_j = \frac{N\Pi_i}{(N+1)(N+2)}, \tag{18}$$

as one can easily check.

Now we replace in (17) the $\mathbf{l}^i_k$'s for $k = M_i + 1, \ldots, K_i$, by the expression given by (16). Then rearranging the terms in the resulting expression, we establish that relation (12) holds for weights ω^i_j defined in the following manner:

$$\omega^i_j = \frac{C_i \tilde{W}^i_j \left(\tilde{\omega}^i_j + \delta^i_j\right)}{W^i_j}, \quad \text{for } j = 1, \ldots, M_i. \tag{19}$$

In (19) C_i is a normalizing constant allowing (5) to hold. Notice that, provided the weight increments δ^i_j are all non-negative, the value of C_i is strictly positive. The values of δ^i_j in turn, simply account for the sum of the contributions of $\mathbf{l}^i_j$ for a given $j \leq M_i$, to the vectors $\mathbf{l}^i_k$ for $k > M_i$, expressed by (16), respectively multiplied by $\tilde{\omega}^i_k$, and the corresponding convex combination coefficient. More specifically we have,

$$\delta^i_j = \sum_{k=M_i+1}^{K_i} \beta^k_j \tilde{\omega}^i_k \tag{20}$$

where $\beta^k_j = 0$ if Q^i_k does not belong to $S_i \cap S_{k_j}$ and $\beta^k_j = \alpha^k_l$ for the pertaining convex combination coefficients in (16), that is, all the α^k_l's such that $m_l = j$. Notice that by construction (5) holds for the weights ω^i_j defined by (19)-(20).

An important result that holds in connection with the above construction is

Theorem 4.1. *The weights generated in the way prescribed in this sub-section are all bounded below by a strictly positive constant ω independent of the mesh step size, whatever the mesh one might consider in the quasi-uniform family of meshes in use.*

Proof. The case where $K_i = M_i$ is trivial and hence we give a detailed proof only for the case $K_i > M_i$.

First of all we note that

$$(N+2)\tilde{\omega}^j_i \geq \big[\min_{1\leq k\leq K_i} l^k_i\big](l^j_i)^{-1} \geq h_{\min} h^{-1}.$$

Next from (20) and (5) we have,

$$\frac{N\Pi_i}{(N+1)(N+2)} = \sum_{j=1}^{M_i} \tilde{W}^j_i \big(\tilde{\omega}^j_i + \delta^j_i\big) C_i. \tag{21}$$

Since

$$\delta^j_i \leq \sum_{k=M_i+1}^{K_i} \tilde{\omega}^k_i \leq \frac{h(K_i - M_i)}{(N+2)h_{\min}}$$

and from (18) it holds that $\displaystyle\sum_{j=1}^{M_i} \tilde{\omega}^j_i \tilde{W}^j_i \leq \frac{N\Pi_i}{(N+1)(N+2)}$, taking into account

(21) we obtain,

$$C_i \geq \frac{1}{1 + \dfrac{M_i(K_i - M_i)h}{k_i h_{\min}}}. \tag{22}$$

It follows from (22) that for a suitable mesh independent constant c_0 we have $C_i \geq c_0\ \forall i$. Indeed $K_i \leq 2M_i$ and for no mesh of the quasi-uniform family of meshes $\{\mathcal{T}_h\}_h$ under consideration, M_i exceeds the value c^{-N}, where c is a mesh independent constant such that $\rho \geq ch$ for every h, ρ being the minimum over the elements of $\mathcal{T}_h$ of the radii of the largest inscribed balls in the elements of $\mathcal{T}_h$ for a given h (cf. [3]).

Finally, since $\delta_i^j > 0$ for all i, j we have,

$$\omega_i^j \geq \frac{c_0 h_{\min} \tilde{W}_i^j}{(N+2)hW_i^j} \geq \frac{c_0}{N+2}\left(\frac{h_{\min}}{h}\right)^{N+1}. \tag{23}$$

It immediately follows from (23) that there exists a suitable mesh independent strictly positive constant c_N such that

$$\omega_i^j \geq c_N \ \ \forall i \in \{1, \ldots I_h\} \quad \text{and} \quad \forall j \in \{1, \ldots, M_i\}. \tag{24}$$

□

Remark 4.2. (24) clearly indicates that c_N may play the role of the parameter ω in the relations (9) and (11). However for very distorted meshes such a value of ω may be largely under-evaluated, and for this reason in practical computations it is advisable to determine this parameter simply as the minimum of all the ω_i^j's for a given mesh, keeping in mind that, according to (24) such a value is necessarily bounded below away from zero independently of the mesh size.

4.2 A second set of weights

Here we consider again the sets E_i^+ and E_i^-. The case where $E_i^- = E_i^+$ is to be treated in the same manner as above. Hence we may assume that $E_i^+ \neq E_i^-$. Distinguishing eventually coincident vectors in E_i^- and E_i^+, we put together all the vectors in both sets. Otherwise stated we are dealing with a set of exactly $2M_i$ vectors not necessarily distinct, say $E_i^{'} := \{\mathbf{e}_j^i\}_{j=1}^{2M_i}$, where the first M_i vectors are those of E_i^+ and the last M_i vectors are given by $\mathbf{e}_{j+M_i}^i := -\mathbf{e}_j^i$, for $j = 1, \cdots, M_i$. Now let the segment $P_i Q_k^i$ be the intersection of $\bar{S}_i$ with the half straight line with origin at P_i and oriented in the sense and direction of $\mathbf{e}_k^i$ for $k > M_i$, Q_k^i being necessarily a point of the boundary of S_i. It follows that $P_i Q_k^i$ either coincides with an edge $P_i P_{k_j}$, where P_{k_j} is a vertex of S_i, or is contained in either a single N-simplex of S_i for $N = 2$ or $N = 3$, or in a common face of exactly two neighboring tetrahedra for $N = 3$. In any case, the vector leading from P_i to Q_k^i, for $k = M_i + 1, \cdots, 2M_i$, still denoted by $\mathbf{l}_k^i$, is a convex combination of at most N edge vectors $\mathbf{l}_j^i$ with $1 \leq j \leq M_i$ pertaining to the same N-simplex of S_i. Let J be the number of such edge vectors, i.e., $J = 1$ if Q_k^i coincides with a given vertex of S_i, $J = 2$ for $N = 3$ only if Q_k^i is contained in a common face of two neighboring tetrahedra of S_i, or $J = N$ for $N = 2$ or $N = 3$ otherwise. Let $P_{k_{m_l}}$, for $l = 1, \cdots, J$ with $1 \leq m_l \leq M_i$, be the vertices of S_i whose convex combination yields $\mathbf{l}_k^i$ for $k = M_i + 1, \cdots, 2M_i$. Here again we denote the modulus of $\mathbf{l}_k^i$ by l_k^i, for $k = M_i + 1, \cdots, 2M_i$ too. In so doing we have:

$$\mathbf{l}_k^i = \sum_{l=1}^{J} \alpha_l^k \mathbf{l}_{m_l}^i, \quad k = M_i + 1, \cdots, 2M_i, \tag{25}$$

where the α_l^k's for $l = 1, \cdots, J$, are coefficients of a convex combination, that is, $0 \leq \alpha_l^k \leq 1, l = 1, \cdots, J$ and $k = M_i + 1, \cdots, 2M_i$, with $\sum_{l=1}^{J} \alpha_l^k = 1$.

Now we assign to each point Q_k^i, $k = M_i + 1, \cdots, 2M_i$ the measure fraction $W_k^i = W_{k-M_i}^i$, and the weight $\tilde{\omega}_k^i = \dfrac{l_{k-M_i}^i}{l_k^i} \tilde{\omega}_{k-M_i}^i$, where $\tilde{\omega}_j^i$ are provisional weights respectively associated with the vertices P_{k_j} of S_i different from P_i satisfying $\tilde{\omega}_k^i + \tilde{\omega}_{k-M_i}^i = \dfrac{1}{N+2}$. Then similarly to the case of (15) we necessarily have:

$$\sum_{j=1}^{2M_i} \tilde{\omega}_j^i W_j^i \mathbf{l}_j^i = \mathbf{0}, \tag{26}$$

together with

$$\sum_{j=1}^{2M_i} \tilde{\omega}_j^i W_j^i = \frac{N \Pi_i}{(N+1)(N+2)}, \tag{27}$$

as one can easily check.

Now we replace in (26) the $\mathbf{l}_k^i$'s for $k = M_i + 1, \cdots, 2M_i$, by the expression given by (25). Then rearranging the terms in the resulting expression, we establish that relation (12) holds for weights ω_j^i defined in the following manner:

$$\omega_j^i = C_i(\tilde{\omega}_j^i + \delta_j^i), \text{ for } j = 1, \ldots, M_i. \tag{28}$$

In (28) C_i is a normalizing constant allowing (5) to hold. Notice that, provided the weight increments δ_j^i are all non-negative, the value of C_i is strictly positive. The values of δ_j^i in turn, simply account for the sum of the contributions of $\mathbf{l}_j^i$ for a given $j \leq M_i$, to the vectors $\mathbf{l}_k^i$ for $k > M_i$, expressed by (25), respectively multiplied by $\tilde{\omega}_k^i$, and the corresponding convex combination coefficient. More specifically we have,

$$\delta_j^i = \sum_{k=M_i+1}^{2M_i} \beta_j^k \tilde{\omega}_k^i \tag{29}$$

where $\beta_j^k = 0$ if Q_k^i does not belong to $S_i \cap S_{k_j}$ and $\beta_j^k = \alpha_l^k$ for the pertaining convex combination coefficients in (25), that is, all the α_l^k's such that $m_l = j$. Notice that by construction (5) holds for the weights ω_j^i defined by (28)-(29).

By arguments in all similar to those in the proof of Theorem 4.1 we can prove,

Theorem 4.3. *The weights generated in the way prescribed in this sub-section are all bounded below by a strictly positive constant ω independent of the mesh step size, whatever the mesh one might consider in the quasi-uniform family of meshes in use.*

Remark 4.4. The pairs of expressions (19)-(20) and (28)-(29) defining two a priori different sets of weights ω_i^j attempt to take into account three main factors playing a role in the influence of the vector $\mathbf{l}_i^j$ in equation (12). First of all its modulus, since the smaller it is, the larger must be the corresponding weight. Next the associated measure fraction W_i^j, which goes in the same sense as l_i^j. Finally the increments δ_i^j, which introduce the necessary adjustment in the weight values, in order to take into account the eventual existence of vectors in L_i roughly opposite to $\mathbf{l}_i^j$, that is, making angles with it close to π. In this sense we may assert that the above described procedure for determining the ω_i^j's, can be viewed as both close to optimal choices of the weights of our method, and straightforward manners to compute them.

5 Implementation and Numerical Aspects

It is not difficult to figure out from the construction of the sets of weights described in the last Section, that both choices give rise to a straightforward implementation of the method. Nevertheless we would like to point out that it is wiser to assemble node by node both the weighted mass matrix and the matrix generated by the discretisation of the convection-diffusion operator, instead of the usual element-by-element procedure. All that is needed for this purpose is a table of node numbers and an associated integer pointer vector having as many components as there are nodes. In the table the series of numbers of the neighbors of each inner node are successively stored, while in the pointer vector the $i - th$ component contains the position in the table corresponding to the first neighbor of the $i - th$ inner node, thereby allowing to uniquely identify all the series of neighbors in the table.

Next we check the performance of the method studied in this paper as compared to a well-established technique to deal with dominant convection in convection-diffusion problems, namely, the least squares formulation. Here the latter is implemented in connection with the Crank-Nicholson scheme for the time-integration, as described in [5]. This comparative study is illustrated by means of some results extracted from [8], in the framework of a test-problem described below, where uniform meshes are used, thereby allowing the use of weights all equal to $1/(N+2)$.

Take Ω to be the unit square $(0, 1) \times (0, 1)$ and a space discretisation based on a uniform $L \times L$ mesh in which every square cell is subdivided into two triangles by means of the diagonal parallel to the line $x_1 = x_2$. We take $T = 0.1$ and $\nu = 10^{-k}$, for $k = 2$ and $k = 5$, and $\mathbf{a} = (1/\pi; 1/\pi)$. For this choice the Péclet number equals $10^k/\pi$. An exact solution is considered to be a function with a double boundary layer in the neighborhood of the edges given by $x_1 = 1$ and $x_2 = 1$. More specifically we take

$$u(x_1, x_2, t) = e^{-t}\big[s(x_1)\,\sin(\pi x_2) + s(x_2)\,\sin(\pi x_1)\big]$$

where for $z \in [0, 1]$,

$$s(z) := \frac{\big[e^{\frac{\pi^2\nu-1}{\pi\nu}} - 1\big]e^{\pi z} + (e^{\pi} - 1)e^{\frac{(\pi^2\nu-1)(1-z)}{\pi\nu}}}{e^{\frac{2\pi^2\nu-1}{\pi\nu}} - 1}.$$

Equation (1) has effectively the above solution, provided the right hand side is given by $f(x_1, x_2, t) = e^{-t}[s(x_1)\cos(\pi x_2) + s(x_2)\cos(\pi x_1)]$, and the prescribed initial and boundary values are those of u.

In self explanatory Tables 1 through 4 we summarize the results generated by both methods in the solution of the above test problem, by showing both L^2 and L^∞ relative errors.

L	Weighted explicit method	Least-squares formulation
16	$0.27550 \times 10^{+0}$	0.74477×10^{-1}
32	$0.13929 \times 10^{+0}$	0.20806×10^{-1}
64	0.68628×10^{-1}	0.53658×10^{-2}
128	0.50245×10^{-1}	0.13526×10^{-2}

Table 1 – Relative error of u in the norm of $L^2(\Omega)$ for t=0.1 with Pé = 0.3183×10^2.

L	Weighted explicit method	Least-squares formulation
16	$0.84379 \times 10^{+0}$	$0.10739 \times 10^{+0}$
32	$0.40427 \times 10^{+0}$	0.33655×10^{-1}
64	$0.21973 \times 10^{+0}$	0.78437×10^{-2}
128	$0.16351 \times 10^{+0}$	0.19133×10^{-2}

Table 2 – Maximum relative error of u in Ω for t=0.1 with Pé = 0.3183×10^2.

Observation of Tables 1 and 2 leads to the conclusion that both methods being compared simulate correctly thicker boundary layers, that is, those corresponding to a moderate Péclet number close to 10^2. Moreover, as one can infer from the above results, our scheme is much less accurate than the least-squares formulation in this case. This is quite natural for the latter is a second order method in the L^2-norm, and empirically in the L^∞-norm too, whereas the former is a quasi first order method in the L^∞-norm (cf. [7]). Nevertheless the least-squares approach failed completely in the case of a thin boundary layer corresponding to Pé roughly equal to 10^5, as clearly indicated in Tables 3 and 4, while our explicit scheme showed a much better behavior. However even so the errors in the maximum norm are not so small. In this respect it is worthwhile commenting that such results are to be expected. Indeed the maximum error values tend to occur precisely in the interior of the thin boundary layer, which cannot be reached by any numerical method, at least not for the degree of mesh refinement used in the above test.

L	Weighted explicit method	Least-squares formulation
16	$0.16682 \times 10^{+0}$	$0.18919 \times 10^{+0}$
32	$0.14754 \times 10^{+0}$	$0.20459 \times 10^{+0}$
64	$0.12560 \times 10^{+0}$	$0.27201 \times 10^{+0}$
128	0.99616×10^{-1}	$0.29258 \times 10^{+0}$

Table 3 – Relative error of u in the norm of $L^2(\Omega)$ for t =0.1 with Pé = 0.3183×10^5.

L	Weighted explicit method	Least-squares formulation
16	$0.33093 \times 10^{+0}$	$0.88493 \times 10^{+0}$
32	$0.32787 \times 10^{+0}$	$0.16663 \times 10^{+1}$
64	$0.32539 \times 10^{+0}$	$0.23959 \times 10^{+1}$
128	$0.32244 \times 10^{+0}$	$0.23286 \times 10^{+1}$

Table 4 – Maximum relative error of u in Ω for t=0.1 with Pé = 0.3183×10^5.

Acknowledgement. The author is thankful to Carlos Tomei from the Department of Mathematics of PUC-Rio for fruitful discussions.

REFERENCES

[1] R.A. Adams, *Sobolev Spaces*, Academic Press, N.Y., (1975).

[2] A.N. Brooks and T.J.R. Hughes, *The streamline upwind/Petrov-Galerkin formulation for convection dominated flows with particular emphasis on the Navier-Stokes equations*, Computer Methods in Applied Mechanics and Engineering, **32** (1982), 199–259.

[3] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Noth-Holland, Amsterdam (1978).

[4] M. Kawahara, N. Takeuchi and Y. Yoshida, *Two step finite element methods for Tsunami wave propagation analysis*, International Journal of Numerical Methods in Engineering, **12** (1978), 331–351.

[5] R. Leal Toledo and V. Ruas, *Numerical analysis of a least-squares finite element method for the time-dependent advection-diffusion equation*, Journal of Computational and Applied Mathematics, **235** (2011), 3615–3631.

[6] V. Ruas and A.P. Brasil Jr., *A stable explicit method for time-dependent convection-diffusion equations*, Proc. ICNAAM, Greece, (2007), 480–483.

[7] V. Ruas, A.P. Brasil Jr. and P. Trales, *An explicit scheme for solving convection-diffusion equations*, Japan Journal of Industrial and Applied Mathematics, **26** (2009), 65–91.

[8] V. Ruas, M. Kischinhevsky and R. Leal Toledo, *Elementos finitos em formulação mista de mínimos quadrados para a simulação da convecção-difusão em regime transiente*, to appear in Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería (January 2013).

[9] G. Strang and G.J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs (1973).

[10] M. Tabata, *Uniform convergence of the upwind finite element approximation for semilinear parabolic problems*, Journal of Mathematics of the Kyoto University, **18-2** (1978), 327–351.