ACADÉMIE
DES SCIENCES
INSTITUT DE FRANCE

Comptes Rendus
Mathématique

Frédéric Hecht and Olivier Pironneau

The Dual Characteristic-Galerkin Method

Volume 362 (2024), p. 1109-1119

Online since: 5 November 2024

https://doi.org/10.5802/crmath.598

CENTRE MERSENNE

The Comptes Rendus. Mathématique *are a member of the*
Mersenne Center for open scientific publishing
www.centre-mersenne.org — e-ISSN : 1778-3569

ACADÉMIE
DES SCIENCES
INSTITUT DE FRANCE

Comptes Rendus. Mathématique
2024, Vol. 362, p. 1109-1119
https://doi.org/10.5802/crmath.598

Research article / *Article de recherche*
Algorithmic and computer tools / *Algorithmes et outils informatiques*

The Dual Characteristic-Galerkin Method

La méthode des caractéristiques-Galerkin duale

Frédéric Hecht [a] **and Olivier Pironneau** [*, a]

[a] LJLL, Boite 187, Sorbonne Université, Place Jussieu, 75005 Paris, France
E-mails: Frederic.hecht@sorbonne-universite.fr, olivier.pironneau@academie-sciences.fr

Abstract. The Dual Characteristic-Galerkin method (DCGM) is conservative, precise and experimentally positive. We present the method and prove convergence and L^2-stability in the case of Neumann boundary conditions. In a 2D numerical finite element setting (FEM), the method is compared to Primal Characteristic-Galerkin (PCGM), Streamline upwinding (SUPG), the Dual Discontinuous Galerkin method (DDG) and centered FEM without upwinding. DCGM is difficult to implement numerically but, in the numerical context of this note, it is far superior to all others.

Résumé. La méthode *Dual Characteristic-Galerkin* (DCGM) est conservative, précise et expérimentalement positive. Nous prouvons la convergence et la stabilité L^2. Dans le cadre numérique des méthodes d'éléments finis (FEM) en 2D, la méthode est comparée à la méthode *Primal Characteristic-Galerkin* (PCGM), au *Streamline upwinding* (SUPG), à la méthode *Dual Discontinuous Galerkin* (DDG) et à une discretisation FEM sans décentrage. La méthode DCGM est difficile à mettre en œuvre numériquement, mais elle est de loin supérieure à toutes les autres dans le cadre étudié dans cette note.

Keywords. Partial differential equations, convection-diffusion, numerical method, finite element method.

Mots-clés. Équations aux dérivées partielles, convection-diffusion, schémas numériques, éléments finis.

2020 Mathematics Subject Classification. 35Q35, 65M06, 65M15, 65M25, 65M60.

Manuscript received 23 October 2023, revised 19 January 2024, accepted 6 December 2023.

Introduction

A good numerical method for the convection-diffusion equation is important in itself but it is also a test bed for more complex systems such as the Navier–Stokes equations. A finite element method (FEM) combined with a first or second order implicit in time discretization without upwinding works only if a CFL condition is satisfied, a severe constraint if the viscous coefficient is small (the method is also known as Arakawa's scheme in meteorology [8]). Hence in the eighties a number of upwinding schemes have been proposed in particular by K. Baba et al [1], J.-P. Benque et al [2] T.J.R. Hughes [7] and O. Pironneau[11]. Later, in the nineties Finite Volume methods and Discontinuous Galerkin methods were proposed for non-solenoidal convective velocities (see for example A. Ern et al [4].)

Recently we were faced with the problem of finding a good method for the computation of the probability density of a process via the Kolmogorov forward equation. Here positivity and

* Corresponding author

conservativity are essential. A more subjective criteria is the numerical diffusivity. It became an opportunity to review the state of the art forty years after the above mentioned methods were proposed, what R. Glowinski would call a rear-guard battle. Nevertheless, the following methods are popular:

- The Primal Characteristic-Galerkin method (PCGM) proposed in [11] is very precise but known to diverge in some cases when the viscosity is zero [14] and it is not conservative. It is convergent when mass-lumping is used [12] but then it is too diffusive.
- The Dual Characteristic-Galerkin method (DCGM) proposed in [2] by J.P. Benque et al. was never shown to converge except possibly when the initial and convected triangulations are intersected.
- T.J.R. Hughes' streamline upwinding method (SUPG) [7], also called Galerkin Least-square upwinding [9], is easy to implement, conservative and convergent but numerically diffusive, even when the *upwinding parameter* is tuned to the problem.

In the present note we study the DCGM with numerical quadrature for the nonlinear integral, prove that it is conservative, L^2-stable and convergent when the diffusion coefficient ν is not zero. Proposition 5, below, shows that the method is $O(h+h^2/\delta t)$ when $\nu \gg h^2/\delta t$; δt is the time step and h is the size of the edges of the triangulation.

The numerical section shows the superiority of DCGM over all 4 above cited methods. But DCGM is difficult to program. Indeed it is hard (but not computer intensive) to find in which element of the triangulation lies a given point, a well known problem of computational geometry [13].

Note also that the paper analyzes only the case of homogeneous Neumann condition. It ends with a numerical test with non-homogenous Dirichlet conditions for the Navier–Stokes equations, but the error analysis does not apply and it seems that it is numerically sensitive to the choice of the time step.

1. The Dual Characteristic-Galerkin Method

Given a real parameter $\nu > 0$, a bounded open set Ω of $\mathbb{R}^d$, $d = 2,3$, a smooth velocity field $\mathbf{a}: \Omega\times(0,T)\to\mathbb{R}^d$ and an initial condition $u^0:\mathbb{R}^d\to\mathbb{R}$, we wish to find $u:\Omega\times(0,T)\to\mathbb{R}$ such that, at all time $t\in(0,T)$,

$$\partial_t u + \mathbf{a}\cdot\nabla u - \nu\Delta u = 0, \quad u(0) = u^0 \text{ in } \Omega, \quad \partial_n u = 0 \text{ on } \partial\Omega. \tag{1}$$

Let $\bar{\mathbf{a}}$ be the extension of $\mathbf{a}$ by zero outside Ω. Define: $\dot{\boldsymbol{\eta}}(t) = \bar{\mathbf{a}}(\boldsymbol{\eta}(t))$, $\boldsymbol{\eta}(0) = \mathbf{x}$ and $\boldsymbol{\eta}^{\pm}(\mathbf{x}) = \boldsymbol{\eta}(\pm\delta t)$. Recall that

$$\partial_t u(\mathbf{x},t) + \mathbf{a}(\mathbf{x})\cdot\nabla u(\mathbf{x},t) = \lim_{\delta t\to 0}\frac{1}{\delta t}[u(\mathbf{x},t) - u(\boldsymbol{\eta}^-(\mathbf{x}), t-\delta t)].$$

We assume that $\nabla\cdot\mathbf{a} = 0$ and $\mathbf{a}\cdot\mathbf{n} = 0$ at the boundary $\Gamma := \partial\Omega$, so that $\boldsymbol{\eta}^{\pm}(\Omega) = \Omega$ and $\det\nabla\boldsymbol{\eta}^{\pm} = 1$. Hence two variational formulations of the problem discretized in time are feasible,

$$\begin{aligned}
&\int_\Omega \left(\frac{1}{\delta t}(u^n\,\hat{u} - u^{n-1}\circ\boldsymbol{\eta}^-\,\hat{u}) + \nu\nabla u^n\cdot\nabla\hat{u}\right) = 0 \quad \forall\hat{u}\in H^1(\Omega), \quad \text{(Primal form)},\\
&\int_\Omega \left(\frac{1}{\delta t}(u^n\,\hat{u} - u^{n-1}\,\hat{u}\circ\boldsymbol{\eta}^+) + \nu\nabla u^n\cdot\nabla\hat{u}\right) = 0 \quad \forall\hat{u}\in H^1(\Omega), \quad \text{(Dual form)}.
\end{aligned}\tag{2}$$

We have used $\boldsymbol{\eta}^+(\boldsymbol{\eta}^-(\mathbf{x})) = \mathbf{x}$ and,

$$\int_\Omega f(\mathbf{x})g(\boldsymbol{\eta}^-(\mathbf{x})) = \int_{\boldsymbol{\eta}^-(\Omega)} g(\mathbf{y})f(\boldsymbol{\eta}^+(\mathbf{y}))/\det\nabla\boldsymbol{\eta}^-(\mathbf{y}) = \int_\Omega g(\mathbf{y})f(\boldsymbol{\eta}^+(\mathbf{y})). \tag{3}$$

A spatial discretization with the Finite Element Method (FEM) of the first line in (2) leads to the Primal Characteristic-Galerkin method (PCGM); on the second line it leads to the Dual Characteristic-Galerkin method (DCGM): finds $u^n \in V_h$ such that

$$\int_\Omega \left(u_h^n \hat{u}_h + \delta t \nu \nabla u_h^n \cdot \nabla \hat{u}_h\right) = \sum_{i\in I} u_h^{n-1}(\boldsymbol{\xi}^i)\hat{u}_h(\boldsymbol{\eta}^i)\omega^i, \quad \forall \hat{u}_h \in V_h, \tag{4}$$

where,

- Ω is polygonal so as to be covered by a triangulation $\bigcup_k T^k$.
- The points $\{\boldsymbol{\xi}^i\}_{i\in I}$ and positive weights $\{\omega^i\}_{i\in I}$ define a quadrature rule which must be exact at least for continuous piecewise-P^2 functions on the triangulation. We assume that the quadrature is defined on triangles so as to write

$$\sum_{i\in I} f(\boldsymbol{\xi}^i)\omega^i := \sum_k \sum_{i\in I(T^k)} f(\boldsymbol{\xi}^i)\omega_k^i, \quad I = \bigcup_k I(T^k). \tag{5}$$

Example 1. In 2D one may choose the quadrature points at the mid edges and $\omega_k^i = \frac{1}{3}$, but more precise formulae are permitted.

- $\boldsymbol{\eta}^i \in \Omega$ is an approximation of $\boldsymbol{\eta}^+$ with $|\boldsymbol{\eta}^i - \boldsymbol{\eta}^+(\boldsymbol{\xi}^i)| \le C\delta t^2$. For example

$$\boldsymbol{\eta}_a^+(\mathbf{x}) = \mathbf{x} + \mathbf{a}(\mathbf{x})\delta t + \frac{\sigma}{2}\delta t^2 \mathbf{a}(\mathbf{x})\cdot\nabla\mathbf{a}(\mathbf{x}), \quad \sigma = 0 \text{ or } 1, \quad \boldsymbol{\eta}^i = \boldsymbol{\eta}_a^+(\boldsymbol{\xi}^i). \tag{6}$$

- V_h is the P^1 continuous finite element space.

Proposition 2. *DCGM conserves mass in the sense that*

$$\int_\Omega u_h^n = \int_\Omega u_h^0, \ \forall n.$$

Proof. Simply replace $\hat{u}_h$ by 1 in the scheme. □

Proposition 3. *Assume that the triangulation is regular, in the sense of [3, p. 131], i.e. for all triangles, the ratio of largest edge to the radius of the inscribed circle is bounded independently of h. Then DCGM is stable:*

$$\|u_h^n\|_{\nu\delta t} \le \left(1 + |\det \underline{\mathbf{A}}|\delta t^2 + C\frac{h^2}{\nu}\right)\|u_h^{n-1}\|_{\nu\delta t}$$

where $\|v\|_{\nu\delta t} := (|v|_0^2 + \delta t\nu|\nabla v|_0^2)^{\frac{1}{2}}$, *$C$ is a generic constant and h is the length of the longest edges in the triangulation.*

Proof. The proof is given in 2D with the quadrature at the mid-edges (Example 1) and scheme (6).

The discrete Cauchy–Schwarz inequality applied to the right hand-side of (4) combined with the choice $\hat{u}_h = u_h^n$ in (4), leads to

$$\|u_h^n\|_{\nu\delta t}^2 \le \left(\sum_{i\in I} u_h^{n-1}(\boldsymbol{\xi}^i)^2\omega^i\right)^{\frac{1}{2}}\left(\sum_{i\in I} u_h^n(\boldsymbol{\eta}^i)^2\omega^i\right)^{\frac{1}{2}} \le \|u_h^{n-1}\|_{\nu\delta t}\left(\sum_{i\in I} u_h^n(\boldsymbol{\eta}^i)^2\omega^i\right)^{\frac{1}{2}}, \tag{7}$$

because the quadrature is exact for $(u_h^{n-1})^2$ and because $|u_h^{n-1}|_0 \le \|u_h^{n-1}\|_{\nu\delta t}$. The map $\boldsymbol{\xi} \to \boldsymbol{\eta}_a^+(\boldsymbol{\xi})$ defined by (6) transforms a triangle T^k of the triangulation into $\hat{T}^k$ and $\{\boldsymbol{\eta}^i, \omega^i\}_{i\in I}$ is a quadrature rule which is almost exact on P^2 functions of $\hat{T}^k$. We will show that, for some C,

$$\sum_k \sum_{i\in I(T^k)} u_h^n(\boldsymbol{\eta}^i)^2\omega_k^i \le \left(1 + C\left(\frac{h^2}{\nu} + \delta t^2\right)\right)\|u_h^n\|_{\nu\delta t}^2. \tag{8}$$

Proof of (8) **in the linear case.** Assume that $\mathbf{a}$ is linear in $\mathbf{x} = (x, y)^T$ with $\nabla\cdot\mathbf{a} = 0$, and consider the case $\sigma = 0$ in (6),

$$\boldsymbol{\eta}^+(\mathbf{x}) = \mathbf{x} + \delta t\mathbf{a}(\mathbf{x}) = \mathbf{x} + \delta t\begin{bmatrix} a_1^0 \\ a_2^0 \end{bmatrix} + \delta t\begin{bmatrix} \partial_x \mathbf{a}_1 x + \partial_y \mathbf{a}_1 y \\ \partial_x \mathbf{a}_2 x - \partial_x \mathbf{a}_1 y \end{bmatrix} = \boldsymbol{\eta}_a^+(\mathbf{x}).$$

It is not quite an isometry because $\det\nabla(\mathbf{x}+\mathbf{a}\delta t) = 1-[(\partial_x\mathbf{a}_1)^2+\partial_y\mathbf{a}_1\partial_x\mathbf{a}_2]\delta t^2$.

Consider the quadrature at the mid edges with weight $\omega^i_k = \frac{1}{3}|T^k|$, the area of T^k. A triangle $(\mathbf{q}^1,\mathbf{q}^2,\mathbf{q}^3)$ is transformed by $\boldsymbol{\eta}^+$ into the triangle $(\widehat{\mathbf{q}}^1,\widehat{\mathbf{q}}^2,\widehat{\mathbf{q}}^3)$ with

$$\widehat{\mathbf{q}}^j = \mathbf{q}^j + \delta t\mathbf{a}^0 + \delta t(\nabla\mathbf{a})^T\mathbf{q}^j.$$

Obviously a mid edge $\frac{1}{2}(\mathbf{q}^{j_1}+\mathbf{q}^{j_2})$ of T^k is mapped into a mid edge of $\widehat{T}^k$. Therefore, the only error is due to the variation of the area of the triangle: $|\widehat{T}^k| = \det\nabla(\mathbf{x}+\delta t\mathbf{a})|T^k|$. Indeed, as $u^n_h(\boldsymbol{\eta}^+)$ is affine on T^k and because of (3),

$$\sum_{i\in I(T^k)} u^n_h(\boldsymbol{\eta}^i)^2\omega^i_k = |(u_h\circ\eta^+)^2|_{0,\hat{T}^k} = (1-\delta t^2\det\nabla\mathbf{a})|u^2_h|_{0,T^k},$$

because the quadrature is exact for P^2 functions; $|f|_{0,T}$ is the integral of f on T.

Proof in the general case. For simplicity we consider the case $\sigma = 0$ in (6). Consider a triangle T^k and a Taylor expansion of $\mathbf{a}$ about $\mathbf{x}^0$, the center of T^k,

$$\mathbf{a}(\mathbf{x}) = \mathbf{a}_0 + \underline{\mathbf{A}}(\mathbf{x}-\mathbf{x}^0) + (\mathbf{x}-\mathbf{x}^0)\otimes(\mathbf{x}-\mathbf{x}^0) : \underline{\boldsymbol{\Psi}}(\mathbf{x}),$$

for some bounded in $\mathbf{x}$ third order tensor $\boldsymbol{\Psi}$. Hence,

$$\boldsymbol{\eta}^+_a(\mathbf{x}) = \boldsymbol{\eta}_l(\mathbf{x}) + \delta t(\mathbf{x}-\mathbf{x}^0)\otimes(\mathbf{x}-\mathbf{x}^0) : \underline{\boldsymbol{\Psi}}(\mathbf{x}) \quad \text{where} \quad \boldsymbol{\eta}_l(\mathbf{x}) := \mathbf{x}+\delta t(\mathbf{a}_0+\underline{\mathbf{A}}(\mathbf{x}-\mathbf{x}^0)) \text{ is affine.}$$

Recall the notation $\boldsymbol{\eta}^i := \boldsymbol{\eta}^+_a(\boldsymbol{\xi}^i)$ and let $\boldsymbol{\eta}^i_l := \boldsymbol{\eta}_l(\boldsymbol{\xi}^i)$. The segment $[\boldsymbol{\eta}^i_l,\boldsymbol{\eta}^i]$ cuts a finite number of edges of the triangulation. Let these intersections be $\{\boldsymbol{\xi}^i_j\}_1^{J-1}$. With the convention that $\boldsymbol{\xi}^i_0 := \boldsymbol{\eta}^i_l$ and $\boldsymbol{\xi}^i_J := \boldsymbol{\eta}^i$, we can write

$$u^n_h(\boldsymbol{\eta}^i)^2 - u^n_h(\boldsymbol{\eta}^i_l)^2 = \sum_{0\le j\le J-1}\left(u^n_h(\boldsymbol{\xi}^i_{j+1})^2 - u^n_h(\boldsymbol{\xi}^i_j)^2\right).$$

Each term is continuously differentiable, so the following Taylor expansion is valid,

$$u^n_h(\boldsymbol{\eta}^i)^2 - u^n_h(\boldsymbol{\eta}^i_l)^2 = 2\sum_{0\le j\le J-1} u^n_h(\mathbf{x}^i_j)\cdot\nabla u^n_j(\mathbf{x}^i_j)(\boldsymbol{\xi}^i_{j+1}-\boldsymbol{\xi}^i_j) \le 2\max_j\left|u^n_h(\mathbf{x}^i_j)\cdot\nabla u^n_j(\mathbf{x}^i_j)\right|\,|\boldsymbol{\eta}^i-\boldsymbol{\eta}^i_l|,$$

where $\mathbf{x}^i_j \in [\boldsymbol{\xi}^i_j,\boldsymbol{\xi}^i_{j+1}]$. Let $\mathbf{x}^i_M = \arg\max_j|u^n_h(\mathbf{x}^i_j)\cdot\nabla u^n_j(\mathbf{x}^i_j)|$. Then we have found $\mathbf{x}^i_M \in [\boldsymbol{\eta}^i,\boldsymbol{\eta}^i_l]$ such that,

$$u^n_h(\boldsymbol{\eta}^i)^2 \le u^n_h(\boldsymbol{\eta}^i_l)^2 + 2|u^n_h(\mathbf{x}^i_M)\cdot\nabla u^n_j(\mathbf{x}^i_M)|\,|\boldsymbol{\eta}^i-\boldsymbol{\eta}^i_l|.$$

As $\nabla\cdot\mathbf{a} = 0$, $\underline{\mathbf{A}}$ is as in the linear case . Hence, $\mathbf{x}\to\eta_l(\mathbf{x})$ being affine, by (7), $\sum_{i\in I(T^k)} u^n_h(\boldsymbol{\eta}^i_l)^2\omega^i_k$ is bounded by $(1-\det\underline{\mathbf{A}}\delta t^2)|u^n_h|^2_{0,T^k}$. Now $|\boldsymbol{\eta}^i-\boldsymbol{\eta}^i_l| = \delta t(\boldsymbol{\xi}^i-\mathbf{x}^0)\otimes(\boldsymbol{\xi}^i-\mathbf{x}^0):\Psi|$, so,

$$\sum_{i\in I(T^k)} u^n_h(\boldsymbol{\eta}^i)^2\omega^i_k \le (1-\det\underline{\mathbf{A}}\delta t^2)|u^n_h|^2_{0,T^k} + h^2\delta t\|\boldsymbol{\Psi}\|_\infty\sum_{i\in I(T^k)} 2|u^n_h(\mathbf{x}^i_M)\cdot\nabla u^n_h(\mathbf{x}^i_M)|\omega^i_k$$

A discrete Cauchy–Schwarz inequality leads to,

$$2\left|u^n_h(\mathbf{x}^i_M)\right|\left|\nabla u^n_h(\mathbf{x}^i_M)\right| \le u^n_h(\mathbf{x}^i_M)^2 + \left|\nabla u^n_h(\mathbf{x}^i_M)\right|^2 \le \frac{1}{\nu\delta t}\left(u^n_h(\mathbf{x}^i_M)^2 + \nu\delta t|\nabla u^n_h(\mathbf{x}^i_M)|^2\right).$$

At the cost of a multiplicative constant we may replace $\mathbf{x}^i_M$ by $\boldsymbol{\xi}^{j(i)}$, the nearest quadrature point in the triangle of $\mathbf{x}^i_M$ and obtain,

$$\sum_k\sum_{i\in I(T^k)} 2|u^n_h(\mathbf{x}^i_M)\cdot\nabla u^n_h(\mathbf{x}^i_M)|\omega^i_k \le \frac{C}{\nu\delta t}\sum_k\sum_{i\in I(T^k)}\left(u^n_h(\boldsymbol{\xi}^{j(i)})^2 + \nu\delta t|\nabla u^n_h(\boldsymbol{\xi}^{j(i)})|^2\right)\omega^i_k \le \frac{C'}{\nu\delta t}\|u^n_h\|^2_{\nu\delta t}.$$

The last inequality holds for a regular triangulation because each quadrature point occurs at most N times, finite, and the ω_k^i differs from $\omega_k^{j(i)}$ at most by the ratio R of areas of triangles:

$$\sum_{k,i\in I(T^k)}\left(u_h^n(\boldsymbol{\xi}^{j(i)})^2+\nu\delta t|\nabla u_h^n(\boldsymbol{\xi}^{j(i)})|^2\right)\omega_k^i \le \sum_{k,i\in I(T^k)}\max\frac{\omega_k^i}{\omega_k^{j(i)}}\left(u_h^n(\boldsymbol{\xi}^{j(i)})^2+\nu\delta t|\nabla u_h^n(\boldsymbol{\xi}^{j(i)})|^2\right)\omega_k^{j(i)}$$
$$\le R\,N\sum_{k,i\in I(T^k)}\left(u_h^n(\boldsymbol{\xi}^i)^2+\nu\delta t|\nabla u_h^n(\boldsymbol{\xi}^i)|^2\right)\omega_k^i.$$

In the end,

$$\sum_k\sum_{i\in I(T^k)}u_h^n(\boldsymbol{\eta}^i)^2\omega_k^i\le\left(1+|\det\underline{\mathbf{A}}|\delta t^2+C\frac{h^2}{\nu}\right)\|u_h^n\|_{\nu\delta t}^2.$$

This proves (8) and completes the proof of Proposition 3. □

1.1. *Error Estimates*

Let $u_e^n\in H^1(\Omega)$ be the solution of the continuous problem (1) discretized in time and with the same $\boldsymbol{\eta}_a^+$ as in the discrete case; then let $u_{eh}^n\in V_h$ be the projection of u_e^n in the sense that

$$\int_\Omega(u_e^n\hat u+\nu\delta t\nabla u_e^n\nabla\hat u)=\int_\Omega u_e^{n-1}\cdot\hat u\circ\boldsymbol{\eta}_a^+,\qquad\forall\hat u\in H^1(\Omega),$$
$$\int_\Omega(u_{eh}^n\hat u_h+\nu\delta t\nabla u_{eh}^n\nabla\hat u_h)=\int_\Omega(u_e^n\hat u_h+\nu\delta t\nabla u_e^n\nabla\hat u_h)\quad\forall\hat u_h\in V_h.\tag{9}$$

Lemma 4. *Let $\epsilon_h^n=u_h^n-u_{eh}^n$ defined by* (9). *Then,*

$$\|\epsilon_h^n\|_{\nu\delta t}^2\le\left(1+C\left(\frac{h^2}{\nu}+\delta t^2\right)\right)\|\epsilon_h^{n-1}\|_{\nu\delta t}^2+Ch^2\|\epsilon_h^{n-1}\|_{\nu\delta t}.\tag{10}$$

Proof. Let Q be the quadrature (5),

$$Q_\Omega(v,w):=\sum_{i\in I}v(\boldsymbol{\xi}^i)w(\boldsymbol{\xi}^i)\omega^i=\sum_k Q_{T^k}(v,w),\quad Q_{T^k}(v,w)=\sum_{i\in I(T^k)}v(\boldsymbol{\xi}^i)w(\boldsymbol{\xi}^i)\omega_k^i.$$

Then $\forall\hat u_h\in V_h$,

$$\int_\Omega\left(\epsilon_h^n\hat u_h+\delta t\nu\nabla\epsilon_h^n\cdot\nabla\hat u_h\right)=Q_\Omega(u_h^{n-1},\hat u_h\circ\boldsymbol{\eta}_a^+)-\int_\Omega u_e^{n-1}\cdot\hat u_h\circ\boldsymbol{\eta}_a^+$$
$$=Q_\Omega(\epsilon_h^{n-1},\hat u_h\circ\boldsymbol{\eta}_a^+)+Q_\Omega(u_{eh}^{n-1},\hat u_h\circ\boldsymbol{\eta}_a^+)-\int_\Omega u_e^{n-1}\cdot\hat u_h\circ\boldsymbol{\eta}_a^+$$

Consequently

$$\|\epsilon_h^n\|_{\nu\delta t}^2=Q_\Omega(\epsilon_h^{n-1},\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+)+Q_\Omega(u_{eh}^{n-1}-u_e^{n-1},\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+)$$
$$+Q_\Omega(u_e^{n-1},\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+)-\int_\Omega u_e^{n-1}\cdot\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+.$$

A discrete Schwartz inequality is applied to the first term on the right and then (8),

$$Q_\Omega(\epsilon_h^{n-1},\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+)\le\left(1+C\left(\frac{h^2}{\nu}+\delta t^2\right)\right)\|\epsilon_h^{n-1}\|_{\nu\delta t}^2$$

The second term is handled in the same way,

$$Q_\Omega(u_{eh}^{n-1}-u_e^{n-1},\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+)\le\left(1+C\left(\frac{h^2}{\nu}+\delta t^2\right)\right)\|\epsilon_h^{n-1}\|_{\nu\delta t}\cdot\|u_{eh}^{n-1}-u_e^{n-1}\|_0$$
$$\le Ch^2\left(1+C\left(\frac{h^2}{\nu}+\delta t^2\right)\right)\|\epsilon_h^{n-1}\|_{\nu\delta t}.$$

Finally the third term is bounded by the quadrature error on $\widehat{T}^k$ for $u_e^{n-1}\circ(\eta^+)^{-1}$,

$$Q_\Omega(u_e^{n-1},\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+)-\int_\Omega u_e^{n-1}\cdot\epsilon_h^{n-1}\circ\boldsymbol{\eta}_a^+\le(1+C\delta t^2)h^2\|u_e^{n-1}\circ(\boldsymbol{\eta}_a^+)^{-1}\|_3\cdot\|\epsilon_h^{n-1}\|_{\nu\delta t}.$$

Let us gather the pieces

$$\|\epsilon_h^n\|_{\nu\delta t}^2\le\left(1+C(\frac{h^2}{\nu}+\delta t^2)\right)\|\epsilon_h^{n-1}\|_{\nu\delta t}^2+Ch^2\|\epsilon_h^{n-1}\|_{\nu\delta t}\qquad\square$$

Proposition 5.

$$\|\epsilon_h^n\|_{\nu\delta t}\le\left(\|\epsilon_h^0\|_{\nu\delta t}+C\frac{h^2}{\delta t}\right)\left(1+C\left(\frac{h^2}{\nu}+\delta t^2\right)\right)^n.\tag{11}$$

Proof. Recurrence (10) is of the type

$$(\varepsilon^n)^2-(\varepsilon^{n-1})^2\le\alpha(\varepsilon^n)^2+\beta\varepsilon^n$$

with $\varepsilon^n=\|\epsilon_h^n\|_{\nu\delta t}$, $\beta=Ch^2$ and $\alpha=C(\frac{h^2}{\nu}+\delta t^2)$. It is rewritten as

$$\varepsilon^n-\varepsilon^{n-1}\le\frac{\varepsilon^{n-1}}{\varepsilon^n+\varepsilon^{n-1}}(\alpha\varepsilon^{n-1}+\beta)\le\alpha\varepsilon^{n-1}+\beta$$
$$\Rightarrow\quad\varepsilon^n\le\varepsilon_0(1+\alpha)^n+Ch^2\sum_{j=0}^{n-1}(1+\alpha)^j\le\varepsilon_0(1+\alpha)^n+\frac{(1+\alpha)^n-1}{\alpha}Ch^2.$$

The result derives from the fact that $n\le T/\delta t$ and $(1+\alpha)^n-1\le n\alpha(1+\alpha)^{n-1}$. $\square$

Remark 6. Notice that the sequence is closed to the solution of the ODE in time $\varepsilon'=\frac{1}{2\delta t}(\alpha\varepsilon+\beta)$,

$$\varepsilon(t)+\frac{\beta}{\alpha}=\left(\varepsilon(0)+\frac{\beta}{\alpha}\right)\exp\left(t\frac{\alpha}{2\delta t}\right),\text{ approximated by }\varepsilon(t)\approx\varepsilon(0)\left(1+t\frac{\alpha}{2\delta t}\right)+t\frac{\beta}{2\delta t}\text{ when }h^2\ll\nu\delta t,$$

because then $\frac{\alpha}{\delta t}\ll1$. So, at best, a tighter argument will only improve the constants in (11).

Remark 7. To derive the total error from ϵ_h^n is standard. The time discretization being first order it produces and extra $O(\delta t)$ term , so the total error is of order $\delta t+\frac{h^2}{\nu}$, provided $h^2<\nu\delta t$. Notice that here too, as for Primal Characterisic-Galerkin methods, δt should not be chosen too small.

2. Numerical Tests

2.1. *The Rotating Gaussian Bell*

A point $\mathbf{x}^0=(\mathbf{x}_1^0,\mathbf{x}_2^0)^T$ convected by $\mathbf{a}(\mathbf{x})=(-\mathbf{x}_2,\mathbf{x}_1)^T$ is in fact rotated at time t to $\mathbf{x}^0(t)=(\mathbf{x}_1^0\cos t+\mathbf{x}_2^0\sin t,-\mathbf{x}_1^0\sin t+\mathbf{x}_2^0\cos t)^T$. Consider

$$u_e(\mathbf{x},t)=\frac{e^{-\frac{r|\mathbf{x}-\mathbf{x}^0(t)|^2}{1+4\nu rt}}}{1+4\nu rt}\tag{12}$$

It verifies (1) and $\partial_n u_e\approx0$ if r is large and ν is small.

A Delaunay–Voronoi mesh generator is used for the triangulations of the unit circle. We tested 3 meshes with 926, 3601 and 14071 vertices, corresponding respectively to $N=100$, 200 and 400 boundary vertices. The corresponding number of time steps chosen are 33, 66 and 133.

The other parameters are $\mathbf{x}_1^0=0.35$, $\mathbf{x}_2^0=0,T=2\pi$, $\nu=10^{-4}$ or 0.01, $r=10$.

2.2. *Convergence Study*

In this section $\nu = 10^{-4}$.

The differential equation is discretized by (6) with $\sigma = 1$. V_h is constructed with the linear continuous triangular finite element method and the nonlinear integral is approximated with the mid-edges as quadrature points of Example 1 or a 9-points quadrature per triangle [5].

Figure 1 shows the convergence rate and Figure 2 shows the Gaussian bell after one turn. It is difficult to see the difference with the exact solution.

A discontinuous function is subject to the rotating field to test the robustness with respect to discontinuity. Results are on Figure 3. Finally, as shown by Figure 4 u_h need not be zero at the boundary. Figures 2, 3 and 4 have been computed with $N = 200$. Table 1 shows the positivity and conservativity of the method.

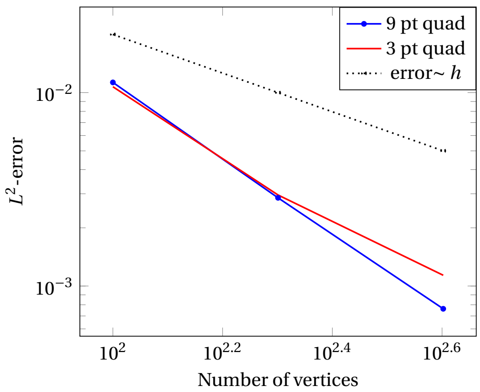

Figure 1. Plot (log-log scales) of L^2 error versus vertices number and effect of quadratures on the precision.

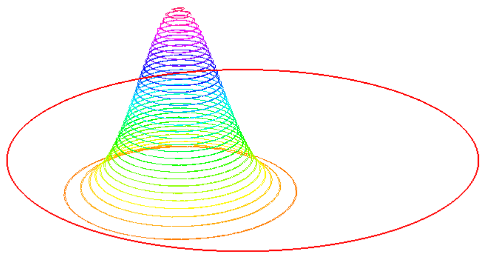

Figure 2. Gaussian Bell after one turn and exact solution. The level lines of both surfaces are very near to each others. Level lines values are as in Fig. 3.

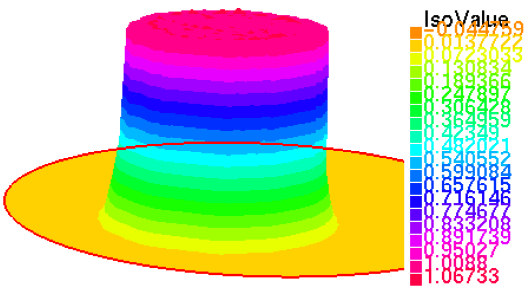

Figure 3. $u^0 = \mathbf{1}_{(x-0.3)^2+y^2<0.15}$ and u_h^T after one turn. Notice there is almost no oscillation and no numerical diffusion.

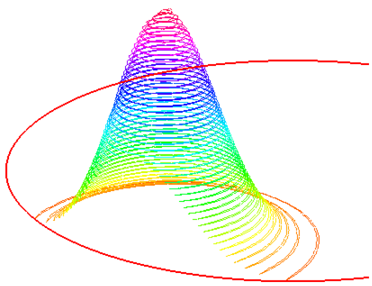

Figure 4. Gaussian bell crossing the boundary, because initially $x_0 = 0.5$, after one turn and exact solution.

Table 1. Positivity, Conservativity and Convergence

N	$\min u_h$	$\max u_h$	$\int_\Omega u_h$	L^2-error
100	−1.13689e-08	0.643741	0.156945	0.0112869
200	1.94281e-11	0.664612	0.156998	0.00282539
400	1.94281e-11	0.665645	0.156962	0.000763338
Exact	1.94281e-11	0.665268	0.156965	0

3. Comparison with other methods

In this section $\nu = 0.01$ and by default $N = 200$.

We ran the same tests with 4 other popular methods: PCGM [11], SUPG [7], DDG [4] and no upwinding [8]. Streamline Upwinding Galerkin (SUPG) reads:

$$\int_\Omega \left(\frac{u_h^n - u_h^{n-1}}{\delta t} + \mathbf{a}\cdot\nabla u \right)(w_h + \alpha \mathbf{a}\cdot\nabla w_h) + \int_\Omega \nu \nabla u_h^n \cdot \nabla w_h = 0$$

for all $w_h \in V_h$; $\alpha = 0.3$ in the numerical test.

With homogeneous Dirichlet conditions the Dual Discontinuous-Galerkin (DDG) methods is:

$$\int_\Omega \left(\left(\frac{u_h^n - u_h^{n-1}}{\delta t} + \mathbf{a}\cdot\nabla u_h^n \right) w_h + \nu \nabla u_h^n \cdot \nabla w_h \right) + \int_E w_h (\alpha |\mathbf{n}\cdot\mathbf{a}| - \frac{1}{2}\mathbf{n}\cdot\mathbf{a})[u_h^n] = 0$$

for all $w_h \in V_h$; $\alpha = 0.5$ in the numerical test. Here E is the set of inner edges and $[b]$ is the jump of b across an edge of E.

Finally the centered method which keeps the convective terms as is

$$\int_\Omega \left(\left(\frac{u_h^n - u_h^{n-1}}{\delta t} + \mathbf{a}\cdot\nabla u_h^n \right) w_h + \nu \nabla u_h^n \cdot \nabla w_h \right) = 0 \quad \forall w_h \in V_h.$$

A CFL condition $\delta t \le c(\nu) h^2$ is necessary for stability, so the method is not viable for small ν.

Figure 5 shows the horizontal cross sections of the Gaussian bell in the x direction after one turn for all 5 methods. Obviously PCGM and DCGM perform better, with the advantage that DCGM is convervative and convergence is proved. The level lines of the Gaussian bell after one turn are shown on Figures 6, 7, 8 and 10 and the positivity and conservativity on Table 2. Finally the convergence rates are shown in Figure 9.

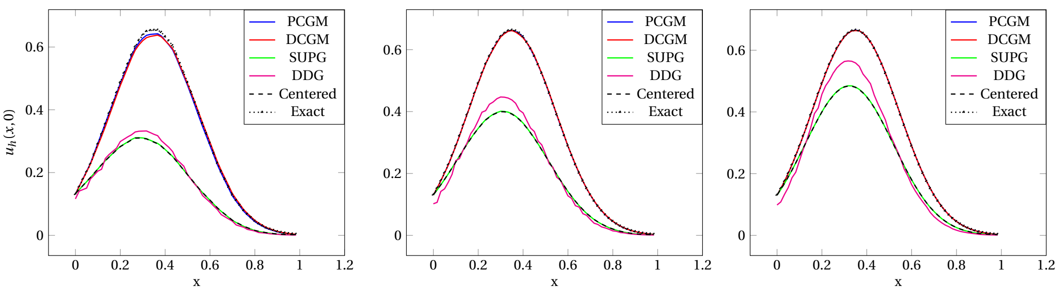

Figure 5. Plot of $x \to u_h(x,0)$ computed by the 5 methods, at $N = 100$ (left), $N = 200$ (middle) and $N = 400$ (right) .

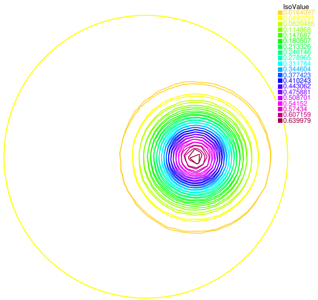

Figure 6. Bell computed with $N = 100$ and with PCGM after one turn and exact solution (level lines are essentially on top of each other).

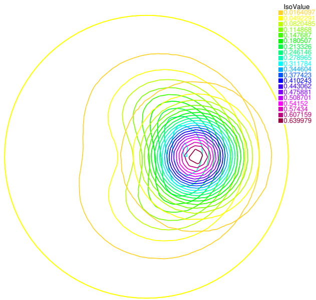

Figure 7. Bell computed with $N = 100$ and with SUPG after one turn and exact solution. Phase error, flatness error and maximum error are visible.

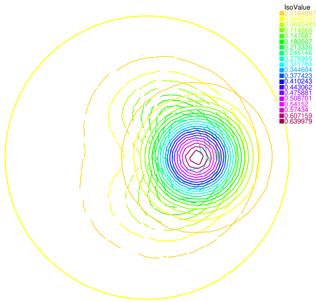

Figure 8. Bell computed with $N = 100$ and with DDG elements after one turn and exact solution. Phase, flatness and maximum error are visible.

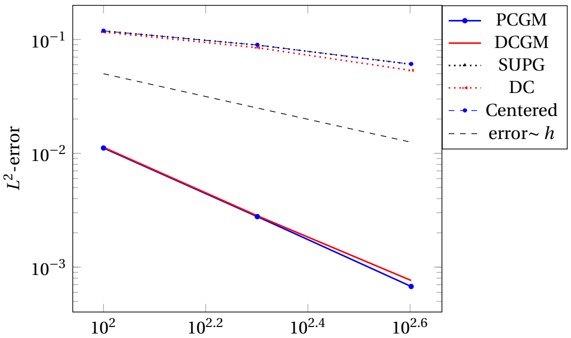

Figure 9. Plot (log-log scales) of L^2 error versus N. Both characteristic methods are equally precise and the other methods (SUPG, DDG, no upwding) are equally coarse.

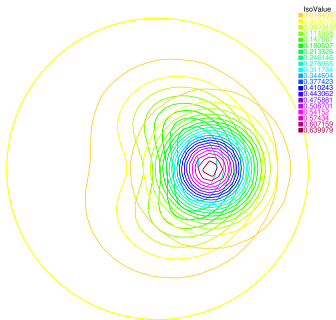

Figure 10. Bell computed with $N = 100$ and with the centered FEM (i.e. without upwinding). There are ten times more time steps to perform a turn. Phase error, maximum error and flatness error are visible.

Table 2. Comparison of the methods at N=200 after one turn.

Method	min u_h	max u_h	$\int_\Omega u_h$	L^2-error
u_e interpolated	1.94281e-11	0.66339	0.156984	
PCGM	1.94281e-11	0.662813	0.156777	0.00277886
DCGM	1.94281e-11	0.664612	0.156998	0.00282539
SUPG	1.94281e-11	0.40193	0.157103	0.0893023
DDG	2.27941e-06	0.448727	0.157102	0.0847009
Centered	1.94281e-11	0.400491	0.157099	0.0894042

4. Application to the Kolmogorov Equation for Heston's Model

Let $\mathbb{E}[f]$ be the expected value of a random f. In quantitative finance Heston's model [6] is,

$$\mathrm{d}X_t = X_t(r\mathrm{d}t + \sqrt{Y_t}\mathrm{d}W_t^1), \quad \mathrm{d}Y_t = \kappa(\theta - Y_t)\mathrm{d}t + \lambda\sqrt{Y_t}\mathrm{d}W_t^2,$$
$$\mathbb{E}[\mathrm{d}W_t^1\mathrm{d}W_t^2] = \rho,\ X_0 = \mathbb{N}(\mu,\sigma),\ Y_0 = \mathbb{N}(\mu',\sigma').$$

It is popular to set the (undiscounted) price of a "Put" to be $P_T = \mathbb{E}(K - X_T)^+$ at time T where K is the "strike". Here the random process $t \to \{X_t, Y_t\}$ is driven by its initial conditions $\{X_0, Y_0\}$ and the two normal Brownian motions $t \to W_t^i,\ i = 1,2$ with correlation ρ. The initial conditions are Gaussian random variables of means μ,μ' and standard deviations σ,σ'. The parameters r,κ,θ and λ are positive real numbers. Kolmogorov's theorem gives the PDF $u \in L^2(\mathbb{R}_+^2)$ of $\{X_t, Y_t\}$: for all $\{x, y, t\} \in \mathbb{R}_+^2 \times (0, T)$,

$$\partial_t u + \nabla\cdot\begin{bmatrix} rxu \\ \kappa(\theta - y)u \end{bmatrix} - \nabla^2 : \left(\begin{bmatrix} x^2y & \lambda xy \\ \lambda xy & \lambda^2 y \end{bmatrix} \frac{u}{2}\right) = 0, \qquad u_{|t=0} = G_{\mu,\sigma}(x)G_{\mu',\sigma'}(y), \tag{13}$$

where G is the Gaussian curve. Then $P_T = \int_{\mathbb{R}_+^2}(K - x)_+ u_T(x, y)$. Computing P_T for large T is a challenge because it is essential to keep having $\int_{\mathbb{R}_+^2} u_t = 1$ for all t and $u(x, y) \geq 0$ for all $x \geq 0, y \geq 0$.

We computed u_T at $T = 10$ with DCGM when $r = 0.03$, $K = 75$, $\mu = 50$, $\kappa = 2$, $\theta = 0.1$, $\lambda = 0.2$, $\rho = -0.5$, $\mu' = 0.75$, $\sigma = 10$, $\sigma' = 0.1$. The results are in Figure 11 after 1500 time iterations and a mesh of 150×150 vertices. No negative values are observed and by construction $\int_{\mathbb{R}_+^2} u = 1$.

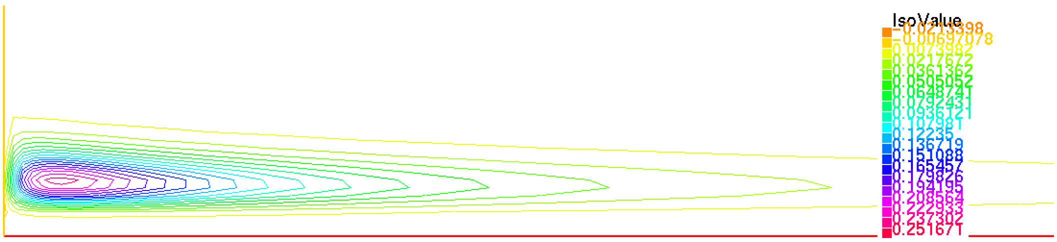

Figure 11. The level lines of the PDF of Heston's model at time T=10.

5. Non Homogeneous Dirichlet Conditions

Equation (3) is wrong when $\mathbf{a}\cdot\mathbf{n}|_\Gamma \neq 0$. To compensate with the fact that $\eta^-(\Omega) \neq \Omega$, a correction must be added (resp. subtracted) outside (resp. inside) Γ if $\mathbf{a}\cdot\mathbf{n}|_\Gamma$ is negative (reps. positive). For Dirichlet conditions $u = u_\Gamma$, we propose to replace (4) by: find $u_h^n - u_\Gamma \in V_{0h}$ such that

$$\int_\Omega (u_h^n \hat{u}_h + \delta t\nu\nabla u_h^n\cdot\nabla\hat{u}_h) - \int_\Gamma \delta t\mathbf{a}\cdot\mathbf{n}u_h^n\hat{u}_h = \sum_{i\in I} u_h^{n-1}(\boldsymbol{\xi}^i)\hat{u}_h(\boldsymbol{\eta}^i)\omega^i, \quad \forall\hat{u}_h \in V_{0h}, \tag{14}$$

This formulation was tested on the Navier–Stokes equations for the backward step problem, using the $P^2 - P^1$ element. Results are on Figure 12. However the results are better without the boundary integral on right, so the generalization to Dirichlet conditions is not straightforward, the problem is open.

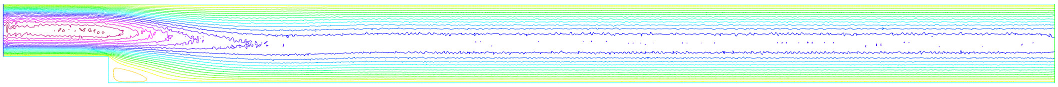

Figure 12. Stationary solution of the Navier–Stokes equation at Reynold 50. The level lines of the horizontal component of the fluid velocity are shown. The color scale is the same as that of Figure 3. The size of the recirculation is 3 times the height of the step as expected [10].

Declaration of interests

The authors do not work for, advise, own shares in, or receive funds from any organization that could benefit from this article, and have declared no affiliations other than their research organizations.

References

[1] K. Baba and M. Tabata, "On a conservative upwind finite element scheme for convective diffusion equations", *RAIRO, Anal. Numér.* **15** (1981), no. 1, pp. 3–25.

[2] J. P. Benque, B. Ibler and G. Labadie, "A finite element method for Navier–Stokes equations", in *Numerical Methods for Non-Linear Problems*, Pineridge Press, 1980, pp. 709–720.

[3] P. G. Ciarlet and J. L. Lions, *Finite element methods (Part 1)*, North-Holland, 1991.

[4] A. Ern and J.-L. Guermond, "Discontinuous Galerkin methods for Friedrichs' systems. I: General theory", *SIAM J. Numer. Anal.* **44** (2006), no. 2, pp. 753–778.

[5] F. Hecht, "New development in freefem++", *J. Numer. Math.* **20** (2012), no. 3-4, pp. 251–265.

[6] S. L. Heston, "A closed-form solution for options with stochastic volatility with applications to bond and currency options", *Rev. Financ. Stud.* **6** (1993), no. 2, pp. 327–343.

[7] T. J. R. Hughes, *The finite element method. Linear static and dynamic finite element analysis*, Prentice Hall, 1987, pp. xxviii+803.

[8] D. C. Jespersen, "Arakawa's method is a finite-element method", *J. Comput. Phys.* **16** (1974), pp. 383–390.

[9] C. Johnson, U. Nävert and J. Pitkäranta, "Finite element methods for linear hyperbolic problems", *Comput. Methods Appl. Mech. Eng.* **45** (1984), pp. 285–312.

[10] K. Morgan, J. Periaux and F. Thomasset, *Analysis of laminar flow over a backward facing step. A GAMM-Workshop (held on January 18-19, 1983, at Bièvres, France)*, Springer, 1984.

[11] O. Pironneau, "On the transport-diffusion algorithm and its applications to the Navier–Stokes equations", *Numer. Math.* **38** (1982), no. 3, pp. 309–332.

[12] O. Pironneau and M. Tabata, "Stability and convergence of a Galerkin-characteristics finite element scheme of lumped mass type", *Int. J. Numer. Methods Fluids* **64** (2010), no. 10-12, pp. 1240–1253.

[13] F. P. Preparata and M. I. Shamos, *Computational geometry*, Springer, 1985, pp. xii+390. An introduction.

[14] E. Süli, "Convergence and nonlinear stability of the Lagrange–Galerkin method for the Navier–Stokes equations", *Numer. Math.* **53** (1988), no. 4, pp. 459–483.