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.

*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 $h|ln h|$ 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 outline 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 \( u(x, t) \) defined in \( \Omega \times [0, \infty) \) such that,

\[
\begin{align*}
  &u_t + a \cdot \nabla u - \nu \Delta u = f \quad \text{in } \Omega \times (0, \infty) \\
  &u = g \quad \text{on } \partial \Omega \times (0, \infty) \\
  &u = u^0 \quad \text{in } \Omega \quad \text{for } t = 0
\end{align*}
\]

(1)

where \( u_t \) represents the first order derivative of \( u \) with respect to \( t \), \( \nu \) is a positive constant and \( a \) 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 \( x \in \Omega \) \( g(x, \cdot) \) is of bounded variation in \( (0, \infty) \). In (1) \( \nu \) represents the inverse of the Péclet number.

Without loss of essential aspects, in all the sequel we assume that \( \Omega \) is an interval if \( N = 1 \), a polygon if \( N = 2 \) or a polyhedron if \( N = 3 \). In so doing we consider a partition \( T_h \) of \( \Omega \) into \( N - 1 \) simplices, with maximum edge length equal to \( h \). We assume that \( T_h \) 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 \( h_{\text{min}} \) as the minimum height of all the elements of \( T_h \) if \( N = 2 \) or 3 and the minimum length of \( K \in T_h \) if \( N = 1 \).

Let \( N_h \) be the number of nodes of \( T_h \), 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 \( \Omega \) and the remaining \( N_h - I_h \) nodes are located on \( \partial \Omega \).

Now for every \( K \in T_h \) we denote by \( P_l(K) \) 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 $\mathcal{T}_h$:

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

$$V^0_h := V_h \cap H^1_0(\bar{\Omega}).$$

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

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

Now let $u^0_h$ be the field of $V^g(0)$ satisfying $u^0_h(P_j) = u^0(P_j)$ 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 $\Omega$ by $f^n(\cdot) = f(\cdot, n\Delta t)$ and $a^n$ in $\Omega$ by $a^n(\cdot) = a(\cdot, n\Delta t)$, for $n = 1, 2, \ldots$, ideally we wish to determine approximations $u^n_h(\cdot)$ 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, ..., we wish to determine $u^n_h \in V^g_n$ of the form

$$u^n_h = \sum_{j=1}^{N_h} u^n_j \varphi_j,$$

where $\varphi_j$ is the canonical basis function of $V_h$ associated with the $j$-th node of $\mathcal{T}_h$, i.e. $P_j, u^n_j \in \mathbb{R}$ being the value of $u^n_h$ at $P_j$. We denote by $S_j$ the support of $\varphi_j$ and by $\Pi_j$ its measure.

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

$$m^L_{ij} u^n_i = \sum_{j=1}^{N_h} \left[ m^n_{ij} - \Delta t a^n_{ij} \right] u^{n-1}_j + \Delta t b^n_i, \text{ for } i = 1, \ldots, I_h, \quad (2)$$

where, setting $a^n_i := a^n(P_i)$ and $f^n_i := f^n(P_i)$, the coefficients $a^n_{ij}$ and $b^n_i$ are given by:

$$a^n_{ij} = \int_{\Omega} \left[ (a^n_i \cdot \nabla) \varphi_j \varphi_i + v \nabla \varphi_j \cdot \nabla \varphi_i \right], \quad (3)$$

$$b^n_i = \int_{\Omega} f^{n-1}_i \varphi_i.$$
Coefficient \( m^L_{ii} \) is the well-known lumped mass diagonal matrix (cf. [9]) given by \( \frac{\Pi_i}{N+1} \). The mass matrix coefficients \( m^W_{ij} \) 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 \( j = 1, 2, \ldots, M_i \) with \( 1 \leq k_j \leq N_h \). Let also \( W_j^i \) be the measure fractions associated with \( P_{k_j} \) given by:

\[
W_j^i = \frac{\text{meas}(S_i \cap S_{k_j})}{N + 1} \quad (4)
\]

and \( \omega_j^i \) be corresponding strictly positive weights satisfying,

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

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

\[
\sum_{j=1}^{M_i} W_j^i = \frac{N \Pi_i}{N + 1} \quad (6)
\]

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

\[
m^W_{ik_j} = \frac{h_{\min}}{v + h_{\min}} \omega_j^i W_j^i \quad \text{for} \ i \neq k_j
\]

(7)

together with

\[
m^W_{ii} = \frac{v}{h_{\min} + v} m^L_{ii} + \frac{h_{\min}}{h_{\min} + v} m^C_{ii}.
\]

(8)

where \( m^C_{ii} \) is the \( i - \text{th} \) diagonal coefficient of the standard consistent mass matrix ([9]) given by

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

Naturally enough, by definition, \( m^W_{ij} = 0 \) 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

\[
\frac{v}{h_{\min} + v} \quad \text{and} \quad \frac{h_{\min}}{h_{\min} + v},
\]

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 \( \Delta t \) 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 \( \Delta t \) go to zero.

First we have to define the following quantities:

- \( A = \sup_{t \in (0, \infty)} \max_{1 \leq i \leq I} h_i |a_i(t)| \);
- \( \omega = \min_{1 \leq i \leq I} \min_{1 \leq j \leq M_i} \omega_{ij} \).

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

**Theorem 1.** If \( \Delta t \) fulfills the condition

\[
\Delta t \leq \frac{\omega h^3_{min}}{(v + h_{min})[Ah_{min} + (N + 1)v]},
\]

then the finite element solution sequence \( \{u^n_h\}_n \) given by \( u^n_h = \sum_{j=1}^{N_h} u^n_j \varphi_j \) 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 \( L^\infty \)-norm of a function \( F \) defined in an open set \( D \) of \( \mathbb{R}^N \), and \( BV[G] \) represents the standard norm of a function \( G(t) \) having bounded variation for \( t \in (0, \infty) \):

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

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 $T_h$ is of the acute type (cf. [10]). Then if $\Delta t$ satisfies the condition
\[
\Delta t \leq \frac{h_{\text{min}}^2}{\nu + h_{\text{min}}} \\min \left[ \frac{\omega}{A'}, \frac{v(N + 2) + 2h_{\text{min}}}{v(N + 1)(N + 2)} \right]
\]  
the finite element solution sequence $\{u^n_h\}_n$ given by $u^n_h = \sum_{j=1}^{N_h} u^n_j \varphi_j$ generated by (2) for $n = 1, 2, \ldots$ satisfies the stability condition (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 $\omega_j$ are suitably chosen, this scheme provides convergent approximations in the maximum norm, as both $h$ and $\Delta t$ go to zero, under the assumption (11) of Theorem 2. For this purpose we use Sobolev spaces $W^{m,\infty}(D)$ equipped with the standard 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 $\mathbb{R}^N$ (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 $T_h$, for $i \in \{1, 2, \ldots, I_h\}$, and $\Psi_j$ 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 $\omega_j$ satisfying (5) such that
\[
\sum_{j=1}^{M_i} \omega_j \Psi_j' \Psi_j = 0.
\]  

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

Theorem 2. (cf. [7]). Let the strictly positive weights $\omega_j$, $\forall i \in \{1, 2, \ldots, I_h\}$ 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 $u_i$ belong to $W^{2,\infty}(\Omega)$. Assume also that $\forall t \in [0, T]$, $a(\cdot, t) \in [W^{1,\infty}(\Omega)]^N$ and $f(\cdot, t) \in W^{1,\infty}(\Omega)$, and $(u_i)_t$ 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 \( \Delta t := \frac{T}{k} \) 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 \( \omega \) is bounded below away from zero independently of \( h \). Then there exists a constant \( C \) independent of \( u \), \( h \) and \( \Delta t \), such that the following estimate applies:

\[
\max_{1 \leq m \leq M_i} \|u^m - u_h^m\|_{0, \infty, \Omega} \leq C h |\ln h| \max_{0 \leq s \leq T} \|u(\cdot, s)\|_{2, \infty, \Omega} + h \|u_t(\cdot, s)\|_{0, \infty, \Omega} + \|a(\cdot, s)\|_{1, \infty, \Omega} \|u(\cdot, s)\|_{1, \infty, \Omega} + \|f(\cdot, s)\|_{1, \infty, \Omega}
\]

(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 \( (\omega_i^1, \omega_i^2) \) associated with inner node \( P_i \) is uniquely defined by the equations:

\[
\begin{aligned}
-\omega_i^1 (l_i^1)^2 + \omega_i^2 (l_i^2)^2 &= 0 \\
\frac{1}{2} [\omega_i^1 l_i^1 + \omega_i^2 l_i^2] &= \frac{l_i^1 + l_i^2}{6},
\end{aligned}
\]

(14)

where \( l_i^1 \) and \( l_i^2 \) are the lengths of the intervals of \( \mathcal{T}_h \) having \( P_i \) as the right and the left end, respectively. This yields \( \omega_i^1 = \frac{l_i^2}{3l_i^1} \) and \( \omega_i^2 = \frac{l_i^1}{3l_i^2} \).

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 $E^+_i := \{e^i_j\}_{j=1}^{M_i}$ be the set of unit vectors corresponding to the vectors $l^i_j$, that is, $e^i_j = l^i_j / l^i_j$, where $l^i_j$ is the modulus of $l^i_j$. Let also $E^-_i := \{-e^i_j\}_{j=1}^{M_i}$, and $E_i = E^+_i \cup E^-_i$. Setting $K_i = \text{card}(E_i)$, 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 $j_1 \in \{1, \ldots, M_i\}$ there is necessarily another $j_2 \in \{1, \ldots, M_i\}$ such that $e^i_{j_1} + e^i_{j_2} = 0$. This implies in particular that $M_i$ must be an even number. Hence we may choose the weights in pairs, say $(\omega^i_{j_1}, \omega^i_{j_2})$, 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:

$$
\begin{align*}
-\omega^i_{j_1} l^i_{j_1} W^i_{j_1} + \omega^i_{j_2} l^i_{j_2} W^i_{j_2} &= 0 \\
\omega^i_{j_1} W^i_{j_1} + \omega^i_{j_2} W^i_{j_2} &= \frac{W^i_{j_1} + W^i_{j_2}}{N + 2},
\end{align*}
$$

for all pairs $(j_1, j_2) \in \{1, \ldots, M_i/2\} \times \{M_i/2 + 1, \ldots, M_i\}$, such that $e^i_{j_1} + e^i_{j_2} = 0$. 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 $e^i_k$, $k = M_i + 1, \ldots, K_i$. By assumption the intersection of $\tilde{S}_i$ with the half straight line with origin at $P_i$ and oriented in the sense and direction of $e^i_k$ for $k > M_i$, cannot be an edge $P_i \tilde{P}_{kj}$ with $j \leq M_i$. Letting the segment $P_i \tilde{Q}^i_k$ be such intersection, where $\tilde{Q}^i_k$ 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 $\tilde{Q}^i_k$, for $k = M_i + 1, \ldots, K_i$, still denoted by $l^i_k$, is a non trivial convex combination of either exactly $N$ edge vectors $l^i_j$ 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 $P_{m_l}$, for $l = 1, \ldots, J$ with $1 \leq m_l \leq M_i$, be the vertices of $S_i$ whose convex combination yields $l^i_k$ for $k = M_i + 1, \ldots, K_i$. Let us denote the modulus of $l^i_k$ by $|l^i_k|$, for $k = M_i + 1, \ldots, K_i$ too. In so doing we have:

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

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

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

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

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

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

(17)

together with

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

(18)
as one can easily check.

Now we replace in (17) the $l^i_k$'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 $\omega^i_j$ defined in the following manner:

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

(19)

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

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

(20)

where $\beta^k_j = 0$ if $Q^i_k$ does not belong to $S_i \cap S_{ij}$ and $\beta^k_j = \alpha^k_i$ for the pertaining convex combination coefficients in (16), that is, all the $\alpha^k_i$’s such that $m_i = j$. Notice that by construction (5) holds for the weights $\omega^i_j$ 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 $\omega$ 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

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

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

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

(21)

Since

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

and from (18) it holds that

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

(21) we obtain,

\[
C_i \geq \frac{1}{1 + \frac{M_i(K_i - M_i)h}{k_i h_{\min}}}.
\]

(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 \(\{T_h\}_h\) under consideration, \(M_i\) exceeds the value \(c^{-N}\), where \(c\) is a mesh independent constant such that \(\rho \geq ch\) for every \(h\), \(\rho\) being the minimum over the elements of \(T_h\) of the radii of the largest inscribed balls in the elements of \(T_h\) for a given \(h\) (cf. [3]).

Finally, since \(\delta^j_i > 0\) for all \(i, j\) we have,

\[
\omega^j_i \geq c_0 h_{\min} \tilde{W}^j_i \geq \frac{c_0}{N + 2} \left( \frac{h_{\min}}{h} \right)^{N+1}.
\]

(23)

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

\[
\omega^j_i \geq c_N \ \forall i \in \{1, \ldots, I_h\} \text{ and } \forall j \in \{1, \ldots, M_i\}.
\]

(24)

**Remark 4.2.** (24) clearly indicates that \(c_N\) may play the role of the parameter \(\omega\) in the relations (9) and (11). However for very distorted meshes such a value of \(\omega\) 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 \(\omega^j_i\)'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 \(E_i^j := \{e^j_i\}_{j=1}^{2M_i}\), where the first \(M_i\) vectors are those of \(E_i^+\) and the last \(M_i\) vectors are given by \(e^{j+M_i}_i := -e^j_i\).
for \( j = 1, \ldots, M_i \). Now let the segment \( P_i Q^i_k \) be the intersection of \( \tilde{S}_i \) with the half straight line with origin at \( P_i \) and oriented in the sense and direction of \( e^i_k \) for \( k > M_i \), \( Q^i_k \) being necessarily a point of the boundary of \( S_i \). It follows that \( P_i Q^i_k \) 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 \( Q^i_k \), for \( k = M_i + 1, \ldots, 2M_i \), still denoted by \( I^i_k \), is a convex combination of at most \( N \) edge vectors \( I^i_j \) 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 \( Q^i_k \) coincides with a given vertex of \( S_i \), \( J = 2 \) for \( N = 3 \) only if \( Q^i_k \) is contained in a common face of two neighboring tetrahedra of \( S_i \), or \( J = N \) for \( N = 2 \) or \( N = 3 \) otherwise. Let \( P_{k_{a_j}} \), for \( j = 1, \ldots, J \) with \( 1 \leq m_1 \leq M_i \), be the vertices of \( S_i \) whose convex combination yields \( I^i_k \) for \( k = M_i + 1, \ldots, 2M_i \). Here again we denote the modulus of \( I^i_k \) by \( I^i_k \), for \( k = M_i + 1, \ldots, 2M_i \) too.

In so doing we have:

\[
I^i_k = \sum_{l=1}^{J} \alpha^i_k I^i_{m_l}, \quad k = M_i + 1, \ldots, 2M_i, \tag{25}
\]

where the \( \alpha^i_k \)'s for \( l = 1, \ldots, J \), are coefficients of a convex combination, that is, \( 0 \leq \alpha^i_k \leq 1 \), \( l = 1, \ldots, J \) and \( k = M_i + 1, \ldots, 2M_i \), with \( \sum_{l=1}^{J} \alpha^i_k = 1 \).

Now we assign to each point \( Q^i_k \), \( k = M_i + 1, \ldots, 2M_i \) the measure fraction \( W^i_s = W^i_{s-M_i} \), and the weight \( \tilde{\omega}^i_k = \frac{I^i_{k-M_i}}{I^i_k} \tilde{\omega}^i_{k-M_i} \), where \( \tilde{\omega}^i_j \) are provisional weights respectively associated with the vertices \( P_{k_j} \) of \( S_i \) different from \( P_i \) satisfying \( \tilde{\omega}^i_k + \tilde{\omega}^i_{k-M_i} = \frac{1}{N+2} \). Then similarly to the case of (15) we necessarily have:

\[
\sum_{j=1}^{2M_i} \tilde{\omega}^i_j W^i_j = 0, \tag{26}
\]

together with

\[
\sum_{j=1}^{2M_i} \tilde{\omega}^i_j W^i_j = \frac{N \Pi_i}{(N+1)(N+2)}, \tag{27}
\]

as one can easily check.

Now we replace in (26) the $l_i^k$’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 $\omega_j^i$ defined in the following manner:

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

(28)

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

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

(29)

where $\beta_j^k = 0$ if $Q_i^k$ does not belong to $S_i \cap S_{kj}$ and $\beta_j^k = \alpha_k^j$ for the pertaining convex combination coefficients in (25), that is, all the $\alpha_k^j$’s such that $m_l = j$. Notice that by construction (5) holds for the weights $\omega_j^i$ 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 $\omega$ 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 $\omega_j^i$ attempt to take into account three main factors playing a role in the influence of the vector $l_i^j$ 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 $W_j^i$, which goes in the same sense as $l_i^j$. Finally the increments $\delta_j^i$, 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 $l_i^j$, that is, making angles with it close to $\pi$. In this sense we may assert that the above described procedure for determining the $\omega_j^i$’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 $\Omega$ 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 $v = 10^{-k}$, for $k = 2$ and $k = 5$, and $a = (1/\pi; 1/\pi)$. 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

\[
\begin{align*}
\mathbf{u}(x_1, x_2, t) &= e^{-t} \left[ s(x_1) \sin(\pi x_2) + s(x_2) \sin(\pi x_1) \right]
\end{align*}
\]

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

\[
\begin{align*}
\mathbf{s}(z) := \left[ e^{\frac{\pi^2 z^2 - 1}{\pi v}} - 1 \right] e^{\pi z} + (e^{\pi} - 1) e^{\frac{\pi^2 z^2 - 1 + (1-z)}{\pi v}}.
\end{align*}
\]
Equation (1) has effectively the above solution, provided the right hand side is given by
\[ f(x_1, x_2, t) = e^{-t} \left[ s(x_1) \cos(\pi x_2) + s(x_2) \cos(\pi x_1) \right], \]
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 \( L^\infty \) relative errors.

<table>
<thead>
<tr>
<th>( L )</th>
<th>Weighted explicit method</th>
<th>Least-squares formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.27550 × 10^0</td>
<td>0.74477 × 10^1</td>
</tr>
<tr>
<td>32</td>
<td>0.13929 × 10^0</td>
<td>0.20806 × 10^1</td>
</tr>
<tr>
<td>64</td>
<td>0.68628 × 10^-1</td>
<td>0.53658 × 10^-2</td>
</tr>
<tr>
<td>128</td>
<td>0.50245 × 10^-1</td>
<td>0.13526 × 10^-2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>( L )</th>
<th>Weighted explicit method</th>
<th>Least-squares formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.84379 × 10^0</td>
<td>0.10739 × 10^0</td>
</tr>
<tr>
<td>32</td>
<td>0.40427 × 10^0</td>
<td>0.33655 × 10^-1</td>
</tr>
<tr>
<td>64</td>
<td>0.21973 × 10^-1</td>
<td>0.78437 × 10^-2</td>
</tr>
<tr>
<td>128</td>
<td>0.16351 × 10^-2</td>
<td>0.19133 × 10^-2</td>
</tr>
</tbody>
</table>

Table 2 – Maximum relative error of \( u \) in \( \Omega \) for \( t=0.1 \) with \( \text{Pé} = 0.3183 \times 10^2 \).

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 Pécellet 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 \( L^\infty \)-norm too, whereas the former is a quasi first order method in the \( L^\infty \)-norm (cf. [7]). Nevertheless the least-squares approach failed completely in the case of a thin boundary layer corresponding to \( \text{Pé} \) 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.

<table>
<thead>
<tr>
<th>$L$</th>
<th>Weighted explicit method</th>
<th>Least-squares formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>$0.16682 \times 10^{+0}$</td>
<td>$0.18919 \times 10^{+0}$</td>
</tr>
<tr>
<td>32</td>
<td>$0.14754 \times 10^{+0}$</td>
<td>$0.20459 \times 10^{+0}$</td>
</tr>
<tr>
<td>64</td>
<td>$0.12560 \times 10^{+0}$</td>
<td>$0.27201 \times 10^{+0}$</td>
</tr>
<tr>
<td>128</td>
<td>$0.99616 \times 10^{-1}$</td>
<td>$0.29258 \times 10^{+0}$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$L$</th>
<th>Weighted explicit method</th>
<th>Least-squares formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>$0.33093 \times 10^{+0}$</td>
<td>$0.88493 \times 10^{+0}$</td>
</tr>
<tr>
<td>32</td>
<td>$0.32787 \times 10^{+0}$</td>
<td>$0.16663 \times 10^{+1}$</td>
</tr>
<tr>
<td>64</td>
<td>$0.32539 \times 10^{+0}$</td>
<td>$0.23959 \times 10^{+1}$</td>
</tr>
<tr>
<td>128</td>
<td>$0.32244 \times 10^{+0}$</td>
<td>$0.23286 \times 10^{+1}$</td>
</tr>
</tbody>
</table>

Table 4 – Maximum relative error of $u$ in $\Omega$ for $t = 0.1$ with $Pé = 0.3183 \times 10^5$.

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

REFERENCES


