Computer Methods in Applied Mechanics and Engineering 110 (1993) 325–342
North-Holland

A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation

Ramon Codina

Escola Tècnica Superior d'Enginyers de Camins, Canals i Ports, Universitat Politècnica de Catalunya, Gran Capità s/n, Mòdul C1 08034 Barcelona, Spain

Received 4 December 1992

To avoid the local oscillations that still remain using the streamline-upwind/Petrov–Galerkin formulation for the scalar convection–diffusion equation, the introduction of a nonlinear crosswind dissipation is proposed. It is shown that the method is less overdiffusive than other discontinuity-capturing techniques and has better numerical behavior. The design of the crosswind diffusion is based on the study of the discrete maximum principle for some simple cases.

1. Introduction

In spite of the simple model represented by the scalar convection–diffusion equation, its numerical solution is still a challenge when convection is highly dominant, in which case the standard Galerkin method fails to give reasonable results. Among the many finite element methods that allow us to circumvent this problem, here we use the streamline-upwind/Petrov–Galerkin (SUPG) method (see, e.g., [1]).

The SUPG formulation does not preclude the presence of overshoots and undershoots in the vicinity of sharp gradients of the solution. Near optimal global error estimates have been obtained and it is also possible to obtain near optimal local error estimates outside a small neighborhood containing the layers [2, 3]. This ensures that the numerical solution will not be globally deteriorated.

In certain situations, not even the small overshooting and undershooting found using the SUPG method are permissible. This happens for instance in the numerical simulation of compressible flow problems, where the solution may develop discontinuities (shocks) whose poor resolution may affect the global stability of the numerical calculations due to the nonlinear nature of the problem. The numerical solution of the convection–diffusion equation with a very high Péclet number provides a good model to develop numerical strategies to remove small oscillations about abrupt layers of the solution.

The reason why overshooting and undershooting appear using the SUPG method is that it is neither a monotone nor a monotonicity preserving method. A numerical method is said to be monotone if the numerical solution for all time steps retains the sign of the previous time step at all the nodes of the spatial mesh. If only the monotonicity of the initial data is maintained, the method is called monotonicity preserving [4]. To design a high order accurate and monotone method is not easy. Godunov's theorem (cf. [4]) establishes that a linear, monotonicity preserving method is at most first order accurate. Therefore, the only feasible way to achieve the goals of high accuracy in regions where the solution is smooth and to avoid oscillations about layers is to design a nonlinear method, that is, a numerical scheme which depends itself on the numerical solution. The main idea of any shock-capturing (or discontinuity-capturing) technique is to increase the amount of numerical dissipation in the neighborhood of layers.

Correspondence to: Dr. R. Codina, Escola Tècnica Superior d'Enginyers de Camins, Canals i Ports, Universitat Politècnica de Catalunya, Campus Nort UPC, Gran Capità s/n, Mòdul C1, 08034 Barcelona, Spain.

Several shock-capturing methods have been developed, both using finite difference and finite element techniques. Here we do not treat the former, for which a vast literature exists. We refer to [4–6] for a review of these methods, with special reference to compressible flow problems and systems of conservation laws. Concerning finite element methods, one of the most common approaches is described in Section 2.

For the particular case of the convection–diffusion equation, a possible way to treat the problem is the satisfaction of the discrete maximum principle (see [7] for a thorough description of early finite element methods designed with this property). In Section 3, we adopt this point of view and consider several particular cases (1D problems with linear elements, linear and multilinear elements in multidimensional problems). Although from these examples the main conclusion is that it is difficult to take the discrete maximum principle as a starting point to design shock-capturing techniques, it provides the underlying idea for the method proposed in Section 4. Assuming that the streamline dissipation introduced by the SUPG method is enough to avoid oscillations in this direction, only the crosswind diffusion has to be increased. For consistency, the new dissipation added must be proportional to the element residual and, for accuracy, it must vanish quickly in regions where the solution is smooth and also where the convective term of the residual is small. The previous study of the discrete maximum principle is used to set the expression of the numerical crosswind dissipation. All these ideas are developed in Section 4. Numerical results using this new approach are presented in Section 5, showing a good resolution of the layers and also excellent convergence properties, in the sense to be explained later.

2. Background

2.1. *Basic formulation*

Let us introduce some notation first. A finite element partition of the computational domain $\Omega \subset \mathbb{R}^{N_{\rm sd}}$ (open, bounded and polyhedral, $N_{\rm sd} = 2$ or 3) is denoted by $\{\Omega^e\}$, the index e ranging from 1 to the number of elements $N_{\rm el}$ and h^e being the characteristic length of the eth element. For simplicity, Dirichlet boundary conditions $\phi = g$ are considered on the whole boundary $\partial\Omega$, ϕ being the unknown function. The space of trial solutions is $\Phi = \{\phi \in H^1(\Omega)$: $\phi = g$ on $\partial\Omega\}$ and the space of test functions $\Psi = H_0^1(\Omega)$. A subscript h is introduced to refer to the discrete finite element problem.

Although we are interested only in the steady-state problem, it is convenient here to consider the transient convection–diffusion equation

$$\frac{\partial \phi}{\partial t} + \boldsymbol{u} \cdot \nabla\phi - k\,\Delta\phi = f \quad \text{in } \Omega \times (0, T) \tag{1}$$

where $\boldsymbol{u}$ is the velocity field, k the diffusion coefficient (assumed throughout to be small) and f is the source term. An initial condition and the boundary condition $\phi = g$ have to be appended to (1).

The finite element formulation for solving (1) that we use is the following: Find $\phi_h \in \Phi_h$ such that

$$\int_\Omega \psi_h \left[\frac{\partial \phi_h}{\partial t} + \boldsymbol{u} \cdot \nabla\phi_h\right] \mathrm{d}\Omega + k \int_\Omega \nabla\psi_h \cdot \nabla\phi_h \,\mathrm{d}\Omega + \sum_{e=1}^{N_{\rm el}} \int_{\Omega^e} \zeta_h \mathcal{R}(\phi_h)\,\mathrm{d}\Omega = \int_\Omega \psi_h f \,\mathrm{d}\Omega \tag{2}$$

for all $\psi_h \in \Psi_h$, where

$$\mathcal{R}(\phi_h) := \boldsymbol{u} \cdot \nabla\phi_h - k\,\Delta\phi_h - f\,, \tag{3}$$

$$\zeta_h = \tau \boldsymbol{u} \cdot \nabla\psi_h\,, \tag{4}$$

$$\tau = \frac{\alpha h}{2|\boldsymbol{u}|}\,, \tag{5}$$

and α is the so-called upwind function, which depends on the element Péclet number

$$\gamma := \frac{|u|h}{2k} . \tag{6}$$

It is understood that all the terms in (3)–(6) are computed for each element. In (2), the term $\partial\phi_h/\partial t$ should also be weighted by ζ_h, although this is immaterial for what follows.

A possible way to determine the expression of the upwind function α is to impose that the solution of the 1D steady-state problem be nodally exact. It is well known that for linear elements this leads to $\alpha = \coth \gamma - 1/\gamma$. The asymptotic approximation $\alpha = \min(\gamma/3, 1)$ is also often used.

Using the same methodology, we have obtained the upwind functions for quadratic elements [8]. In this case, two upwind functions are needed, one for the extreme nodes and another for the central nodes. It is also possible to use a unique upwind function, which is much simpler although not so accurate. Moreover, if the asymptotic approximation is employed, the upwind functions may be taken as

$$\alpha = \alpha_0 \min(\gamma/3, 1) , \tag{7}$$

with $\alpha_0 = 1$ for linear elements and $\alpha_0 = 0.5$ for quadratic elements using the SUPG approach. Hereafter, we use this choice for α.

2.2. Time relaxation

Once a finite element basis (shape functions) has been chosen, problem (2) may be recast in matrix form as

$$\boldsymbol{M}\dot{\boldsymbol{\phi}} + \boldsymbol{K}\boldsymbol{\phi} = \boldsymbol{F} , \tag{8}$$

the dot denoting the temporal derivative and $\boldsymbol{\phi}$ being the vector of nodal unknowns of the function ϕ_h. Let Δt be the time step size (constant or not) of a discretization of the time interval and let a superscript n indicate an approximation of the vector $\boldsymbol{\phi}$ at $t^n = n\,\Delta t$. The forward Euler scheme applied to (8) (with an appropriate initial condition) reads as follows: Given $\boldsymbol{\phi}^n$ compute $\boldsymbol{\phi}^{n+1}$ as

$$\boldsymbol{\phi}^{n+1} = \boldsymbol{\phi}^n + \Delta t\, \boldsymbol{M}_d^{-1}(\boldsymbol{F}^n - \boldsymbol{K}\boldsymbol{\phi}^n) , \tag{9}$$

where we have introduced the diagonal ('lumped') matrix $\boldsymbol{M}_d$, approximation to the matrix $\boldsymbol{M}$. This approximation may be obtained by using a nodal quadrature rule to evaluate the integrals appearing in the expression of $\boldsymbol{M}$ when the standard finite element basis is chosen. If this is done, the influence of the SUPG perturbation in the mass matrix is negligible (zero for uniform meshes and solenoidal velocities). This is why it has not been included in weighting of the temporal derivative in (2).

Algorithm (9) is only conditionally stable. We have obtained the stability limits for the SUPG formulation using quadratic elements for the 1D problem [9]. It is found that the stability condition is

$$c \leqslant \frac{\gamma}{8(1+\alpha\gamma)} , \tag{10}$$

which in the advective limit ($\gamma \to \infty$) and in the diffusive limit ($|u| \to 0$) reduces to

$$c \leqslant \frac{1}{8\alpha_0} \quad \text{and} \quad \Delta t \leqslant \frac{h^2}{16k} , \tag{11}$$

respectively. In (10) and (11), $c := u\,\Delta t/h$ is the Courant number.

The stability limit (10) is in clear contrast with what one finds for linear elements:

$$c \leq \frac{\gamma}{1+\alpha\gamma} . \tag{12}$$

For the same number of nodal points, the critical time step will be larger using linear elements than using quadratics (four times, according to the expression of the Courant number we use). Nevertheless, it is shown in [9] that quadratic elements are more diffusive (in time) and the total number of time steps needed to reach the steady-state is similar to that needed for linear elements.

In 2D, we do the following. For each node of the mesh, the critical time step size is computed as

$$\Delta t = s \frac{\Delta t_\sigma \, \Delta t_\nu}{\Delta t_\sigma + \Delta t_\nu} , \tag{13}$$

where Δt_σ is computed using the general 1D expression and Δt_ν using the 1D diffusive limit. Subscripts σ and ν refer to the streamline and the crosswind directions, respectively. The element length needed to compute (13) is taken as the minimum of the elements surrounding the node under consideration. In (13), s is a safety factor that in most cases may be taken as $s = 1$ (always using the SUPG formulation). We use local time stepping and thus (13) is used for each equation of the algorithm (9). Otherwise, Δt could be taken as the minimum for all the nodes of the expression given by (13).

In Section 5, when talking about the convergence rate of a certain method, we refer to the convergence towards the steady-state solution. The residual for the nth time step is

$$\text{Residual} = \frac{|\boldsymbol{\phi}^{n+1} - \boldsymbol{\phi}^{n}|}{|\boldsymbol{\phi}^{n+1}|} . \tag{14}$$

Algorithm (9) is viewed in what follows as an iterative procedure to obtain the steady-state solution. The possible nonlinearity of the scheme is dealt with in the same iterative loop. This approach is widespread in computational fluid dynamics, especially in the numerical simulation of inviscid flows.

2.3. A shock-capturing technique

The basic idea behind the SUPG method is to introduce numerical diffusion along the streamlines in a consistent manner. The streamline as the best upwind direction was questioned by Mizukami and Hughes in [10], and an upwind scheme satisfying the maximum principle was especially developed for linear triangular elements. Another monotone algorithm was presented by Rice and Schnipke in [11] for bilinear quadrangular elements. Both methods are restricted to the elements for which they were designed and no generalization seems easy. Since the oscillations observed using the SUPG formulation were placed in directions normal to the gradient of transported quantity, Hughes et al. proposed to introduce another diffusion in this direction [1], in a similar way to that proposed by Davis for finite differences [12]. This new diffusion is consistently introduced as another term in the weighting functions called discontinuity capturing. Extensions to systems were studied in [13]. The method was initially adopted by Johnson and Szepessy in [14] using space-time finite element discretizations and, with a slight modification, used to prove convergence for the inviscid Burgers equation to an entropy solution in [15]. Other discontinuity capturing terms have been proposed, although all of them keeping the SUPG (or Galerkin/least-squares [16]) terms (see, e.g., [17–20]).

In [1], it is proposed to replace ζ_h given by (4) by

$$\zeta_h = \tau_1 \boldsymbol{u} \cdot \nabla\psi_h + \tau_2 \boldsymbol{u}_\| \cdot \nabla\psi_h , \tag{15}$$

where $\tau_1 = \tau$, $\tau_2 = \max(0, \tau_\| - \tau)$ (method DC2 in [1]) and $\tau_\|$ is computed as indicated in (5) but using $\boldsymbol{u}_\|$, the projection of $\boldsymbol{u}$ onto $\nabla\phi_h$. For $|\nabla\phi_h| \neq 0$, it is given by

$$u_{\|}=\frac{u\cdot\nabla\phi_h}{|\nabla\phi_h|^2}\,\nabla\phi_h\,. \tag{16}$$

Clearly, $u_{\|}\cdot\nabla\phi_h=u\cdot\nabla\phi_h$. The problem that may arise using this method becomes clear if we consider the effect of the second term in (15) when it is multiplied by $\mathcal{R}(\phi_h)$:

$$(\tau_2 u_{\|}\cdot\nabla\psi_h)\mathcal{R}(\phi_h)=\left[\tau_2\frac{u\cdot\nabla\phi_h}{|\nabla\phi_h|^2}\,\mathcal{R}(\phi_h)\right]\nabla\psi_h\cdot\nabla\phi_h\,. \tag{17}$$

The bracketed term acts as a diffusion. It may happen that $\operatorname{sgn}[(u\cdot\nabla\phi_h)\mathcal{R}(\phi_h)]=-1$ and thus a negative numerical diffusion may have been introduced. This problem is circumvented if the method proposed by Galeão and Dutra do Carmo [19] is employed. The basic idea of this method is to choose a vector u_r instead of $u_{\|}$ so that the upwind direction given by a linear combination of u and u_r is as close as possible to u, the difference $u-u_r$ satisfying the differential equation within each element. This yields

$$u_r:=\frac{\mathcal{R}(\phi_h)}{|\nabla\phi_h|^2}\,\nabla\phi_h\,. \tag{18}$$

Instead of (17), we now have

$$(\tau_2 u_r\cdot\nabla\psi_h)\mathcal{R}(\phi_h)=\tau_2\frac{\mathcal{R}(\phi_h)^2}{|\nabla\phi_h|^2}\,\nabla\psi_h\cdot\nabla\phi_h\,, \tag{19}$$

and it τ_2 is computed as indicated before, the isotropic diffusion introduced by this method is given by

$$k_{\text{iso}}=\frac{1}{2}\,\alpha_c h\,\frac{|\mathcal{R}(\phi_h)|}{|\nabla\phi_h|}\,, \tag{20}$$

with the function α_c calculated using the vector u_r.

Summarizing, this method consists of adding a diffusion proportional to the discrete residual of the differential equation within each element. This is also what Johnson et al. proposed in [15]. The same approach was used by Shakib [20] and now seems to be the canonical form of the shock-capturing diffusion to be introduced to preclude local oscillations (see, e.g., [21]). In the following, we question why this new diffusion should be isotropic.

3. The discrete maximum principle

3.1. Previous results

For the continuous steady-state problem, it is well known that the maximum principle holds (see, e.g., [22]), that is, the solution attains its maximum at the boundary when the source term f is non-positive. The question is whether this property is also inherited by the discrete problem in a sense to be explained later.

The discrete maximum principle (DMP, for short) has important consequences concerning the convergence properties of the numerical scheme and, in particular, in uniform convergence. If the components of u belong to $L^\infty(\Omega)$, f and the extension of the Dirichlet data g belong to the Sobolev space $W^{1,p}(\Omega)$, with $2\le N_{\text{sd}}<p$, then it is known (cf. [23]) that ϕ belongs to $L^\infty(\Omega)$. The same stability estimate is true for the finite element solution ϕ_h if the DMP holds. Moreover, for linear (simplicial) elements and using triangulations of strictly acute type (a concept to be defined in the following), Ciarlet and Raviart [23] proved that ϕ_h converges to ϕ in $L^\infty(\Omega)$. If it is further assumed that $\phi\in W^{2,p}(\Omega)$, then convergence is $\mathrm{O}(h)$. This estimate is not optimal, in the sense that one would hope

to obtain $O(h^2)$ for linear elements. For the particular case of self-adjoint problems, but also considering the effect of numerical integration, Wahlbin [24] proved an $O(h^{3-\varepsilon})$ estimate using quadratic elements, referring to the work of Nitsche [25] for an $O(h^{2-\varepsilon})$ estimate for linear elements. In both cases, ε is a positive number arbitrarily small.

The DMP is an important property of the numerical scheme, since it ensures monotonicity (for the steady-state problem) and that no spurious oscillations will appear, not even in the vicinity of sharp layers. Moreover, uniform convergence and pointwise stability estimates can be proved, as mentioned above.

In the next subsection, a sufficient condition is stated for an abstract discrete problem and applied to several particular cases using finite elements. Unfortunately, this condition is too restrictive to decide if the finite element method satisfies the DMP or not.

3.2. A sufficient condition for the discrete problem

Let N_{tp} be the total number of nodes of the finite element mesh and N_{fp} the number of interior nodes. The finite element discretization of the problem leads to an algebraic system of the form

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \,, \tag{21}$$

where $\boldsymbol{x}$ stands for the vector containing the nodal unknowns x_i, $i = 1, \ldots, N_{\mathrm{tp}}$. The values x_i, $i = N_{\mathrm{fp}} + 1, \ldots, N_{\mathrm{tp}}$ are known from the Dirichlet boundary conditions. Matrix $\boldsymbol{A}$, whose components are denoted a_{ij}, has dimensions $N_{\mathrm{fp}} \times N_{\mathrm{tp}}$ and the vector $\boldsymbol{b}$ coming from the source term has components b_i, $i = 1, \ldots, N_{\mathrm{fp}}$.

Our purpose now is to give a condition on the matrix $\boldsymbol{A}$ from which it will be possible to ensure that the DMP holds, viz.,

$$\max_{i=1,\ldots,N_{\mathrm{tp}}} \{x_i\} = x_m \,, \quad \text{with } N_{\mathrm{fp}} + 1 \leqslant m \leqslant N_{\mathrm{tp}} \,. \tag{22}$$

First, let us introduce the following definition [23, 26]: the matrix $\boldsymbol{A}$ is of non-negative type if the following conditions hold:

$$a_{ij} \leqslant 0 \,, \quad \text{for } i \neq j, \quad i = 1, \ldots, N_{\mathrm{fp}}, \quad j = 1, \ldots, N_{\mathrm{tp}} \,, \tag{23}$$

$$\sum_{j=1}^{N_{\mathrm{tp}}} a_{ij} \geqslant 0 \,, \quad i = 1, \ldots, N_{\mathrm{fp}} \,. \tag{24}$$

Let us call $\boldsymbol{A}_r = [a_{ij}]$, $i, j = 1, \ldots, N_{\mathrm{fp}}$. The sufficient condition mentioned above, which we prove for completeness, is the following (cf. [23, 26]).

THEOREM 1. Assume that $\boldsymbol{A}$ *is of nonnegative type,* $\boldsymbol{A}_r$ *is nonsingular and* $b_i \leqslant 0$, $i = 1, \ldots, N_{\mathrm{fp}}$. *Then the discrete maximum principle as expressed in (22) holds.*

PROOF. Let us write the jth equation of the linear system (21) as

$$a_{jj} x_j = b_j - \sum_{\substack{k=1 \\ k \neq j}}^{N_{\mathrm{tp}}} a_{jk} x_k \,. \tag{25}$$

From conditions (23) and (24), together with the fact that $\boldsymbol{A}_r$ is nonsingular, it follows that

$$a_{jj} > 0, \quad 1 + \sum_{\substack{k=1 \\ k \neq j}}^{N_{\mathrm{tp}}} \frac{a_{jk}}{a_{jj}} \geqslant 0 \,. \tag{26}$$

Since $b_j \leqslant 0$, $-a_{jk}/a_{jj} \geqslant 0$ for all $k \neq j$ and from (25) and (26), we obtain

$$\begin{aligned} x_j &= \frac{b_j}{a_{jj}} - \sum_{\substack{k=1\\k\neq j}}^{N_{\mathrm{tp}}} \frac{a_{jk}}{a_{jj}} x_k \leqslant \frac{b_j}{a_{jj}} - \max_{k\neq j}\{x_k\} \sum_{\substack{k=1\\k\neq j}}^{N_{\mathrm{tp}}} \frac{a_{jk}}{a_{jj}} \\ &\leqslant \frac{b_j}{a_{jj}} + \max_{k\neq j}\{x_k\} \leqslant \max_{k\neq j}\{x_k\} . \end{aligned} \tag{27}$$

Let us argue by contradiction and suppose that

$$\begin{aligned} &\max_{k\neq j}\{x_k\} = x_m , \quad \text{with } 1 \leqslant m \leqslant N_{\mathrm{fp}} , \\ &x_k < x_m , \quad k = N_{\mathrm{fp}} + 1, \dots , N_{\mathrm{tp}} . \end{aligned} \tag{28}$$

From (27), we have that $x_j \leqslant x_m$ and using the same argument as that to arrive at (27) for this m,

$$x_m \leqslant \max_{k\neq m}\{x_k\} = x_j ,$$

so that $x_m = x_j$. Without loss of generality, suppose that $m = 1$ and $j = 2$. Since we know that $x_1 = x_2$, we can eliminate the first equation in (21) and consider the system $\boldsymbol{A}'\boldsymbol{x}' = \boldsymbol{b}'$, with

$$\boldsymbol{x}' = (x_2, \dots , x_{N_{\mathrm{tp}}})^{\mathrm{t}} , \qquad \boldsymbol{b}' = (b_2, \dots , b_{N_{\mathrm{fp}}})^{\mathrm{t}} ,$$

$$a'_{i,1} = a_{i+1,1} + a_{i+1,2} , \quad i = 1, \dots , N_{\mathrm{fp}} - 1 ,$$

$$a'_{i,j} = a_{i+1,j+1} , \quad i = 1, \dots , N_{\mathrm{fp}} - 1, \ j = 2, \dots , N_{\mathrm{tp}} - 1 .$$

Since $\boldsymbol{A}_r$ is nonsingular, $\boldsymbol{A}'$ is of nonnegative type and (26) also holds for its components. Using the same argument as before, we find that $x_2 = x_j$ for some j, that we may take as $j = 3$. Repeating this process we arrive at the conclusion that $x_1 = x_2 = \cdots = x_{N_{\mathrm{fp}}}$ and in the last step, analogous to (27),

$$x_{N_{\mathrm{fp}}} \leqslant \max_{k = N_{\mathrm{fp}}+1,\dots,N_{\mathrm{tp}}} \{x_k\} ,$$

which is a contradiction with the second statement in (28). Therefore, this assumption (28) must be false, that is, condition (22) holds true. □

3.3. *Some particular cases*

We now check the hypothesis of Theorem 1 for three different particular discretizations using finite elements. For all these cases, it is assumed that the problem is well posed and thus the matrix $\boldsymbol{A}_r$ is nonsingular. The condition $b_i \leqslant 0$, $i = 1, \dots , N_{\mathrm{fp}}$, is very easy to verify and therefore our main concern is to check if the matrix A is of nonnegative type, that is, to verify conditions (23) and (24). This last condition is trivial to prove when standard finite elements are used. In fact, using the SUPG formulation and denoting by $\psi_{h,j}$ the shape function associated with the node j, we have that $\Sigma_{j=1}^{N_{\mathrm{tp}}} \psi_{h,j} \equiv 1$ and condition (24) holds with equal sign. So we only have to check condition (23). Moreover, since the assembly operator is linear, it has to be verified only for the element matrices and this is not difficult to do for some particular cases.

3.3.1. One-dimensional problem using linear elements

The first case we consider is the one-dimensional convection–diffusion equation using linear uniform elements of length h. For this case, it is possible to obtain explicit bounds for the upwind function α in order to satisfy the discrete maximum principle. The continuous problem is

$$\begin{aligned} & u\frac{\mathrm{d}\phi}{\mathrm{d}x} - k\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = f(x)\,, \quad 0 < x < l\,, \\ & \phi(0) = \phi_0\,, \qquad \phi(l) = \phi_l\,, \end{aligned} \tag{29}$$

with $f(x) \leqslant 0$. For this case, the signum of u is considered included in γ and also in α.

PROPOSITION 1. Assume that problem (29) is discretized using the Galerkin method and an artificial diffusion $k_a = \alpha hu/2$ is introduced. Then the numerical scheme satisfies the discrete maximum principle (DMP) iff the function α is such that

$$|\alpha| \geqslant 1 - \frac{1}{|\gamma|}\,. \tag{30}$$

If the SUPG method is employed and f is piecewise constant, then the DMP holds provided α verifies condition (30) and

$$|\alpha| \leqslant 1\,. \tag{31}$$

PROOF. The off-diagonal components of the element stiffness matrix $\boldsymbol{K}^e$ are

$$K_{12}^e = -\frac{k}{h}(1 + \alpha\gamma - \gamma)\,, \qquad K_{21}^e = -\frac{k}{h}(1 + \alpha\gamma + \gamma)\,,$$

using both the Galerkin method with artificial diffusion and the SUPG formulation. Requiring $K_{12}^e \leqslant 0$ and taking into account that $\mathrm{sgn}(\alpha) = \mathrm{sgn}(\gamma)$ leads to

$$\alpha\gamma = |\alpha||\gamma| \geqslant \gamma - 1\,, \qquad |\alpha| \geqslant \mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}\,.$$

Condition $K_{21}^e \leqslant 0$ yields

$$\alpha\gamma = |\alpha||\gamma| \geqslant -\gamma - 1\,, \qquad |\alpha| \geqslant -\mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}\,.$$

Both $K_{12}^e \leqslant 0$ and $K_{21}^e \leqslant 0$ iff

$$|\alpha| \geqslant \max\left\{\mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}, -\mathrm{sgn}(\gamma) - \frac{1}{|\gamma|}\right\} = 1 - \frac{1}{|\gamma|}\,.$$

Let us now check that $b_i \leqslant 0$. For the Galerkin method with artificial diffusion this is obvious, since $\psi_{h,i} \geqslant 0$ for all i. Thus, the DMP holds for this case. If the SUPG formulation is employed and f is piecewise constant, with value f_i in the element $[(i-1)h, ih]$, we have

$$b_i = \int_0^l \left(\psi_{h,i} + \frac{1}{2}\alpha h \frac{\mathrm{d}\psi_{h,i}}{\mathrm{d}x}\right) f \,\mathrm{d}x$$
$$= f_i \int_{(i-1)h}^{ih} (\psi_{h,i} + \tfrac{1}{2}\alpha)\,\mathrm{d}x + f_{i+1} \int_{ih}^{(i-1)h} (\psi_{h,i} - \tfrac{1}{2}\alpha)\,\mathrm{d}x\,, \tag{32}$$

for $i = 1, \ldots, N_{\mathrm{el}} - 1$. For arbitrary values of f_i and f_{i+1} both integrals must be nonnegative in order to ensure that $b_i \leqslant 0$. Since their values are $h/2 + \alpha h/2$ and $h/2 - \alpha h/2$ this will only happen if condition (31) holds. □

REMARK 1. In [8] it is proved that the SUPG method using the optimal upwind function must give nodally exact results when f is piecewise constant, but not for arbitrary source functions f. Now we see that in the general case, it is not even possible to satisfy the DMP, since from (32) it is observed that it is possible to choose $f \leqslant 0$ such that $b_i > 0$ for any $\alpha > 0$. When f is piecewise linear, it is easy to show that $|\alpha| \leqslant 2/3$ is needed in order to have $b_i \leqslant 0$, but if this condition is fulfilled, it is possible to violate (30) for certain values of the Péclet number γ.

REMARK 2. Condition (30) can be found using very different arguments. It can be shown [27] that it is the condition under which no oscillations appear in the numerical solution. In [9] it is proved that it is the condition that ensures that the diffusive stability limit of the forward Euler scheme in time dominates the convective limit.

3.3.2. Multidimensional problem using simplicial linear elements

Suppose now that the domain Ω is discretized using linear N_{sd}-simplices. Under a certain condition on the finite element partition $\{\Omega^e\}$, it is also possible in this case to obtain bounds for the upwind function in order to satisfy the DMP. The results presented here are an extension of the work of Kikuchi [26].

For each element e, let us introduce the following notation: ρ^e is the supremum of the diameter of the spheres inscribed in Ω^e, κ^e is the minimum perpendicular length of Ω^e, λ^e is the maximum perpendicular length of Ω^e, h^e is the diameter of Ω^e and $h = \max_e\{h^e\}$.

As usual, the finite partition is assumed to be regular [28]. This means that there exists a positive constant C_1 such that

$$\min_e \frac{\rho^e}{h^e} \geqslant C_1 \quad \text{as } h \to 0\,, \tag{33}$$

and in particular this implies that there exists another positive constant C_2 such that

$$\min_e \frac{\kappa^e}{\lambda^e} \geqslant C_2 \quad \text{as } h \to 0\,. \tag{34}$$

Let us also introduce the constant

$$\sigma^e := \max_{i \neq j} \frac{\nabla\psi^e_{h,i} \cdot \nabla\psi^e_{h,j}}{|\nabla\psi^e_{h,i}|\,|\nabla\psi^e_{h,j}|} = \max_{i \neq j} \cos(\nabla\psi^e_{h,i}, \nabla\psi^e_{h,j})\,. \tag{35}$$

The main restriction of what follows is that we assume the finite element partition $\{\Omega^e\}$ to be of strictly acute type. By definition (cf. [23, 26]), this means that there exists a constant $\sigma_0 > 0$ such that

$$\sigma^e \leqslant -\sigma_0\,, \quad e = 1, \ldots, N_{\mathrm{el}}\,. \tag{36}$$

For two-dimensional problems ($N_{\mathrm{sd}} = 2$) this happens only if all the angles of the triangles are $<\pi/2$. Observe that $\sigma^e \geqslant -1$ and therefore $\sigma_0 \leqslant 1$.

To prove the following result, we make a simplification. The velocity $\boldsymbol{u}$ is considered constant within each element. Since $\nabla\phi_h$ is piecewise constant, the vector $\boldsymbol{u}_\|$ defined in (16) is also piecewise constant. The Péclet number computed with this vector is denoted by $\gamma_\|$.

PROPOSITION 2. Under the assumptions stated above, assume that the steady-state convection–diffusion equation, with $f \leqslant 0$, is discretized using the Galerkin method and an isotropic artificial diffusion,

$$k_a^e = \frac{1}{2}\alpha^e h^e |\boldsymbol{u}_\|^e| , \tag{37}$$

is introduced within each element. Then the numerical scheme satisfies the DMP if the function α^e is such that

$$\alpha^e \geqslant \frac{C}{N_{\text{sd}}+1} - \frac{1}{\gamma_\|^e} , \tag{38}$$

for a certain constant C.

PROOF. Since $\psi_{h,i} \geqslant 0$ for each node i, it is clear that $b_i = \int_\Omega \psi_{h,i} f \, \mathrm{d}\Omega \leqslant 0$. As before, since the assembly operator is linear, it suffices to prove condition (23) for each element stiffness matrix. Let us denote one of them by $\boldsymbol{K}^e$ and let $\boldsymbol{K}_\|^e$ be the matrix calculated with $\boldsymbol{u}_\|$. We have that

$$\boldsymbol{K}^e\boldsymbol{\phi}^e = \boldsymbol{K}_\|^e\boldsymbol{\phi}^e ,$$

and therefore it is enough to check (23) for $\boldsymbol{K}_\|^e$. Observe first that [23]

$$(\lambda^e)^{-1} \leqslant |\nabla\psi_{h,i}^e| \leqslant (\kappa^e)^{-1} , \tag{39}$$

$$\int_{\Omega^e} \psi_{h,i}^e \, \mathrm{d}\Omega = \frac{\text{meas}(\Omega^e)}{N_{\text{sd}}+1} , \tag{40}$$

for all $i = 1, \ldots, N_{\text{sd}}$.

The ij component of the matrix $\boldsymbol{K}_\|^e$, for $i \neq j$, can be bounded as follows:

$$\begin{aligned}
(k_\|^e)_{ij} &= (k + k_a^e)\int_{\Omega^e} \nabla\psi_{h,i}^e \cdot \nabla\psi_{h,j}^e \, \mathrm{d}\Omega + \int_{\Omega^e} \psi_{h,i}^e \boldsymbol{u}_\|^e \cdot \nabla\psi_{h,j}^e \, \mathrm{d}\Omega \\
&\leqslant (k + k_a^e)\,\text{meas}(\Omega^e)\sigma^e |\nabla\psi_{h,i}^e||\nabla\psi_{h,j}^e| + |\boldsymbol{u}_\|^e||\nabla\psi_{h,j}^e| \frac{\text{meas}(\Omega^e)}{N_{\text{sd}}+1} \\
&\leqslant (k + k_a^e)\,\text{meas}(\Omega^e)\sigma^e \frac{1}{(\lambda^e)^2} + |\boldsymbol{u}_\|^e| \frac{\text{meas}(\Omega^e)}{\kappa^e(N_{\text{sd}}+1)} \\
&= (\lambda^e)^{-2}\,\text{meas}(\Omega^e)\left[\frac{(\lambda^e)^2}{\kappa^e}\frac{|\boldsymbol{u}_\|^e|}{N_{\text{sd}}+1} + \left(k + \frac{1}{2}\alpha^e h^e |\boldsymbol{u}_\|^e|\right)\sigma^e\right] ,
\end{aligned}$$

where we have used (39) and (40). Taking into account that

$$\frac{(\lambda^e)^2}{\kappa^e} \leqslant \frac{\lambda^e}{\kappa^e} h^e \leqslant \frac{h^e}{C_2} ,$$

it follows that

$$(k_{\|}^{e})_{ij} \leqslant \frac{1}{2}(\lambda^{e})^{-2}\,\mathrm{meas}(\Omega^{e})|\boldsymbol{u}_{\|}^{e}|h^{e}\left[\frac{2}{C_2(N_{\mathrm{sd}}+1)}+\left(\frac{2k}{h^{e}|\boldsymbol{u}_{\|}^{e}|}+\alpha^{e}\right)\sigma^{e}\right].$$

Therefore, the condition $(k_{\|}^{e})_{ij} \leqslant 0$ holds if

$$\frac{2}{C_2(N_{\mathrm{sd}}+1)}+\left(\frac{1}{\gamma_{\|}^{e}}+\alpha^{e}\right)\sigma^{e} \leqslant 0\,.$$

Since $\sigma^{e}<0$, this is equivalent to

$$\alpha^{e} \geqslant \frac{(-2/C_2\sigma^{e})}{N_{\mathrm{sd}}+1}-\frac{1}{\gamma_{\|}^{e}}\,.$$

The proposition follows for $C=-2/C_2\sigma^{e}$. □

REMARK 3. As for the one-dimensional problem, an upper bound for the upwind function α^{e} is needed when the SUPG formulation is used in order to have $b_i \leqslant 0$ for the case of a piecewise constant source f, that is, $\alpha^{e} \leqslant C'$ for a certain constant C'. Let us prove this for each elemental contribution to the vector $\boldsymbol{b}$. If f^{e} is the value of f in element e,

$$\begin{aligned} b_i^{e} &= f^{e}\int_{\Omega^{e}}\left(\psi_{h,i}^{e}+\frac{\alpha^{e}h^{e}}{2|\boldsymbol{u}^{e}|}\boldsymbol{u}^{e}\cdot\nabla\psi_{h,i}^{e}\right)\mathrm{d}\Omega \\ &= f^{e}\left[\frac{\mathrm{meas}(\Omega^{e})}{N_{\mathrm{sd}}+1}+\frac{\alpha^{e}h^{e}}{2|\boldsymbol{u}^{e}|}\boldsymbol{u}^{e}\cdot\nabla\psi_{h,i}^{e}\,\mathrm{meas}(\Omega^{e})\right]. \end{aligned} \tag{41}$$

On the other hand, using the fact that $\lambda^{e} \geqslant \rho^{e}$, we obtain

$$\begin{aligned} \frac{1}{N_{\mathrm{sd}}+1}+\frac{\alpha^{e}h^{e}}{2|\boldsymbol{u}^{e}|}\boldsymbol{u}^{e}\cdot\nabla\psi_{h,i}^{e} &\geqslant \frac{1}{N_{\mathrm{sd}}+1}-\frac{\alpha^{e}h^{e}}{2}|\nabla\psi_{h,i}^{e}| \geqslant \frac{1}{N_{\mathrm{sd}}+1}-\frac{\alpha^{e}h^{e}}{2\kappa^{e}} \\ &\geqslant \frac{1}{N_{\mathrm{sd}}+1}-\frac{\alpha^{e}h^{e}}{2\kappa^{e}}\frac{\lambda^{e}}{\rho^{e}} \geqslant \frac{1}{N_{\mathrm{sd}}+1}-\frac{\alpha^{e}}{2C_1C_2}\,, \end{aligned}$$

and from (41) it follows that $b_i^{e} \leqslant 0$ provided that

$$\alpha^{e} \leqslant \frac{2C_1C_2}{N_{\mathrm{sd}}+1}\,.$$

But even though $b_i^{e} \leqslant 0$ the DMP may not hold since the SUPG formulation only introduces streamline diffusion and thus Proposition 2 does not guarantee the satisfaction of the DMP for this case. This argument is the basic idea of the method introduced in the next section.

3.3.3. *A two-dimensional problem using bilinear elements*

Now we shall analyse a very simple two-dimensional example using bilinear finite elements. Let the domain Ω be discretized using equal elements of length h_x in the x-direction and length h_y in the y-direction, and let $\boldsymbol{u}=(u,0)$, with $u \geqslant 0$.

Our purpose in analyzing this case is again to gain insight into the role played by the upwind function in the satisfaction of the DMP. For that, we assume that numerical diffusion is added both in the x- and the y-directions, so that the final diffusion to be considered in each direction is

$$k_x = k+\tfrac{1}{2}\alpha_x u h_x\,, \qquad k_y = k+\tfrac{1}{2}\alpha_y u h_y\,. \tag{42}$$

We want to study under which conditions on the functions α_x and α_y and also on the ratio $\eta := h_y/h_x$ it is possible to satisfy the DMP. As in the previous cases, this reduces to checking condition (23) for the element stiffness matrix $\boldsymbol{K}^e$, that shall be split into the diffusive contribution $\boldsymbol{K}^e_d$ and the convective contribution $\boldsymbol{K}^e_c$, with components

$$(k^e_d)_{ij} = \int_{\Omega^e} \left(k_x \frac{\partial \psi^e_{h,i}}{\partial x} \frac{\partial \psi^e_{h,j}}{\partial x} + k_y \frac{\partial \psi^e_{h,i}}{\partial y} \frac{\partial \psi^e_{h,i}}{\partial y} \right) \mathrm{d}x \, \mathrm{d}y \, , \qquad (k^e_c)_{ij} = \int_{\Omega^e} u \psi_{h,i} \frac{\partial \psi^e_{h,j}}{\partial x} \, \mathrm{d}x \, \mathrm{d}y \, .$$

Working out the explicit expression of $(k^e_d)_{ij}$ and $(k^e_c)_{ij}$, it is found that

$$\boldsymbol{K}^e_d = \begin{bmatrix} \frac{1}{3}\left(\eta k_x + \frac{1}{\eta} k_y\right) & \frac{1}{6}\left(-2\eta k_x + \frac{1}{\eta} k_y\right) & \frac{1}{6}\left(-\eta k_x - \frac{1}{\eta} k_y\right) & \frac{1}{6}\left(\eta k_x - \frac{2}{\eta} k_y\right) \\ & \frac{1}{3}\left(\eta k_x + \frac{1}{\eta} k_y\right) & \frac{1}{6}\left(\eta k_x - \frac{2}{\eta} k_y\right) & \frac{1}{6}\left(-\eta k_x - \frac{1}{\eta} k_y\right) \\ & & \frac{1}{3}\left(\eta k_x + \frac{1}{\eta} k_y\right) & \frac{1}{6}\left(-2\eta k_x + \frac{1}{\eta} k_y\right) \\ \text{Symm} & & & \frac{1}{3}\left(\eta k_x + \frac{1}{\eta} k_y\right) \end{bmatrix} ,$$

$$\boldsymbol{K}^e_c = \frac{u h_y}{12} \begin{bmatrix} -2 & 2 & 1 & -1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ -1 & 1 & 2 & -2 \end{bmatrix} .$$

Let us first consider $u = 0$ and thus $\alpha_x = \alpha_y = 0$, $k_x = k_y = k$. Requiring condition (23) to hold yields $\sqrt{2}/2 \leqslant \eta \leqslant \sqrt{2}$. Hence, even without convection the elements cannot be too elongated to ensure that the DMP will hold.

Suppose now that $\eta = 1$. It is easy to show that (23) yields

$$\alpha_x - \frac{1}{2}\alpha_y \geqslant 1 - \frac{1}{2\gamma} \, , \qquad \alpha_x + \alpha_y \geqslant 1 - \frac{2}{\gamma} \, , \qquad \alpha_x - 2\alpha_y \leqslant -1 + \frac{1}{\gamma} \, . \tag{43}$$

Consider the case $\alpha_y = 0$. The last condition in (43) leads to $\alpha_x \leqslant -1 + 1/\gamma$, which is incompatible with the first, $\alpha_x \geqslant 1 - 1/2\gamma$, for high values of the Péclet number. Therefore, it is impossible to satisfy condition (23) adding only streamline diffusion.

Suppose now that $\alpha_x = \alpha_y = \alpha$. From (43) it is found that $\alpha \geqslant 2 - 1/\gamma$ is needed if condition (23) is to be verified. Again we observe, as in the previous cases, that the form of the upwind function must be $C - 1/\gamma$ for a certain constant C.

3.4. Discussion

Several conclusions may be drawn from the application of Theorem 1 to the cases considered above. The first is that this result is limited, in the sense that it does not apply to many cases of interest. Maybe the most obvious one is the one-dimensional problem using quadratic elements, for which it is proved in [8] that it is possible to obtain nodally exact solutions for piecewise linear source functions, and therefore (22) must be true (although if ϕ_h is piecewise quadratic it may have local extrema between two nodes). However, the stiffness matrix for this case is not of nonnegative type. Nevertheless, for linear elements we have obtained bounds for the upwind function (conditions (30) and (31)) that ratify what its behavior must be in terms of γ predicted by convergence analysis [1, 16].

The next two cases (linear simplicial elements and a simple problem using bilinear elements) provide several conclusions. The first is that the stiffness matrix will not be of nonnegative type if the mesh is distorted, in particular, if the triangulation is not of strictly acute type when simplices are used and if the aspect ratio is large using bilinear elements. Also, we have observed that the streamline diffusion is not enough to end up with a matrix of nonnegative type, but an addition of crosswind diffusion is also needed. This is perhaps the most salient result. When an isotropic artificial diffusion $k_a = \alpha h|\boldsymbol{u}|/2$ is introduced, α must be greater than $C - 1/\gamma_{\|}$ for a certain constant C. Upper bounds for α have been obtained when the SUPG method is employed.

4. A discontinuity-capturing crosswind-dissipation

From the discussion of the previous section, it is clear that the streamline diffusion introduced by the SUPG formulation is not enough to obtain a monotone scheme and an additional crosswind diffusion is needed. The method discussed in Section 2 introduces this new dissipation but also modifies the streamline diffusion. From the expression of the upwind function that we use, this streamline diffusion satisfies all the general requirements we have found in Section 3 for some particular cases (up to the choice of the algorithmic constants). The main idea of the method to be introduced here is to keep unaltered the diffusion in the direction of the streamlines and to modify only the crosswind diffusion.

The new crosswind dissipation (CD, from now onwards) must satisfy two conditions. To avoid excessive overdamping, it must be small in regions where convective effects are not very important, that is, where $|\boldsymbol{u} \cdot \nabla\phi_h|$ is small. For consistency, it must be proportional to the element residual defined in (9). Guided by the results of the previous section, the magnitude of the CD could be taken within each element as

$$k_c^e = \tfrac{1}{2}\alpha_c^e h^e \frac{|\mathcal{R}(\phi_h)|}{|\nabla\phi_h|} \tag{44}$$

when $|\nabla\phi_h| \neq 0$ and zero otherwise. The function α_c^e is taken as

$$\alpha_c^e = \max\{0, C - 1/\gamma_{\|}^e\} . \tag{45}$$

This ensures that $k_c^e = 0$ when $|\boldsymbol{u}^e \cdot \nabla\phi_h|$ is small. The question that remains is how to choose the constant C. From 2D numerical experiments we have found that $C \approx 0.7$ for linear and bilinear elements, and $C \approx 0.35$ for quadratic and biquadratic elements are effective. For the first case, it is observed that this corresponds to (38) with the constant in this expression equal to 2.

Having defined the magnitude of the CD by (44) and the function α_c^e by (45), the description of the method is now complete. In order to introduce the CD, a term

$$\sum_{e=1}^{N_{\text{el}}} \int_{\Omega^e} \frac{1}{2}\alpha_c^e h^e \frac{|\mathcal{R}(\phi_h)|}{|\nabla\phi_h|} \nabla\psi_h \cdot \left(\boldsymbol{I} - \frac{1}{|\boldsymbol{u}|^2}\boldsymbol{u}\otimes\boldsymbol{u}\right) \cdot \nabla\phi_h \, \mathrm{d}\Omega \tag{46}$$

must be added to the left-hand side of (2), $\boldsymbol{I}$ being the unit tensor.

From the expressions of the τ^e in terms of the upwind function given by (7) and from (45) it follows that the CD will always be smaller than the streamline diffusion introduced by the SUPG formulation. The total diffusion ellipsoid in 2D is schematically represented in Fig. 1.

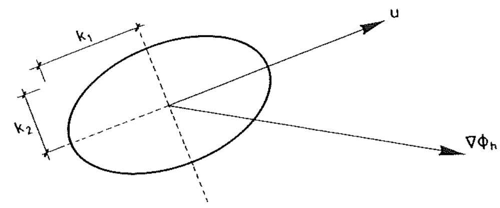

Fig. 1. Total diffusion ellipsoid. The values of k_1 and k_2 are $k_1 = k + \alpha h|\boldsymbol{u}|/2$, $k_2 = k + \alpha_c h|\mathcal{R}(\phi_h)|/2|\nabla\phi_h|$.

Let us close this section mentioning some results obtained by Johnson et al. in [29] and slightly improved by Niijima [30]. They analyzed a 2D model problem using linear elements and the SUPG formulation and introducing a constant crosswind diffusion $k_c = \max(k, h^{3/2})$, k being the physical diffusion. For this choice of k_c, the global L^2-estimates of the SUPG method do not deteriorate, since they are $O(h^{3/2})$. Suppose that $k \leq h^{3/2}$. Under certain regularity assumptions on the data, the main results of the quoted references are

$$\begin{array}{rll} & \text{Ref. [29]} & \text{Ref. [30]} \\ |(\phi - \phi_h)(x_0, y_0)| = & O(h^{5/4} \log^{3/2}(1/h)) & O(h^{11/8} \log(1/h))\,, \\ \|\phi_h\|_{L^\infty(\Omega)} = & O(h^{-1/4} \log^{3/2}(1/h)) & O(h^{-1/8} \log(1/h))\,, \\ \|\phi - \phi_h\|_{L^1(\Omega)} = & O(h^{1/2} \log^{5/2}(1/h)) & O(h^{5/8} \log^2(1/h))\,, \end{array}$$

where (x_0, y_0) is a point in Ω. All these theoretical results seem to confirm that the introduction of a crosswind diffusion might improve the numerical solution. This idea is further strengthened by the numerical experiments presented next.

5. Numerical examples

We have chosen two simple test problems to compare the accuracy and the numerical efficiency of the original SUPG method and SUPG plus the introduction of a shock-capturing dissipation, either isotropic (labelled ID in what follows) given by (20) or only as a crosswind diffusion (labelled CD). The transient relaxation described in Section 2.2 is used to obtain the steady-state solution.

The first example is the classical problem of the propagation of a discontinuity in the unit square. The Dirichlet prescription g is 1 in the upper boundary $y = 1$ and in part of the lateral boundary $x = 0$, and zero on the rest of the boundary. The diffusion coefficient is taken as $k = 10^{-8}$ and the velocity is constant, with modulus 1 and parallel to $(1, -2)$. We have solved this problem using a mesh of 20×20 Q_1 (bilinear) elements and an unstructured mesh of 200 P_2 (quadratic triangles) elements, with 437 nodal points. The triangles of this last mesh have also been split into four P_1 (linear) subtriangles.

Results using the P_1 element are shown in Fig. 2. It is observed that the ID and the CD methods yield very similar answers, and they indeed remove the oscillations appearing using the SUPG method. This also happens using the Q_1 and the P_2 elements (results not shown). The important point is the convergence shown in Figs. 3–5. It is observed that the convergence rate of the CD method is very similar to that of the original SUPG formulation, whereas it is substantially deteriorated for the ID method.

Let us discuss the numerical cost of the calculations referring to the different shock-capturing techniques and the different elements employed. Consider first the bilinear element, for which the 2×2 Gauss–Legendre integration rule has been employed. The total CPU of the computation using the SUPG, the ID and the CD methods, as well as the CPU time per iteration needed on a CONVEX-C3 computer are given in Table 1.

It is observed that the CD method needs an amount of computer time per iteration similar to the ID method (even smaller in this case). For the linear triangle using a three-point integration rule, the results are as shown in Table 2. Table 3 gives the results for the quadratic triangle using a four-point integration rule.

Table 1

Method	CPU (s)	CPU per iteration ($\times 10^{-3}$)
SUPG	6.74	79.29
ID	18.99	94.95
CD	7.60	89.41

Table 2

Method	CPU (s)	CPU per iteration ($\times 10^{-3}$)
SUPG	13.15	125.40
ID	44.29	147.63
CD	17.18	149.39

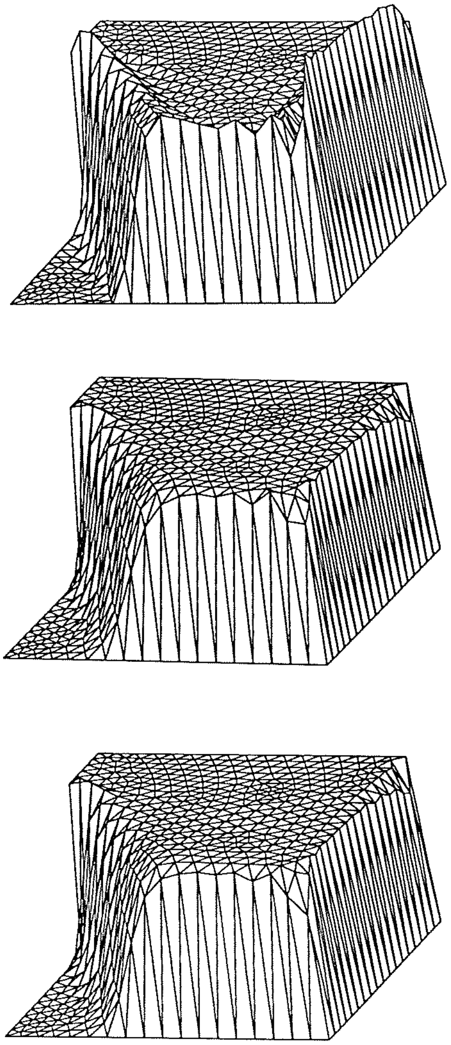

Fig. 2. Results using the P_1 element for the propagation of a discontinuity. From the top to the bottom: SUPG, ID and CD methods.

Table 3

Method	CPU (s)	CPU per iteration ($\times 10^{-3}$)
SUPG	9.04	37.67
ID	12.70	42.33
CD	8.27	41.35

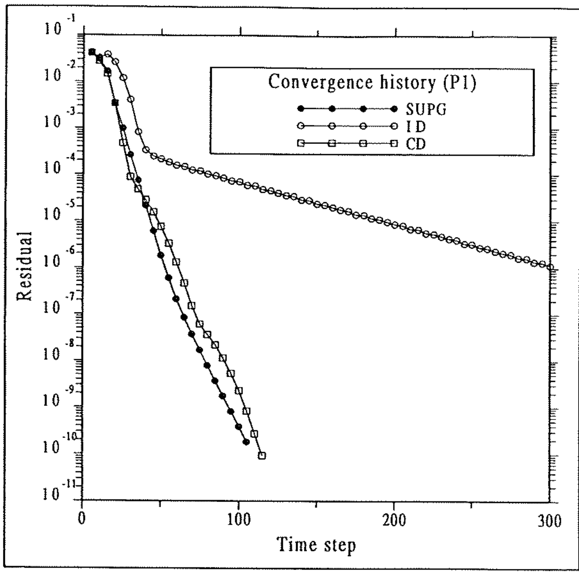

Fig. 3. Convergence history for the P_1 element for the propagation of a discontinuity.

Concerning the element employed in the calculation, the smallest CPU time per iteration is needed using the quadratic triangle, whereas the largest is needed using the linear triangle. This is due to the total number of integration points of the finite element mesh, since using an explicit scheme to advance in time the cost of updating the unknowns depends on the number of nodal points, not on the type of element. It is also observed that the total CPU time is similar (although depending on the method employed) using the bilinear element and the quadratic one, even though many more time steps are needed to reach the steady state for the latter.

As a second example, we consider the problem with data $\Omega = (0,1) \times (0,1)$, $\boldsymbol{u} = (0,1)$, $k = 10^{-8}$, $f = 1$, $g = 0$. The computational domain is discretized using a uniform mesh of 20×20 bilinear elements. Results are shown in Fig. 6. It is observed that the CD method is much less overdiffusive than the ID method. In both cases, the oscillations of the SUPG formulation are removed. From Fig. 7, it is also observed that the convergence history of the CD method is similar to that of the SUPG method, and much better than using the ID approach.

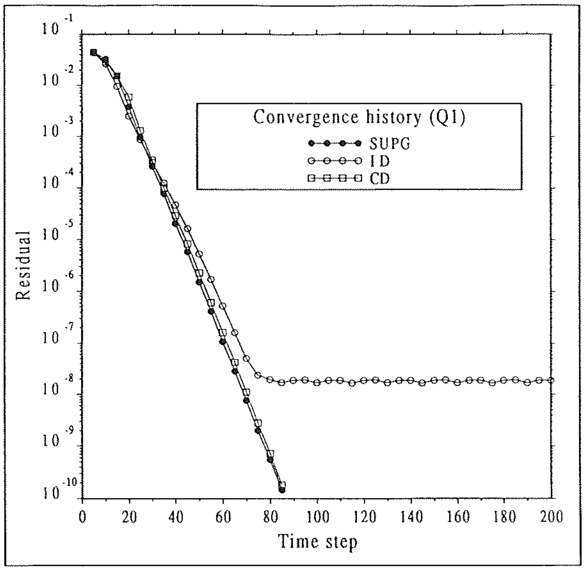

Fig. 4. Convergence history for the Q_1 element for the propagation of a discontinuity.

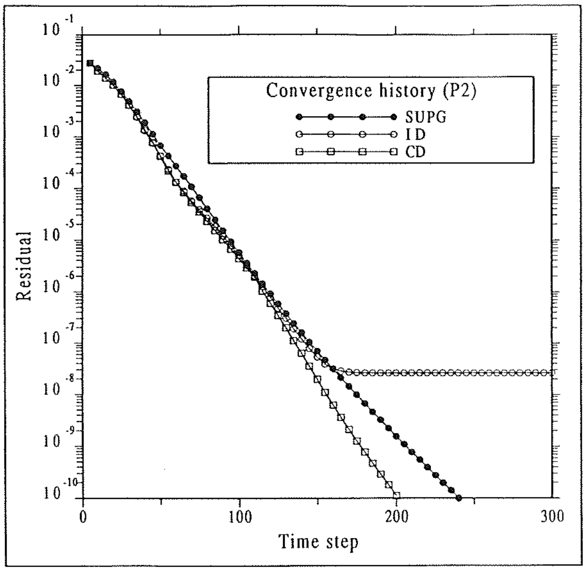

Fig. 5. Convergence history for the P_2 element for the propagation of a discontinuity.

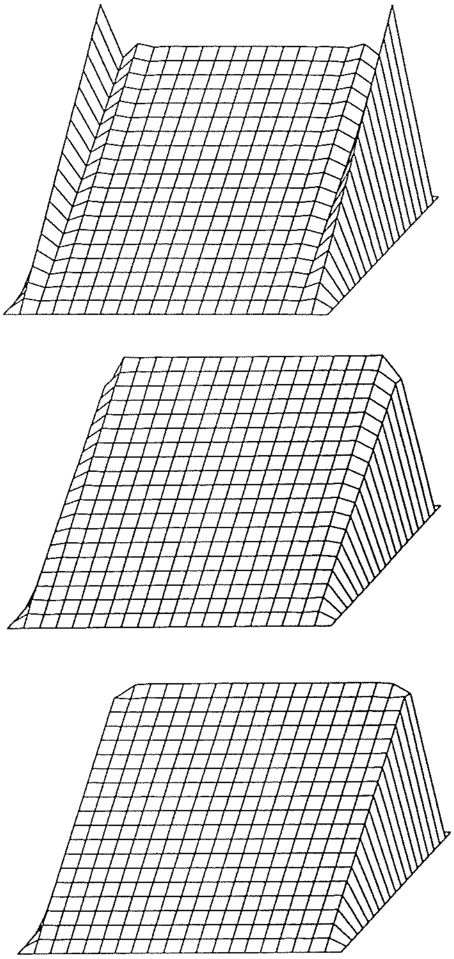

Fig. 6. Results for the problem with a source term. From the top to the bottom: SUPG, ID and CD methods.

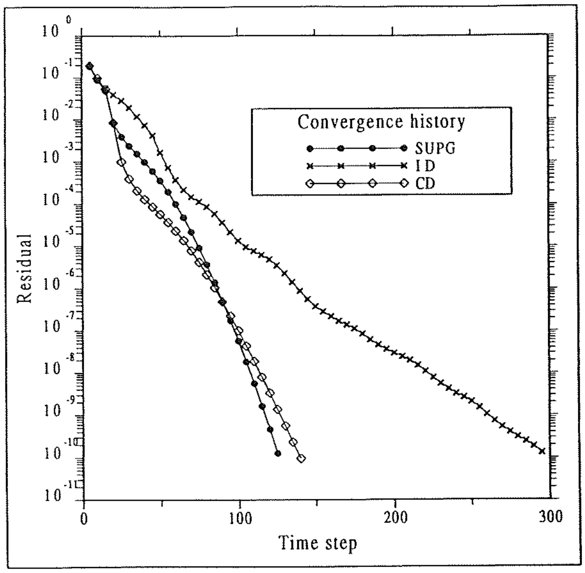

Fig. 7. Convergence history for the problem with a source term.

6. Conclusions

The aim of this paper has been to discuss which is the appropriate shock-capturing modification of the basic SUPG formulation. A common approach is to introduce an isotropic diffusion proportional to the residual within each element. The crucial question is why should this new dissipation be isotropic if the streamline diffusion introduced by the SUPG method seems to be enough along the streamlines. This last point is confirmed not only by numerical experiments, but also by the study of the discrete maximum principle in some simple cases.

Having in mind the idea that only a modification of the crosswind diffusion is necessary, the discrete maximum principle provides the theoretical grounds for the design of the new dissipation. Assuming first that an isotropic diffusion is added to the Galerkin formulation, for two particular cases, it has been shown that this dissipation can be taken as indicated by (44), with the function α_c given by (45). The bound $\alpha_c \geqslant C - 1/\gamma_{\|}$ is the sharpest we have been able to obtain. Observe that $\gamma_{\|}$ is the smallest of the possible pseudo-Péclet numbers that can be computed with vectors $\boldsymbol{v}$ such that $\boldsymbol{v} \cdot \nabla\phi_h = \boldsymbol{u} \cdot \nabla\phi_h$. The question that remains open is, as for most numerical methods, the election of the algorithmic constants. Our choice has been based on numerical experimentation. Also, it has been noticed that the upwind function of the SUPG method may not be arbitrarily large, in contrast with what happens using a purely artificial dissipation.

Concerning the practical behavior of this new approach, from the numerical experiments presented in the last section, several conclusions may be drawn. We have seen that this new method, compared to the introduction of an isotropic diffusion, has a much better convergence rate towards the steady-state when a transient relaxation is used to solve the nonlinear discrete problem, is less overdiffusive and has a similar computational cost. Therefore, we believe that its numerical performance makes it an attractive shock-capturing technique.

References

[1] T.J.R. Hughes, M. Mallet and A. Mizukami, A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, Comput. Methods Appl. Mech. Engrg. 54 (1986) 341–355.

[2] C. Johnson, U. Nävert and J. Pitkäranta, Finite element methods for linear hyperbolic equations, Comput. Methods Appl. Mech. Engrg. 45 (1984) 285–312.

[3] U. Nävert, A finite element method for convection–diffusion problems, Thesis, Chalmers University of Technology, Göteborg, Sweden, 1982.

[4] R.J. LeVeque. Numerical Methods for Conservation Laws (Birkhäuser, Boston, 1990).

[5] C. Hirsch, Numerical Computation of Internal and External Flows, Vols. 1 and 2 (Wiley, New York, 1990).

[6] E.S. Oran and J.P. Boris, Numerical Simulation of Reactive Flow (Elsevier, Amsterdam, 1987).

[7] T. Ikeda, Maximum Principle in Finite Element Models for Convection–Diffusion phenomena (North-Holland/Kinokuniya, Amsterdam, 1983).

[8] R. Codina, E. Oñate and M. Cervera, The intrinsic time for the streamline upwind/Petrov–Galerkin formulation using quadratic elements, Comput. Methods Appl. Mech. Engrg. 94 (1992) 239–262.

[9] R. Codina, Stability analysis of the forward Euler scheme for the convection–diffusion equation using the SUPG formulation in space, Internat. J. Numer. Methods Engrg. 36 (1993) 1445–1464.

[10] A. Mizukami and T.J.R. Hughes, A Petrov–Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle, Comput. Methods Appl. Mech. Engrg. 50 (1985) 181–193.

[11] J.G. Rice and R.J. Schnipke, A monotone streamline upwind finite element method for convection-dominated flows, Comput. Methods Appl. Mech. Engrg. 48 (1985) 313–327.

[12] S.F. Davis, A rotationally biased upwind difference scheme for the Euler equations, J. Comput. Phys. 39 (1981) 164–178.

[13] T.J.R. Hughes and M. Mallet, A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems, Comput. Methods Appl. Mech. Engrg. 58 (1986) 329–336.

[14] C. Johnson and A. Szepessy, On the convergence of a finite element method for a nonlinear hyperbolic conservation law, Math. Comput. 49 (1987) 427–444.

[15] C. Johnson, A. Szepessy and P. Hansbo, On the convergence of shock-capturing streamline finite element methods for hyperbolic conservation laws, Technical report 1987-21, Mathematics Department, Chalmers University of Technology, Göteborg, 1987.

[16] T.J.R. Hughes, L.P. Franca and G.M. Hulbert, A new finite element formulation for computational fluid dynamics: VIII.

The Galerkin/least-squares method for advective-diffusive equations, Comput. Methods Appl. Mech. Engrg. 73 (1989) 173–189.

[17] T.E. Tezduyar and Y.J. Park, Discontinuity-capturing finite element formulations for nonlinear convection-diffusion-reaction equations, Comput. Methods Appl. Mech. Engrg. 59 (1986) 307–325.

[18] E.G. Dutra do Carmo and A.C. Galeão, Feedback Petrov–Galerkin methods for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 88 (1991) 1–16.

[19] A.C. Galeão and E.G. Dutra do Carmo, A consistent approximate upwind Petrov–Galerkin method for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 68 (1988) 83–95.

[20] F. Shakib, Finite element analysis of the compressible Euler and Navier–Stokes equations, Ph.D. Thesis, Stanford University, 1988.

[21] C. Johnson, A new approach to algorithms for convection problems which are based on exact transport + projection, Comput. Methods Appl. Mech. Engrg. 100 (1992) 45–62.

[22] R. Courant and D. Hilbert, Methods of Mathematical Physics, Vol. 2 (Wiley/Interscience, New York, 1962).

[23] P.G. Ciarlet and P.-A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput. Methods Appl. Mech. Engrg. 2 (1973) 17–31.

[24] L.B. Wahlbin, Maximum norm error estimates in the finite element method with isoparametric quadratic elements and numerical analysis, RAIRO Anal. Numer. 12 (1978) 173–262.

[25] J.A. Nitsche, L^∞-convergence for finite element approximations, 2, Conference on Finite Elements, Rennes, France, 1975.

[26] F. Kikuchi, Discrete maximum principle and artificial viscosity in finite element approximations of convective diffusion equations, ISAS Report No. 550, Vol. 42, No. 5, Tokyo, 1977.

[27] I. Christie, D.F. Griffiths, A.R. Mitchell and O.C. Zienkiewicz, Finite element methods for second order differential equations with significant first derivatives, Internat. J. Numer. Methods Engrg. 10 (1976) 1389–1396.

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

[29] C. Johnson, A.H. Schatz and L.B. Wahlbin, Crosswind smear and pointwise errors in streamline diffusion finite element methods, Math. Comp. 49 (1987) 25–38.

[30] K. Niijima, Pointwise error estimates for a streamline-diffusion finite element scheme, Numer. Math. 56 (1990) 707–719.