A posteriori analysis of a positive streamwise invariant discretization of a convection-diffusion equation

by Christine Bernardi[1], Tomás Chacón Rebollo[2], and Marco Restelli[2]

Abstract: We consider the finite element discretization of a convection-diffusion equation, where the convection term is handled via a fluctuation splitting algorithm. We prove a posteriori error estimates which allow us to perform mesh adaptivity in order to optimize the discretization of these equations. Numerical results confirm the interest of such an approach.

Résumé: Nous considérons une discrétisation par éléments finis d'une équation de convection–diffusion, où un algorithme de décentrage est utilisé pour traiter le terme de convection. Nous prouvons des estimations d'erreur a posteriori qui permettent d'adapter le maillage pour optimiser la discrétisation de ces équations. Des résultats numériques confirment l'intérêt d'une telle approche.

[1] Laboratoire Jacques-Louis Lions, C.N.R.S. & Université Pierre et Marie Curie - Paris 6,
boîte 187, 4 place Jussieu, 75252 Paris Cedex 05, France.
bernardi@ann.jussieu.fr

[2] Departamento de Ecuaciones Diferenciales y Análisis Numerico, Universidad de Sevilla,
Tarfia s/n, 41012 Sevilla, Spain.
chacon@numer.us.es, restelli@numer.us.es

1. Introduction.

We are interested in the discretization of the following convection-diffusion equation, set in a bounded connected domain Ω in $\mathbb{R}^2$ or $\mathbb{R}^3$,

$$-\alpha\,\Delta p + (\boldsymbol{u}\cdot\nabla)p = f \qquad \text{in } \Omega, \tag{1.1}$$

and provided with appropriate boundary conditions, in both cases where the diffusion term or the convection term is the leading one. The discretization that we propose relies on Lagrange finite element methods of low order. However, as standard for convection dominated problems, a further upwind scheme must be inserted into the discrete problem in order to have good convergence properties.

After the pioneering papers [15] and [10] where the method of characteristics is introduced, a large amount of work has been performed concerning the construction of new upwind schemes and their analysis. We quote among others the discontinuous Galerkin methods [7], several kinds of finite volume discretizations [9][19], the N scheme [14] and the positive streamwise invariant (PSI) algorithm introduced in [8]. We use here the abstract framework proposed in [5] for handling the more general fluctuation splitting methods, also called residual distribution methods [1], which includes the PSI scheme and other second-order non-linear residual-distribution schemes (see also [5] for a more general bibliography on these upwind schemes).

On the opposite, it seems that very few works deal with the a posteriori analysis of such methods (see [2] for the method of characteristics and [19] for finite volume methods). In this paper, we perform the a posteriori analysis of the PSI method, in the framework proposed in [5]. We observe that the techniques of proof are very similar to their analogue in the case of a simple Galerkin method and we establish optimal upper bounds of the error as a function of residual type error indicators (see [18] for the description of these indicators). Conversely, these indicators are bounded from above by the local error. So our estimates are fully optimal and the error indicators constitute a very efficient tool for mesh adaptivity.

We next describe our adaptivity strategy. The numerical experiments that we present are aimed to prove that mesh adaptivity is a very efficient tool for computing some solutions in realistic cases, for instance when they present boundary layers. They confirm the interest of using the PSI algorithm in strongly convection dominated regimes.

An outline of the paper is as follows:
• In Section 2, we present the abstract framework for the fluctuation splitting methods. Next, we recall from [5, §6 & 7] the main results concerning the well-posedness of the discrete problem and the convergence of its solution.
• Section 3 is devoted to the a posteriori analysis of the discrete problem.
• Numerical experiments are presented in Section 4.

2. The continuous and discrete problems.

Let Ω be a connected bounded open set in $\mathbb{R}^d$, $d = 2$ or 3, with a Lipschitz–continuous boundary $\partial\Omega$. The generic point in Ω is denoted by $\boldsymbol{x}$, while the generic point on $\partial\Omega$ is denoted by $\boldsymbol{\tau}$. As usual, $\boldsymbol{n}$ stands for the unit outward normal vector to Ω on $\partial\Omega$. With standard notation for the Sobolev spaces, we introduce a vector function $\boldsymbol{u}$ in $W^{1,r}(\Omega)^3$, $r > \frac{d}{2}$, and we make the further assumption that

$$\operatorname{div}\boldsymbol{u} = 0 \quad \text{in } \Omega. \tag{2.1}$$

We are thus in a position to introduce a partition of $\partial\Omega$ without overlap into Γ_- and Γ_+, with

$$\Gamma_- = \big\{\boldsymbol{\tau} \in \partial\Omega;\ (\boldsymbol{u}\cdot\boldsymbol{n})(\boldsymbol{\tau}) < 0\big\}, \qquad \Gamma_+ = \big\{\boldsymbol{\tau} \in \partial\Omega;\ (\boldsymbol{u}\cdot\boldsymbol{n})(\boldsymbol{\tau}) \geq 0\big\},$$

and we assume from now on that Γ_- has a positive measure in $\partial\Omega$. We are interested in the following equation

$$\begin{cases} -\alpha\,\Delta p + (\boldsymbol{u}\cdot\nabla)p = f & \text{in } \Omega, \\[2mm] p = 0 & \text{on } \Gamma_-, \\[2mm] \alpha\,\partial_n p = 0 & \text{on } \Gamma_+, \end{cases} \tag{2.2}$$

where the diffusion coefficient α is a positive constant. The unknown is the scalar function p and the datum f is assumed to be regular enough. Note that we have chosen to work with zero boundary conditions only for simplicity.

Let $H^1_*(\Omega)$ denote the space

$$H^1_*(\Omega) = \Big\{q \in H^1(\Omega);\ q = 0 \text{ on } \Gamma_-\Big\}. \tag{2.3}$$

It is readily checked that, when the boundaries of Γ_- and Γ_+ are smooth enough, system (2.2) admits the equivalent variational formulation

Find p in $H^1_*(\Omega)$ such that

$$\forall q \in H^1_*(\Omega), \quad a(p,q) = \langle f, q\rangle, \tag{2.4}$$

where the bilinear form $a(\cdot,\cdot)$ is defined by

$$a(p,q) = \alpha \int_\Omega (\mathbf{grad}\, p)(\boldsymbol{x}) \cdot (\mathbf{grad}\, q)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_\Omega \big((\boldsymbol{u}\cdot\nabla)\,p\big)(\boldsymbol{x})\,q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{2.5}$$

while $\langle\cdot,\cdot\rangle$ stands for the duality pairing between $H^1_*(\Omega)$ and its dual space $H^1_*(\Omega)'$.

Owing to the assumption on $\boldsymbol{u}$, we deduce from standard Sobolev imbeddings that the form $a(\cdot,\cdot)$ is continuous on $H^1(\Omega) \times H^1(\Omega)$. Moreover, we have for any function q in $H^1_*(\Omega)$

$$a(q,q) = \alpha\,|q|^2_{H^1(\Omega)} + \frac{1}{2}\int_{\Gamma_+}(\boldsymbol{u}\cdot\boldsymbol{n})(\boldsymbol{\tau})\,q^2(\boldsymbol{\tau})\,d\boldsymbol{\tau} \geq \alpha\,|q|^2_{H^1(\Omega)}.$$

Thus, using the Poincaré–Friedrichs inequality (indeed, Γ_- has a positive measure), we obtain the following ellipticity property: There exists a positive constant c such that

$$\forall q \in H^1_*(\Omega), \quad a(q,q) \geq c\,\alpha\,\|q\|^2_{H^1(\Omega)}. \tag{2.6}$$

2

All this together with the Lax–Milgram lemma yields that, for any datum f in $H^1_*(\Omega)'$, problem (2.4) has a unique solution.

We assume from now on that Ω is a polygon ($d = 2$) or a polyhedron ($d = 3$). Let $(\mathcal{T}_h)_h$ be a regular family of triangulations of Ω (by triangles or tetrahedra), in the sense that:

- For each h, $\overline{\Omega}$ is the union of all elements of $\mathcal{T}_h$;
- The intersection of two different elements of $\mathcal{T}_h$, if not empty, is a vertex or a whole edge or a whole face of both of them;
- The ratio of the diameter h_K of any element K of $\mathcal{T}_h$ to the diameter of its inscribed circle or sphere is smaller than a constant σ independent of h.

As usual, h stands for the maximum of the diameters h_K, $K \in \mathcal{T}_h$. We make the further and non restrictive assumption that both $\overline{\Gamma}_-$ and $\overline{\Gamma}_+$ are the union of whole edges ($d = 2$) or whole faces ($d = 3$) of elements of $\mathcal{T}_h$. In what follows, c, c', ... are generic constants that may vary from line to line but are always independent of h.

In view of the finite element discretization, we introduce the space

$$\mathbb{X}_h = \left\{ q_h \in H^1(\Omega); \ \forall K \in \mathcal{T}_h, \ q_h|_K \in \mathcal{P}_1(K) \right\},$$

where $\mathcal{P}_1(K)$ is the space of restrictions to K of affine functions with d variables. We also need its subspace

$$\mathbb{X}_h^* = \mathbb{X}_h \cap H^1_*(\Omega).$$

In view of the fluctuation splitting method, we introduce another finite-dimensional space $\mathbb{Y}_h^*$ made of piecewise constant functions. We assume that the dimension of $\mathbb{Y}_h^*$ is the same as the dimension of $\mathbb{X}_h^*$ and moreover we introduce a one-to-one linear mapping Π_h from $\mathbb{X}_h^*$ onto $\mathbb{Y}_h^*$.

Next, assuming that the datum f belongs to $L^2(\Omega)$, we consider the discrete problem
Find p_h in $\mathbb{X}_h^*$ such that

$$\forall q_h \in \mathbb{X}_h^*, \quad a_h(p_h, q_h) = \int_\Omega f(\boldsymbol{x})(\Pi_h q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{2.7}$$

where the bilinear form $a_h(\cdot, \cdot)$ is defined by

$$a_h(p_h, q_h) = \alpha \int_\Omega (\mathbf{grad}\, p_h)(\boldsymbol{x}) \cdot (\mathbf{grad}\, q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \int_\Omega \big((\boldsymbol{u} \cdot \nabla)\, p_h\big)(\boldsymbol{x})\, (\Pi_h q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{2.8}$$

To go further, we now describe the links between the spaces $\mathbb{X}_h^*$ and $\mathbb{Y}_h^*$ in the framework of the fluctuation splitting method. Let $\boldsymbol{b}_j$, $1 \le j \le J$, denote all the vertices of the elements of $\mathcal{T}_h$. With these vertices, we associate the standard Lagrange functions φ_j: For $1 \le j \le J$, the function φ_j belong to $\mathbb{X}_h$ and satisfies

$$\varphi_j(\boldsymbol{b}_j) = 1 \qquad \text{and} \qquad \varphi_j(\boldsymbol{b}_i) = 0, \quad 1 \le i \le J, \ i \ne j. \tag{2.9}$$

The functions φ_j, $1 \le j \le J$, form a basis of $\mathbb{X}_h$ and without restriction we assume that the φ_j, $1 \le j \le J^*$, form a basis of $\mathbb{X}_h^*$ (with $J^* < J$).

The fluctuation splitting method relies on the choice of linearly independent functions λ_j, $1 \le j \le J^*$, which are piecewise constant and have a small support in a neighbourhood

of $\boldsymbol{b}_j$ ("small" means contained in the union of triangles or tetrehedra in $\mathcal{T}_h$ that contain $\boldsymbol{b}_j$). Then, the operator Π_h is defined by

$$\Pi_h q_h = \sum_{j=1}^{J^*} q_h(\boldsymbol{b}_j)\,\lambda_j, \qquad (2.10)$$

and it is readily checked that it is one-to-one from $\mathbb{X}_h^*$ onto the space $\mathbb{Y}_h^*$ spanned by the functions λ_j, $1 \le j \le J^*$. Of course, since our aim is to construct upwind schemes, the operator Π_h, hence the functions λ_j, depend on the convection vector $\boldsymbol{u}$. They can also depend on the fluxes $\int_K \boldsymbol{u}(\boldsymbol{x}) \cdot (\nabla p_h)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, and in this case we use the notation $\Pi_h(p_h)$ and $\lambda_j(p_h)$ when needed (note that in this case the discrete problem is nonlinear).

For each K in $\mathcal{T}_h$, we denote by $\boldsymbol{b}_{j_K}$ the vertices of K, where j_K is the global index corresponding to the local index j, $1 \le j \le d+1$, and by λ_j^K the restriction of the function λ_{j_K} to K. From now on, we assume that the λ_j^K are zero when $\boldsymbol{b}_j$ is not a vertex of K.

Example 2.1. The N scheme [14]

For each K in $\mathcal{T}_h$, we introduce the normal vector to the edge opposite to $\boldsymbol{b}_{j_K}$

$$\boldsymbol{n}_j^K = d\,\mathrm{meas}(K)\,\mathbf{grad}\,\varphi_{j_K}.$$

Denoting by $\overline{\boldsymbol{u}}^K$ the mean value of $\boldsymbol{u}$ on K, we set

$$a_j^K = \frac{1}{d}\,\overline{\boldsymbol{u}}^K \cdot \boldsymbol{n}_j^K, \quad (a_j^K)^+ = \max\{a_j^K, 0\}, \quad (a_j^K)^- = \min\{a_j^K, 0\}, \quad b^K = \sum_{j=1}^{d+1}(a_j^K)^-.$$

Finally, for any function ρ_h in $\mathbb{X}_h^*$, we define the functions $\lambda_j^K(\rho_h)$ by

$$\lambda_j^K(\rho_h) = \frac{1}{\int_K \boldsymbol{u}(\boldsymbol{x}) \cdot (\nabla \rho_h)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}} \sum_{i=1}^{d+1}(a_j^K)^+ b^K (a_i^K)^- \big(\rho_h(\boldsymbol{b}_j^K) - \rho_h(\boldsymbol{b}_i^K)\big). \qquad (2.11)$$

Despite their complex form, the functions λ_j^K are easy to compute, for instance when $\boldsymbol{u}$ is constant on K. Note however that the local flux $\int_K \boldsymbol{u}(\boldsymbol{x}) \cdot (\nabla \rho_h)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ can vanish and, in this case, the functions λ_j^K are not defined.

Example 2.2. The PSI method

With the previous notation, the functions λ_j^K are now defined by

$$\lambda_j^K(\rho_h) = \frac{1 - \mu_j^K}{\int_K \boldsymbol{u}(\boldsymbol{x}) \cdot (\nabla \rho_h)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}} \sum_{i=1}^{d+1}(a_j^K)^+ b^K (a_i^K)^- \big(\rho_h(\boldsymbol{b}_j^K) - \rho_h(\boldsymbol{b}_i^K)\big), \qquad (2.12)$$

where the constants μ_j^K are chosen such that

$$0 \le \mu_j^K \le 1, \quad 1 \le j \le d+1, \qquad \text{and} \qquad \sum_{j=1}^{d+1}\lambda_j^K(\rho_h) = 1,$$

and also such that all λ_j^K, $K \in \mathcal{T}_h$, are bounded independently of ρ_h. It can be noted that the only difference between the N scheme and the PSI method relies on the choice of the

constants μ_j^K which are added to stabilize the discretization. Examples of constants μ_j^K satisfying the previous properties are given in [1] for instance.

To conclude, we recall from [5, §6 & 7] the main results concerning the well-posedness of problem (2.7) and the convergence of the discrete solution p_h to p. We are led to make a few assumptions for that.

Assumption 2.3. The following property holds for each K in $\mathcal{T}_h$

$$\lambda_j^K \geq 0, \quad 1 \leq j \leq d+1, \qquad \text{and} \qquad \sum_{j=1}^{d+1} \lambda_j^K = 1. \tag{2.13}$$

We now introduce the matrices linked to our discrete problem: D stands for the matrix with coefficients $\int_\Omega (\mathbf{grad}\,\varphi_i)(\boldsymbol{x}) \cdot (\mathbf{grad}\,\varphi_j)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $1 \leq i,j \leq J^*$, and, for any ρ_h in $\mathbb{X}_h^*$, $C(\rho_h)$ stands for the matrix with coefficients $\int_\Omega \big((\boldsymbol{u} \cdot \nabla)\,\varphi_i\big)(\boldsymbol{x})\,\big(\lambda_i(\rho_h)\big)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $1 \leq i,j \leq J^*$. We need a further definition.

Definition 2.4. A matrix A with coefficients A_{ij}, $1 \leq i,j \leq N$, is a M-matrix if

$$A_{ii} > 0, \quad 1 \leq i \leq N, \qquad \text{and} \qquad A_{ij} \leq 0, \quad 1 \leq i,j \leq N,\ i \neq j,$$

$$A_{ii} \geq \sum_{j \neq i} |A_{ij}|.$$

Assumption 2.5. For any ρ_h in $\mathbb{X}_h^*$ and any real number $\alpha > 0$, the matrix $\alpha D + C(\rho_h)$ is a M-matrix.

Assumption 2.6. The mapping: $\rho_h \mapsto C(\rho_h)$ is continuous from $\mathbb{X}_h^*$ into the space of square real matrices with dimension $J^* \times J^*$.

Indeed, the following results are established in [5, Thms 6.1 & 7.1].

Proposition 2.7. *If Assumptions* 2.3, 2.5 *and* 2.6 *are satisfied, for any datum f in $L^2(\Omega)$, problem* (2.7) *has a solution p_h in $\mathbb{X}_h^*$ for h small enough. Moreover, this solution satisfies*

$$\|p_h\|_{H^1(\Omega)} \leq c\,\|f\|_{L^2(\Omega)}. \tag{2.14}$$

Proposition 2.8. *If Assumptions* 2.3, 2.5 *and* 2.6 *are satisfied, any sequence of solutions $(p_h)_h$ of problem* (2.7) *converges to the solution p of problem* (2.4) *strongly in $H^1(\Omega)$ when h tends to zero.*

We refer to [5, Thm 7.3] for an explicit a priori estimate of the error between p and p_h under the previous assumptions: The scheme is first-order acurate in the $H^1(\Omega)$-norm.

We also observe that the N scheme and the PSI method both satisfy Assumption 2.3. On the other hand, they satisfy Assumption 2.5 only when the angles between edges $(d=2)$ or the dihedral angles between faces $(d=3)$ of all elements of $\mathcal{T}_h$ are $\leq \frac{\pi}{2}$. And the PSI method always satisfies Assumption 2.6, but the N-scheme does not. To conclude, let us remark that a relevant feature of the schemes verifying Assumptions 2.3, 2.5 and 2.6 is the verification of the discrete maximum principle, although this only occurs with piecewise affine discretizations.

3. A posteriori analysis of the discrete problem.

We need some further notation. With each K in $\mathcal{T}_h$, we associate
(i) the set $\mathcal{E}_K^0$ of all edges ($d = 2$) or faces ($d = 3$) of K which are not contained in $\partial\Omega$;
(ii) the sets $\mathcal{E}_K^+$ of all edges ($d = 2$) or faces ($d = 3$) of K which are contained in $\overline{\Gamma}_+$;
(iii) the domain ω_K equal to the union of all elements of $\mathcal{T}_h$ that share at least an edge ($d = 2$) or a face ($d = 3$) with K.
For each e in $\mathcal{E}_K^0$, we denote by $[\cdot]_e$ the jump through e (the introduction of a vector normal to e is necessary to make precise the sign of this jump, however we do not need it in what follows). For each e in $\mathcal{E}_K^0$ and in $\mathcal{E}_K^+$, h_e stands for the length ($d = 2$) or diameter ($d = 3$) of e.

We introduce a piecewise constant approximation f_h of the datum f: For each K in $\mathcal{T}_h$, we set

$$f_h|_K = \frac{1}{\text{meas}(K)} \int_K f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

We also need an approximation $\boldsymbol{u}_h$ of $\boldsymbol{u}$. Several choices are possible for that:
• Assuming that $\boldsymbol{u}$ is continuous on $\overline{\Omega}$, we can define $\boldsymbol{u}_h$ as the Lagrange interpolate of $\boldsymbol{u}$ with values in $\mathbb{X}_h$;
• A more consistent choice consists in taking $\boldsymbol{u}_h$ in the Raviart–Thomas finite element space [16]

$$\mathbb{Z}_h = \left\{ \boldsymbol{v}_h \in H(\text{div}, \Omega); \ \forall K \in \mathcal{T}_h, \ \boldsymbol{v}_h|_K \in \mathcal{P}_{RT}(K) \right\},$$

where $H(\text{div}, \Omega)$ is the domain of the divergence operator in $L^2(\Omega)^d$ and $\mathcal{P}_{RT}(K)$ stands for the space of restrictions to K of polynomials of the form $\boldsymbol{\mu} + \nu \boldsymbol{x}$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\nu \in \mathbb{R}$. Indeed, using the corresponding interpolation operator (see [16]) leads to a function $\boldsymbol{u}_h$ which is exactly divergence-free.

We are now in a position to define the error indicators: For each K in $\mathcal{T}_h$,

$$\eta_K = h_K \left\| f_h - (\boldsymbol{u}_h \cdot \nabla) p_h \right\|_{L^2(K)} + \sum_{e \in \mathcal{E}_K^0} \alpha h_e^{\frac{1}{2}} \left\| [\partial_n p_h]_e \right\|_{L^2(e)} + \sum_{e \in \mathcal{E}_K^+} \alpha h_e^{\frac{1}{2}} \left\| \partial_n p_h \right\|_{L^2(e)}. \quad (3.1)$$

It can be noted that all the functions which appear in the norms are polynomials with low degree, so that these indicators are very easy to compute once the discrete solution is known.

We successively prove an upper and a lower bound of the error as a function of these indicators. We set:

$$\varepsilon_K = h_K \left\| \big((\boldsymbol{u} - \boldsymbol{u}_h) \cdot \nabla \big) p_h \right\|_{L^2(K)}. \quad (3.2)$$

Proposition 3.1. *If Assumption 2.3 is satisfied, the following a posteriori estimate holds for the error between the solution p of problem (2.4) and any solution p_h of problem (2.7)*

$$\alpha \left\| p - p_h \right\|_{H^1(\Omega)} \leq c \left(\sum_{K \in \mathcal{T}_h} \left(\eta_K^2 + \varepsilon_K^2 + h_K^2 \left\| f - f_h \right\|_{L^2(K)}^2 \right) \right)^{\frac{1}{2}}. \quad (3.3)$$

Proof: Using (2.4) yields the residual equation, for any q un $H_*^1(\Omega)$,

$$a(p - p_h, q) = \int_\Omega f(\boldsymbol{x}) q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - a(p_h, q).$$

6

Then, it follows from (2.7) that, for any q_h in $\mathbb{X}_h^*$,

$$a(p - p_h, q) = \int_\Omega f(\boldsymbol{x})(q - \Pi_h q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$- \alpha \int_\Omega (\mathbf{grad}\, p_h)(\boldsymbol{x}) \cdot \big(\mathbf{grad}\,(q - q_h)\big)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int_\Omega \big((\boldsymbol{u} \cdot \nabla)\, p_h\big)(\boldsymbol{x})\,(q - \Pi_h q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

By integrating by parts on each K, we observe that, since p_h is affine on K,

$$\alpha \int_K (\mathbf{grad}\, p_h)(\boldsymbol{x}) \cdot \big(\mathbf{grad}\,(q - q_h)\big)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \sum_{e \in \mathcal{E}_K^0 \cup \mathcal{E}_K^+} \alpha \int_e (\partial_n p_h)(\boldsymbol{\tau})(q - q_h)(\boldsymbol{\tau}) \, \mathrm{d}\boldsymbol{\tau}.$$

By also inserting the approximations f_h and $\boldsymbol{u}_h$ in the previous equation, we obtain, with the appropriate signs for the jumps $[\partial_n p_h]_e$,

$$a(p - p_h, q)$$

$$= \sum_{K \in \mathcal{T}_h} \Bigg(\int_K (f - f_h)(\boldsymbol{x})(q - \Pi_h q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int_K \big(((\boldsymbol{u} - \boldsymbol{u}_h) \cdot \nabla)\, p_h\big)(\boldsymbol{x})\,(q - \Pi_h q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$+ \int_K \big(f_h - (\boldsymbol{u}_h \cdot \nabla)\, p_h\big)(\boldsymbol{x})\,(q - \Pi_h q_h)(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$- \frac{1}{2} \sum_{e \in \mathcal{E}_K^0} \alpha \int_e [\partial_n p_h]_e(\boldsymbol{\tau})(q - q_h)(\boldsymbol{\tau}) \, \mathrm{d}\boldsymbol{\tau} - \sum_{e \in \mathcal{E}_K^+} \alpha \int_e (\partial_n p_h)(\boldsymbol{\tau})(q - q_h)(\boldsymbol{\tau}) \, \mathrm{d}\boldsymbol{\tau} \Bigg).$$

Next, we take q_h equal to the image of q by a Clément type regularization operator, see [6]. We recall from [3, Chap. IX, Thms 3.7, 3.8 & Cor. 3.9] that this operator satisfies, for all K in $\mathcal{T}_h$ and e in $\mathcal{E}_K^0$ or $\mathcal{E}_K^+$

$$\|q - q_h\|_{L^2(K)} + h_K\,|q_h|_{H^1(K)} \le c\, h_K\,|q|_{H^1(\Delta_K)}, \qquad \|q - q_h\|_{L^2(e)} \le c\, h_e^{\frac{1}{2}}\,|q|_{H^1(\Delta_K)},$$

where Δ_K is the union of elements of $\mathcal{T}_h$ that intersect K. On the other hand, it is proven in [5, Lemma 5.2] that, owing to Assumption 2.3, for any q_h in $\mathbb{X}_h^*$ and K in $\mathcal{T}_h$,

$$\|q_h - \Pi_h q_h\|_{L^2(K)} \le c\, h_K\,|q_h|_{H^1(K)}.$$

Combining all this with Cauchy–Schwarz inequalities yields

$$a(p - p_h, q) \le c \sum_{K \in \mathcal{T}_h} \big(\eta_K + \varepsilon_K + h_K\,\|f - f_h\|_{L^2(K)}\big)\,|q|_{H^1(\Delta_K)},$$

whence (note that each K in $\mathcal{T}_h$ is only included in a finite number of $\Delta_{K'}$, where "finite" means bounded as a function of the regularity parameter σ of the family of triangulations)

$$a(p - p_h, q) \le c \Bigg(\sum_{K \in \mathcal{T}_h} \big(\eta_K^2 + \varepsilon_K^2 + h_K^2\,\|f - f_h\|_{L^2(K)}^2\big) \Bigg)^{\frac{1}{2}} |q|_{H^1(\Omega)}.$$

To conclude, we take q equal to $p - p_h$ and use the ellipticity property (2.6) of the form $a(\cdot, \cdot)$.

Remark 3.2. When the diffusion term is the leading one, it is a little deceipful that the term Δp_h disappears in the definition of η_K (since p_h is piecewise affine). A remedy to that is proposed in [20] for instance: It consists in introducing an auxiliary unknown by re-interpolating **grad** p_h in an enriched space. We do not present this improvement in this paper for the sake of simplicity and also because fluctuation splitting methods are specially appropriate for convection dominated equations.

Remark 3.3. When $|p_h|_{H^1(\Omega)}$ is bounded as a function of $\|f\|_{L^2(\Omega)}$ (see (2.14)), we have

$$\varepsilon_K \le c \, h_K \, \|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^\infty(K)^d}. \tag{3.4}$$

So, at least when the function $\boldsymbol{u}$ is smooth enough, the quantity ε_K is much smaller than η_K.

Proposition 3.4. *The following bound holds for each error indicator defined in* (3.1), $K \in \mathcal{T}_h$,

$$\eta_K \le c \left(\|p - p_h\|_{H^1(\omega_K)} + \sum_{\kappa \in \mathcal{T}_h, \kappa \subset \omega_K} \left(\varepsilon_\kappa + h_\kappa \, \|f - f_h\|_{L^2(\kappa)} \right) \right). \tag{3.5}$$

Proof: We now write the residual equation in a slightly modified form: For each q in $H^1_*(\Omega)$,

$$a(p - p_h, q) = \sum_{K \in \mathcal{T}_h} \left(\int_K (f - f_h)(\boldsymbol{x}) q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int_K \left(((\boldsymbol{u} - \boldsymbol{u}_h) \cdot \nabla) \, p_h \right)(\boldsymbol{x}) \, q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right.$$

$$+ \int_K \left(f_h - (\boldsymbol{u}_h \cdot \nabla) \, p_h \right)(\boldsymbol{x}) \, q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

$$\left. - \frac{1}{2} \sum_{e \in \mathcal{E}_K^0} \alpha \int_e [\partial_n p_h]_e(\boldsymbol{\tau}) q(\boldsymbol{\tau}) \, \mathrm{d}\tau - \sum_{e \in \mathcal{E}_K^+} \alpha \int_e (\partial_n p_h)(\boldsymbol{\tau}) q(\boldsymbol{\tau}) \, \mathrm{d}\tau \right).$$

We bound successively the three terms in η_K thanks to appropriate choices of q in this equation.

1) Denoting by ψ_K the bubble function on K (equal to the product of the barycentric coordinates associated with the vertices of K), we first take q equal to q_K, with

$$q_K = \begin{cases} (f_h - (\boldsymbol{u}_h \cdot \nabla) p_h) \, \psi_K & \text{in } K, \\ 0 & \text{in } \Omega \setminus K. \end{cases}$$

Since the function q_K vanishes on ∂K, we obtain

$$\|(f_h - (\boldsymbol{u}_h \cdot \nabla) p_h) \, \psi_K^{\frac{1}{2}}\|_{L^2(K)}^2 \le c \, \|p - p_h\|_{H^1(K)} \|q_K\|_{H^1(K)}$$

$$+ \|f - f_h\|_{L^2(K)} \|q_K\|_{L^2(K)} + \|((\boldsymbol{u} - \boldsymbol{u}_h) \cdot \nabla) p_h\|_{L^2(K)} \|q_K\|_{L^2(K)}.$$

Next, we use the inverse inequalities [18, Lemma 3.3], valid for any polynomial φ_K of fixed degree,

$$\|\varphi_K\|_{L^2(K)} \le c \, \|\varphi_K \, \psi_K^{\frac{1}{2}}\|_{L^2(K)}, \qquad \|\varphi_K \, \psi_K\|_{L^2(K)} \le c' \, \|\varphi_K \, \psi_K^{\frac{1}{2}}\|_{L^2(K)},$$

and also

$$\|\varphi_K \, \psi_K\|_{H^1(K)} \le c \, h_K^{-1} \, \|\varphi_K \, \psi_K\|_{L^2(K)}.$$

8

By using this in the previous inequality and multiplying it by h_K, we derive

$$h_K \,\|f_h - (\boldsymbol{u}_h \cdot \nabla) p_h\|_{L^2(K)} \le c \left(\alpha \, \|p - p_h\|_{H^1(K)} + h_K \, \|f - f_h\|_{L^2(K)} + \varepsilon_K \right). \qquad (3.6)$$

2) With each edge or face e of a triangle K, we associate a lifting operator $\mathcal{L}_{e,K}$
- which maps polynomials on e vanishing on ∂e into polynomials on K vanishing on $\partial K \backslash e$,
- which is constructed from a fixed lifting operator on the reference triangle or tetrahedron by affine transformation.

Next, for each edge or face e in $\mathcal{E}_K^0$ which is contained in another element K' of $\mathcal{T}_h$, we define the function q_e by

$$q_e = \begin{cases} \mathcal{L}_{e,\kappa}\big([\partial_n p_h]_e \, \psi_e\big) & \text{in } \kappa \in \{K, K'\}, \\ 0 & \text{in } \Omega \backslash (K \cup K'), \end{cases}$$

where ψ_e stands for the bubble function on e. Taking q equal to this function q_e in the residual equation, we obtain

$$\alpha \, \|[\partial_n p_h]_e \, \psi_e^{\frac{1}{2}}\|_{L^2(e)}^2$$
$$\le \|f_h - (\boldsymbol{u}_h \cdot \nabla) p_h\|_{L^2(K \cup K')} \|q_e\|_{L^2(K \cup K')} + c \, \|p - p_h\|_{H^1(K \cup K')} \|q_e\|_{H^1(K \cup K')}$$
$$+ \|f - f_h\|_{L^2(K \cup K')} \|q_e\|_{L^2(K \cup K')} + \|\big((\boldsymbol{u} - \boldsymbol{u}_h) \cdot \nabla\big) p_h\|_{L^2(K \cup K')} \|q_e\|_{L^2(K \cup K')}.$$

It can be checked from the construction of $\mathcal{L}_{e,\kappa}$ [18, Lemma 3.3] that it satisfies, for any polynomial φ_e of fixed degree,

$$\|\mathcal{L}_{e,\kappa}\varphi_e\|_{L^2(\kappa)} \le c \, h_e^{\frac{1}{2}} \, \|\varphi_e\|_{L^2(e)}.$$

By combining this, the previous inverse inequalities and similar ones (with ψ_K replaced by ψ_e and K replaced by e) and using (3.6), we derive

$$\alpha h_e^{\frac{1}{2}} \, \|[\partial_n p_h]_e\|_{L^2(e)} \le c \left(\|p - p_h\|_{H^1(K \cup K')} + \Big(\sum_{\kappa \in \{K, K'\}} (h_\kappa^2 \, \|f - f_h\|_{L^2(\kappa)}^2 + \varepsilon_\kappa^2)\Big)^{\frac{1}{2}} \right). \quad (3.7)$$

3) Similarly, for each edge or face e in $\mathcal{E}_K^+$, we define the function q_e^+ by

$$q_e^+ = \begin{cases} \mathcal{L}_{e,K}\big((\partial_n p_h) \, \psi_e\big) & \text{in } K, \\ 0 & \text{in } \Omega \backslash K, \end{cases}$$

and we take q equal to q_e^+ in the residual equation. Exactly the same arguments as in part 2) of the proof yield

$$\alpha h_e^{\frac{1}{2}} \, \|\partial_n p_h\|_{L^2(e)} \le c \left(\|p - p_h\|_{H^1(K)} + h_K \, \|f - f_h\|_{L^2(K)} + \varepsilon_K \right). \qquad (3.8)$$

The desired result follows from (3.6) to (3.8).

As a consequence of Propositions 3.1 and 3.4, the error $\alpha \, \|p - p_h\|_{H^1(\Omega)}$ is equivalent to the Hilbertian sum of the indicators

$$\eta_h = \Big(\sum_{K \in \mathcal{T}_h} \eta_K^2 \Big)^{\frac{1}{2}}, \qquad (3.9)$$

up to the terms involving the data and the convection vector. Moreover, the equivalence constants only depend on the regularity parameter σ of the family of triangulations. On the other hand, estimate (3.5) is local. This leads us to think that the η_K, $K \in \mathcal{T}_h$, give a good representation of the local error and thus are an efficient tool for mesh adaptivity.

Remark 3.5. It can be noted that estimate (3.3) holds when only Assumption 2.3 is satisfied. Thus, even if the convergence of the discrete solution is not proved for some algorithms, for instance for the N scheme, computing the η_K allows us to check numerically this convergence or the non-convergence of the method.

4. Numerical experiments.

We first describe the very simple strategy which is used in what follows. Let η^* be a fixed tolerance.

Initialization step: We choose an initial triangulation $\mathcal{T}_h^0$ such that

$$\left(\sum_{K \in \mathcal{T}_h^0} \left(\varepsilon_K^2 + h_K^2 \, \|f - f_h\|_{L^2(K)}^2 \right) \right)^{\frac{1}{2}} \leq \eta^*. \tag{4.1}$$

Indeed, this quantity appears in the right-hand side of (3.3) and, since the datum f and the convection vector $\boldsymbol{u}$ are known (see also (3.4)), enforcing this condition is rather easy.

Adaptation step: Assuming that the triangulation $\mathcal{T}_h^n$ is known, we compute a discrete solution p_h^n, the corresponding error indicators η_K^n and their mean value $\overline{\eta}^n$. Then, we perform mesh adaptivity in the usual way: The diameter of any element in the new triangulation $\mathcal{T}_h^{n+1}$ which contains or is contained in an element K of $\mathcal{T}_h^n$ is equal to a constant ρ times the diameter of K times the ratio $\overline{\eta}^n/\eta_K^n$. We refer to [11, Chap. 21] for the way of constructing such a mesh.

This step can be iterated until the Hilbertian sum of the indicators η_K^n becomes smaller than η^* (when possible) or a finite number of times.

The numerical experiments that follow rely on the PSI method described in Example 2.2, we refer to [1] and [8] for the choice of the coefficients μ_j^K. They are performed on the finite element code FreeFEM++, in dimension $d = 2$ for simplicity, see [12] for the detailed description of this code. We also work in the more general case of nonhomogeneous boundary conditions

$$p = p_- \quad \text{on } \Gamma_-, \qquad \alpha\, \partial_n p = r_+ \quad \text{on } \Gamma_+. \tag{4.2}$$

We refer to [18, §1.1] for the (very weak) modifications induced by these new conditions in the a posteriori analysis.

CASE OF KNOWN SOLUTIONS

We first consider equation (1.1) in the square domain $\Omega =]0, 1[^2$, with coefficients and data

$$\alpha = 1.15, \qquad \boldsymbol{u} = \begin{pmatrix} 7.5 \\ 20 \end{pmatrix}, \qquad f = 0, \tag{4.3}$$

provided with nonhomogeneous Dirichlet boundary conditions on Γ_- and homogeneous Neumann conditions on Γ_+. Moreover, the Dirichlet condition p_- is zero on the edge $x = 0$ and piecewise affine on the edge $y = 0$, with support in $[a - \delta, b + \delta]$, equal to 1 on $[a, b]$ ($a = 0.3 + 0.02\sqrt{3}$, $b = 0.6 + 0.02\sqrt{3}$, $\delta = 0.1$). The exact solution can be easily evaluated by Fourier expansion. A countour plot of this solution is given in Figure 1.
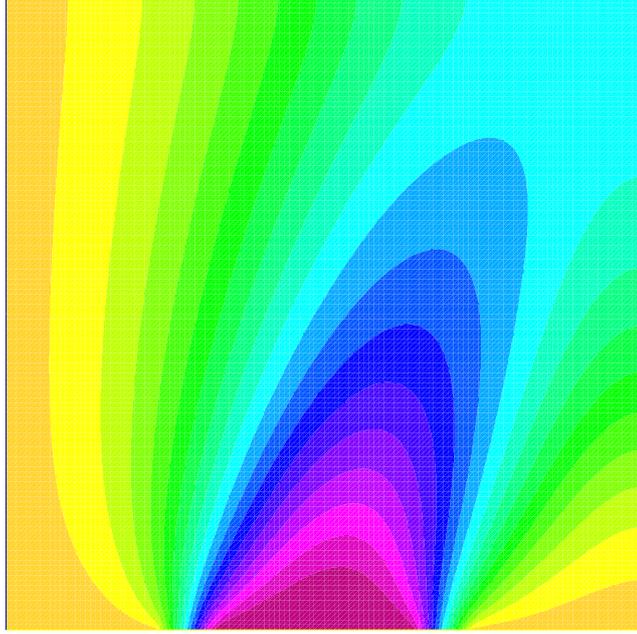
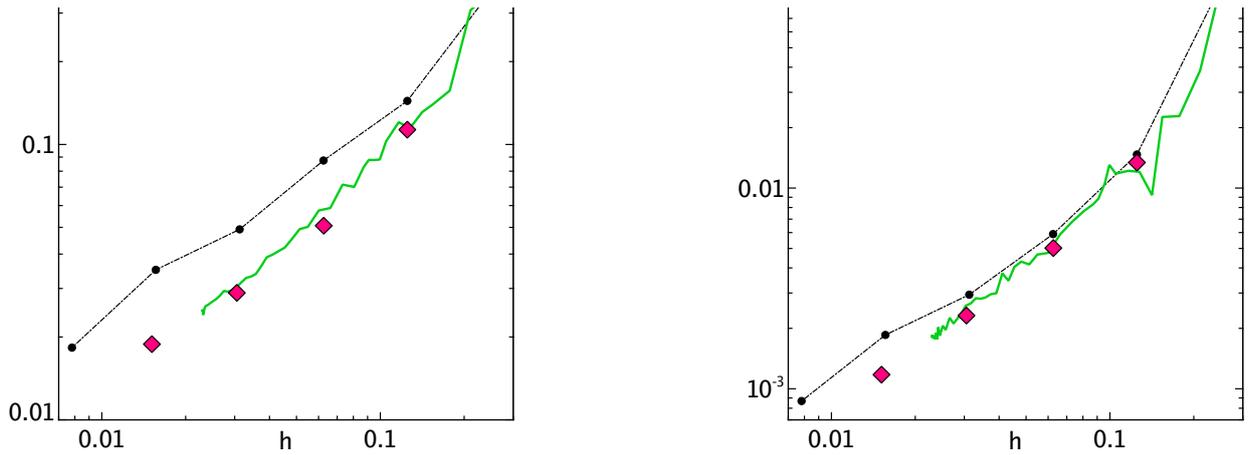**Figure 1:** The exact solution

Figure 2: The curves of errors in $H^1(\Omega)$ and $L^2(\Omega)$

In what follows, we use a new definition of h, which is more significative in the context of mesh adaptivity: h stands for the quantity $J^{-\frac{1}{2}}$, where J is the number of vertices of the triangles in $\mathcal{T}_h$ (see Section 2). Figure 2 presents the curves of the errors in $H^1(\Omega)$ (left part) and in $L^2(\Omega)$ (right part) as a function of h, in the following three situations:
1) for uniformly refined meshes (black line),
2) for adapted meshes according to the previous strategy, with $\rho = 1$ (red diamonds),
3) for adapted meshes according to the previous strategy, but now with $\rho = 0.85$ (green line).
It can be noted that mesh adaptivity is less efficient for diminishing the error in the $L^2(\Omega)$-norm than in the $H^1(\Omega)$-norm. This is in good coherence with the a posteriori analysis performed in Section 3 which only deals with the $H^1(\Omega)$-norm.

Figure 3 presents the final adapted mesh after 30 iterations of the adaptation step with $\rho = 1$. Figure 4 presents the chart of the local error divided by its mean value (left part) and of the error indicators also divided by their mean value (right part) computed

11

on this mesh.
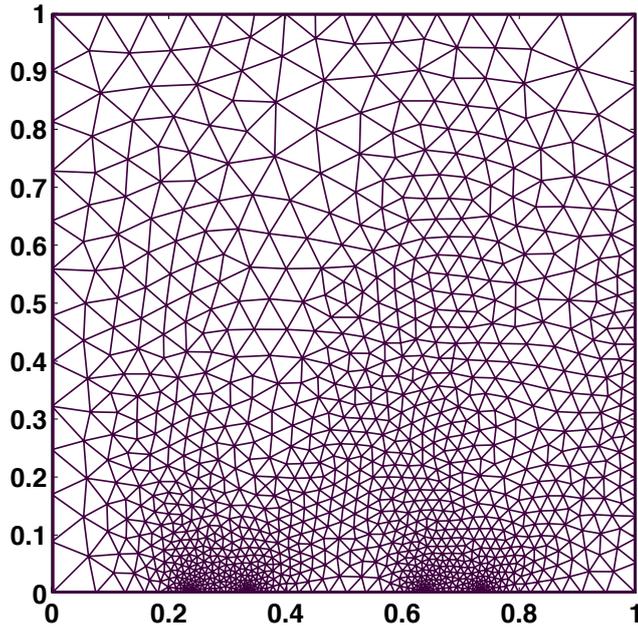
Figure 3: The final adapted mesh
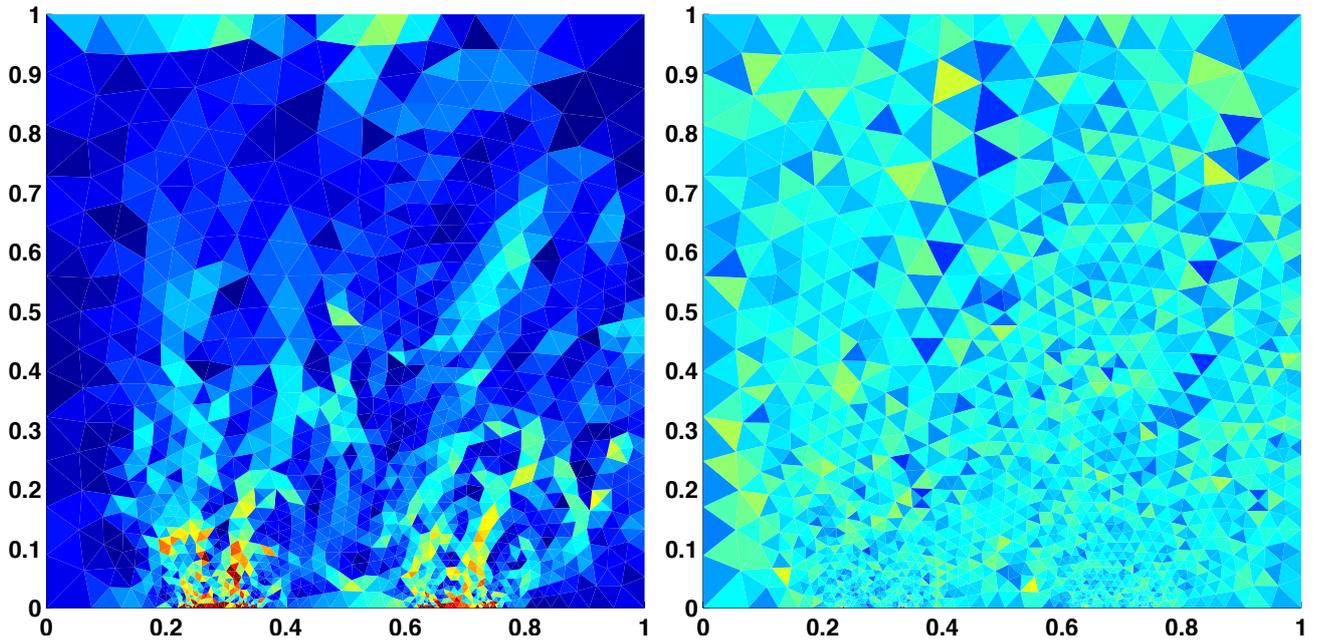
Figure 4: The local errors and error indicators

To conclude, Figure 5 presents the ratio of the standard deviation of the error indicators to their mean value as a function of the number of iterations N of the adaptation step. The fact that this ratio tends to zero when N increases proves the efficiency of our adaptivity strategy, at least in this simple case. Figure 6 presents the effectivity index, i.e., the ratio of the Hilbertian sum of the indicators defined in (3.9) to the error in $H^1(\Omega)$-norm, also as a function of the number of iterations N. Even if it highly oscillates for low values of N, it tends to a limit when N tends to infinity, which is in good coherence with

our analysis, see Propositions 3.1 and 3.4.
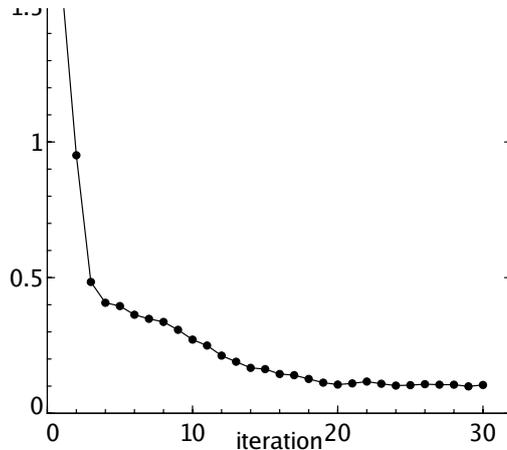
Figure 5: The deviation of the indicators with respect to their mean value
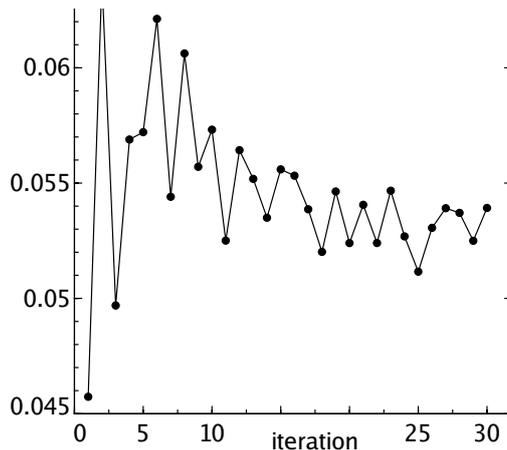
Figure 6: The effectivity index

CASE OF UNKNOWN SOLUTIONS

We now work with a test case proposed in [17]. The domain Ω is the rectangle $]-1, 1[\times]0, 1[$. The coefficients and data are given by

$$\alpha = 10^{-4}, \qquad \boldsymbol{u} = \begin{pmatrix} 2y(1-x^2) \\ -2x(1-y^2) \end{pmatrix}, \qquad f = 0, \tag{4.4}$$

so that the problem is strongly advection-dominated. This problem is provided with homogeneous Neumann boundary conditions $r_+ = 0$ on Γ_+, while the Dirichlet boundary conditions on Γ_- are given by

$$p_-(x, y) = 1 + \tanh\big(50(2x+1)\big) \qquad \text{on } \Gamma_- =]-1, 0[\times\{0\}, \tag{4.5}$$

so that a sharp front is transported by the velocity $\boldsymbol{u}$. Figure 7 presents a reference solution computed on a very fine mesh.
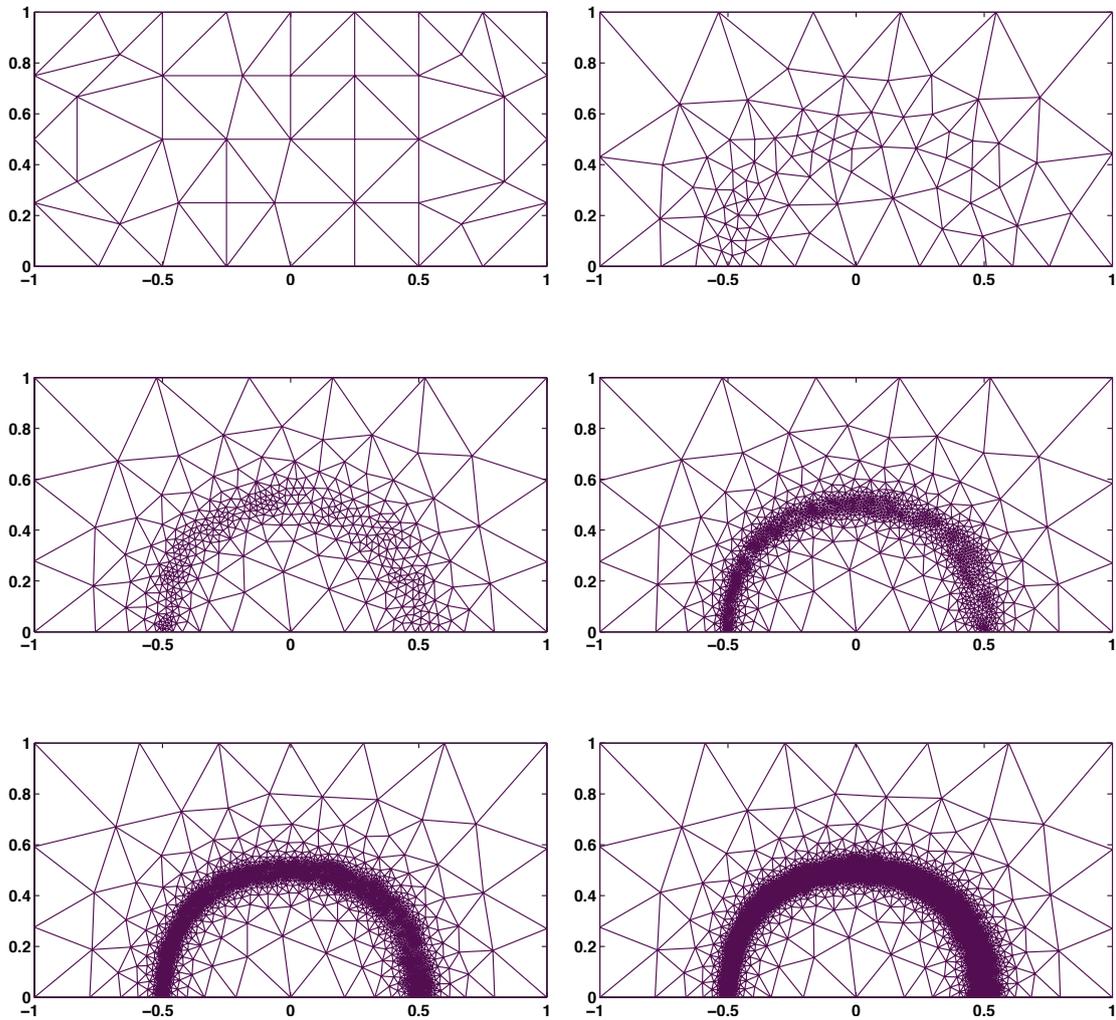
13
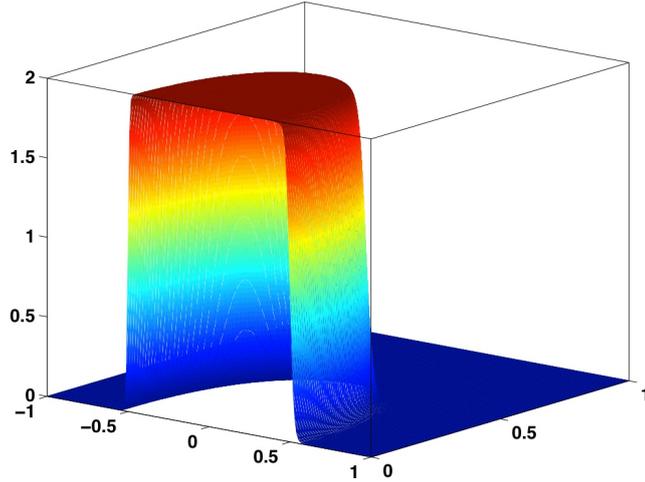
Figure 7: The reference solution

Figure 8: The adapted meshes

We now compute the triangulations $\mathcal{T}_h$ by iterating the adaptation step with the constant ρ equal to 0.9. Figure 8 presents from left to right and top to bottom the initial mesh next the resulting adapted meshes, after 5, 10, 15, 20 and 25 iterations of the adaptation step. It can be noted that the final meshes are refined exactly at the location

of the sharp front, which proves the efficiency of our strategy.

Finally, Figure 9 presents the Hilbertian sum of the indicators η_h (circles), and the errors in the norms of $H^1(\Omega)$ (stars) and $L^2(\Omega)$ (crosses) as a function of h (defined as previously as equal to $J^{-\frac{1}{2}}$) in bilogarithmic scales, all these quantities being divided by their mean value. The expected order for uniform refinement is 1 in the $H^1(\Omega)$ norm, so it seems that clustering the points around the sharp gradient region yields a higher convergence rate.
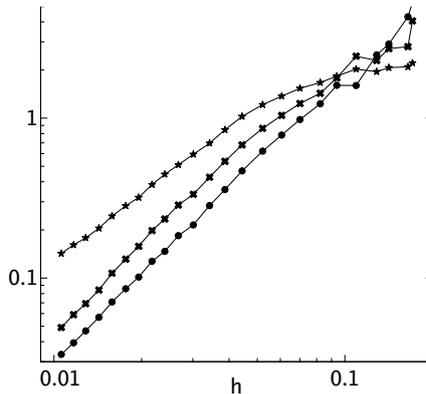
Figure 9: The Hilbertian sum of the error indicators as a function of h.

For the last test case, the domain is $[0,1]^2$ and we still use the coefficient $\alpha = 10^{-4}$ and data $f = 0$ and $r_+ = 0$. We now consider the veloicity $\boldsymbol{u} = (-\partial_y\psi, \partial_x\psi)^T$, where is the stream function ψ is given by

$$\psi(x,y) = -y - \frac{1}{3}x^2(1-2y) + \frac{1}{5\pi}x\cos(k\pi x^2)\sin(p\pi y), \qquad (4.6)$$
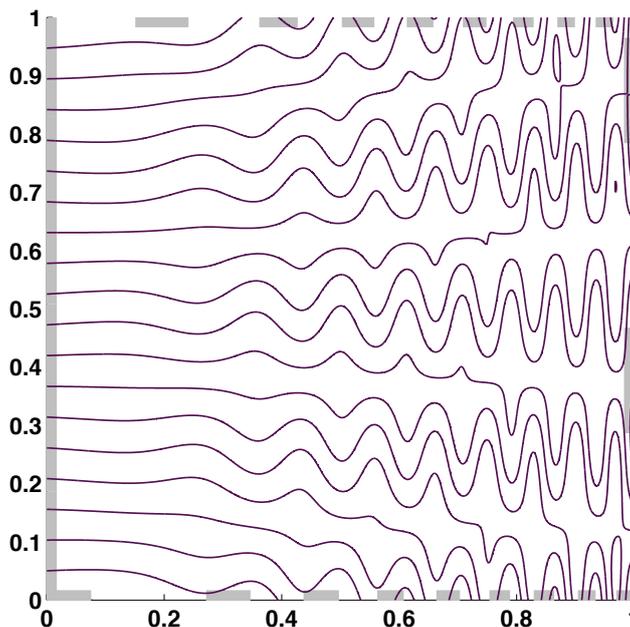
with $k = 16$ and $p = 4$.

Figure 10: The stream lines of the velocity $\boldsymbol{u}$

15

Such a velocity field is chosen in order to mimic a turbulent flow, and the corresponding stream lines are plotted in Figure 10. The Dirichlet boundary conditions on Γ_- are

$$p_-(x,y) = \begin{cases} y^5(1-y)^5, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.7}$$

Due to the choice of the velocity field, Γ_- and Γ_+ are not connected (the inflow boundary Γ_- is indicated by a grey thick line on Figure 10). Hence, to simplify the treatment of the boundary conditions, the Dirichlet boundary conditions are enforced weakly employing the method by Nitsche, for which we refer to [13] and [4], which allows the use of computational meshes such that the intersection $\overline{\Gamma}_- \cap \overline{\Gamma}_+$ does not necessarily coincide with a subset of the mesh nodes. In fact, the initial mesh considered in the mesh adaptation process uses 16 uniform subdivisions for each side of the computational domain, so that the small scale inflow/outflow pattern is not resolved.

The computed solution after 20 iterations of the adaption step is depicted in Figure 11, and the corresponding adapted grid is shown in Figure 12 (left part), together with a zoom on the right upper part of the domain (right part).
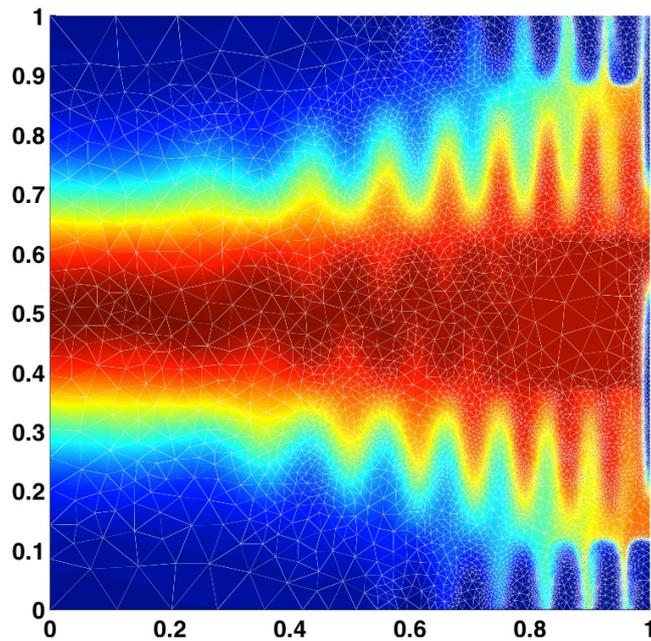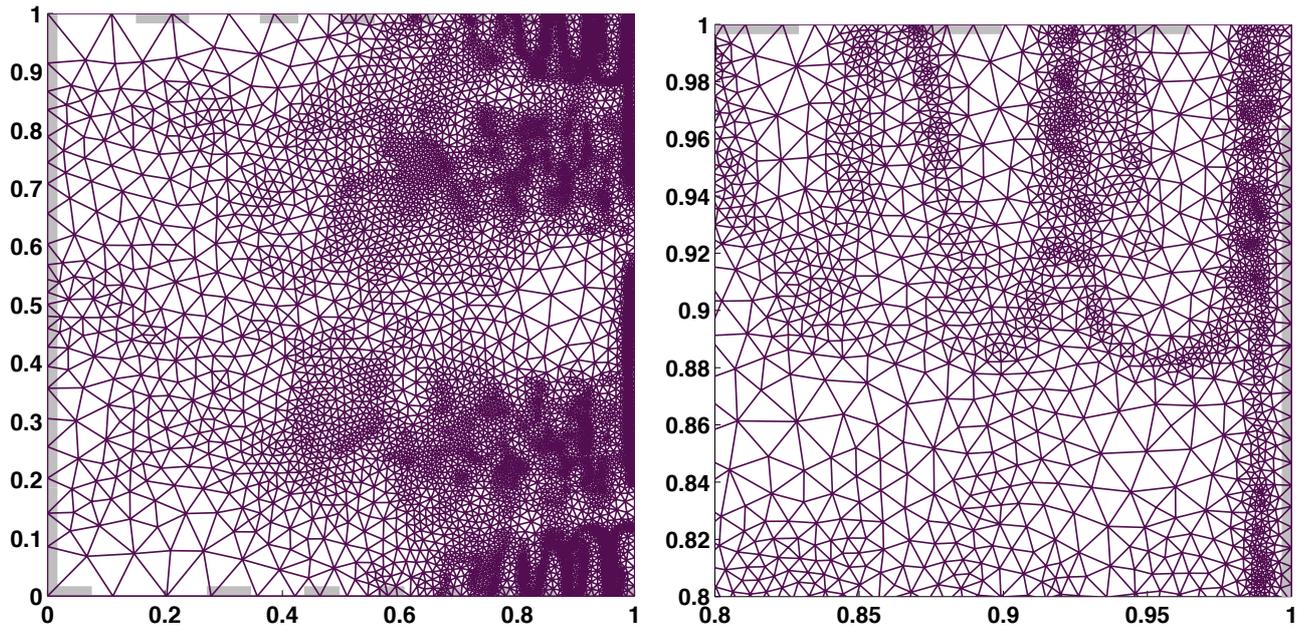

Figure 11: The computed solution

16

Figure 12: The adapted mesh

Even in these rather hard situations, the PSI method seems very efficient to solve the advection-dominated problems and is fully compatible with mesh adaptivity.

References

[1] R. Abgrall, M. Mezine — Construction of second-order accurate monotone and stable residual distribution schemes for steady problems, *J. Comput. Physics* **195** (2004), 474–507.

[2] C. Bernardi, F. Hecht, R. Verfürth — A posteriori error analysis of the method of characteristics, to appear in *Math. Models and Methods in Applied Sciences*.

[3] C. Bernardi, Y. Maday, F. Rapetti — *Discrétisations variationnelles de problèmes aux limites elliptiques*, Collection "Mathématiques et Applications" **45**, Springer-Verlag (2004).

[4] E. Burman — A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty, *SIAM J. Numer. Anal.* **43** (2005), 2012–2033.

[5] T. Chacón Rebollo, M. Gómez Mármol, G. Narbona Reina — Numerical analysis of the PSI solution of advection–diffusion problems through a Petrov–Galerkin formulation, *Math. Models and Methods in Applied Sciences* **17** (2004), 1905–1936.

[6] P. Clément — Approximation by finite element functions using local regularization, *R.A.I.R.O. Anal. Numér.* **9** R2 (1975), 77–84.

[7] B. Cockburn — An introduction to the discontinuous Galerkin method for advection-dominated problems, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, Lecture Notes in Math. **1697**, Springer (1998), pp. 151–268.

[8] H. Deconinck, R. Struijs, G. Bourgeois. P.L. Roe — Compact advection schemes on unstructued meshes, *Comput. Fluid Dynamics*, VKI Lecture Series **1993-04** (1993).

[9] B. Després — Lax theorem and finite volume schemes, *Math. Comput.* **73** (2004), 1203–1234.

[10] J. Douglas, Jr., T.F. Russell — Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures, *SIAM J. Numer. Anal.* **19** (1982), 871–885.

[11] P.J. Frey, P.-L. George — *Maillages, applications aux éléments finis*, Hermès (1999).

[12] F. Hecht, O. Pironneau — FreeFem++, see www.freefem.org.

[13] J. Nitsche — Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind, *Abh. Math. Sem. Univ. Hamburg* **36** (1971), 9–15.

[14] B. Perthame — Convergence of N-schemes for linear advection equations, *Trends in Applications of Mathematics to Mechanics*, Pitman Monogr. Surveys Pure Appl. Math. **77** (1985), pp. 323–333.

[15] O. Pironneau — On the transport-diffusion algorithm and its applications to the Navier–Stokes equations, *Numer.* Math. **38** (1982), 309–332.

[16] P.-A. Raviart, J.-M. Thomas — A mixed finite element method for second order elliptic problems, *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Mathematics **606**, Springer (1977), pp. 292–315.

[17] R.M. Smith, A. G. Hutton — The numerical treatment of advection: A performance comparison of current methods, *Numer. Heat Transfer, Part A: Applications* **5** (1982), 439–461.

[18] R. Verfürth — *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley & Teubner (1996).

[19] M. Vohralík — Residual flux-based a posteriori error estimates for finite volume and related locally conservative methods, *Numer.* Math. **111** (2008), 121–158.

[20] M. Vohralík — Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients, submitted for publication (2008).