Numerical analysis of the PSI solution of advection-diffusion problems through a Petrov-Galerkin formulation

Tomás Chacón Rebollo*, Macarena Gómez Mármol†, Gladys Narbona Reina‡

2nd April 2007

Abstract

In this paper we introduce an analysis technique for the solution of the steady advection-diffusion equation by the PSI (Positive Streamwise Implicit) method. We formulate this approximation as a non-linear finite element Petrov-Galerkin scheme, and use tools of functional analysis to perform a convergence, error and maximum principle analysis. We prove that the scheme is first-order accurate in H^1 norm, and well-balanced up to second order for convection-dominated flows. We give some numerical evidence that the scheme is only first order accurate in L^2 norm. Our analysis also holds for other non-linear Fluctuation Splitting schemes that can be built from first-order monotone schemes by the Abgrall and Mezine's technique introduced in [2].

1 Introduction

We revisit in this paper the problem set by the accurate numerical solution of flow problems in advection dominated regimes. It is well known that this is a challenging problem due to the need of combining high-order accuracy at steady state with maximum principle. Both requirements are essential to obtain numerical solutions that are useful for scientific and engineering applications.

This problem has received several ways of solution by means of methods that necessarily must be non-linear , due to Godunov's Theorem, to comply with both requirements of being of high order at steady state and verifiying the maximum principle (See Toro [32]). Let us

*Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla. C/ Tarfia, s/n. 41080 Sevilla, Spain (chacon@us.es).

†Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla. C/ Tarfia, s/n. 41080 Sevilla, Spain (macarena@us.es).

‡Departamento de Matemática Aplicada I, Universidad de Sevilla. Avda. Reina Mercedes 2. 41012 Sevilla, Spain (gnarbona@us.es).

mention the method of Characteristics (See Pironneau [24], Suli [30]), and more recently the Discontinuous Galerkin (See Cockburn [11], Cockburn and Shu [12]) and Fluctuation Splitting or Residual Distribution methods.

Fluctuation Splitting methods originated in the work by Roe and Sidilkover where the N (narrow) scheme was identified as the linear scheme with the smaller numerical diffusion for the solution of transient advection equations on rectangular grids. This scheme was extended to a large class of linear and non-linear compact schemes to solve hyperbolic systems of conservation laws by piecewise affine finite element discretizations on triangular grids by the same authors, and also by Deconinck, Struijs and co-workers (See [13], [21], [28], [29] for instance) and more recently by Abgrall and co-workers (See [1], [2] and [3]).

The convergence of the N scheme for transient linear advection schemes was analysed by Perthame and co-workers in [22] and [23], where its strong L^2 convergence was proved. This analysis is based upon an intrinsic interpretation of the N-scheme as a *finite volume scheme*, and upon the obtention of weak bounded variation estimates as for the convergence of conservation laws on triangular grids.

One of the most successful non-linear Fluctuation Splitting schemes is the PSI (Positive Streamwise Implicit) method, introduced in [13]. This is an extension of the N-scheme to second-order for steady state. It is monotone and is particularly accurate in zones of strong gradients or discontinuities of the solution.

Our purpose in this paper is to analyse the solution of the steady advection-diffusion equation when the advection operator is discretized by the PSI method and the diffusion operator is discretized by the standard Galerkin approximation. Our main contribution is to formulate this approximation as a non-linear *finite element Petrov-Galerkin scheme*, and to use the tools of functional analysis adapted to this kind of formulation to perform a convergence, error and maximum principle analysis.

We consider the advection-diffusion problem in this paper as a model problem, where we introduce the basic aspects of our analysis. This analysis may be used as a basis for several further developments. At first, to analyse the solution of unsteady convection-diffusion problems as an straightforward extension of the present analysis.

Also, the PSI method may be used in the solution of Navier-Stokes equations by piecewise affine Finite Elements, to obtain a positive solver of the convection operator. This yields a robust solver with excellent stability properties. The analysis of this solver so as some relevant numerical tests will appear in a forthcoming paper.

The paper is organised as follows. In Section 2 we set an abstract Petrov-Galerkin discretization for the advection-diffusion equation, satisfying some general hypotheses that are stated in Section 3. Section 4 proves that the PSI and other non-linear Fluctuation Splitting methods may be formulated in the abstract framework set in Section 2. Section 5 develops some technical tools that are used in following Sections to prove existence and quasi-uniqueness results for discrete problems, and in Section 7 to perform a convergence and error analysis. Section 8 is devoted to prove the maximum principle and to obtain L^r-estimates. Finally, in Section 9 some numerical evidence is given that the PSI method has an overall first order in L^r norms, while a second order well-balanced property for advection-dominated regimes is proved.

2 The advection-diffusion problem

In this section we introduce an approximation of the stationary advection-diffusion problem by a non-linear Petrov-Galerkin Finite Element method. This will be an abstract discrete variational formulation for the PSI (Positive Streamwise Implicit) method.

Let Ω be a bounded domain of $\mathbb{R}^d$ ($d = 2$ or 3) with Lipschitz boundary Γ. We consider a measurable subset Γ^- of Γ with non-zero measure, and set $\Gamma^+ = \Gamma \setminus \Gamma^-$. Denote by n the unit normal to Γ, outer to Ω. We consider the following stationary advection-diffusion problem:

$$\left\{ \begin{array}{rl} u \cdot \nabla \rho - \nu \Delta \rho = f & \text{in } \Omega \\ \rho = g & \text{on } \Gamma^- \\ \nu \dfrac{\partial \rho}{\partial n} = 0 & \text{on } \Gamma^+, \end{array} \right. \tag{1}$$

where ρ is a physical magnitude (a "tracer ") transported by the velocity field $u : \bar{\Omega} \to \mathbb{R}^d$, and ν is the diffusion coefficient of ρ. Also, $f : \bar{\Omega} \to \mathbb{R}^d$ is the source term and g is the Dirichlet data for ρ on the inflow boundary Γ^-.

We assume that the velocity field satisfies $u \in L^q(\Omega)^d$ for some $q > d$ and $\nabla \cdot u = 0$. Then, the trace on Γ of the normal velocity $u \cdot n$ belongs to $L^r(\Gamma)^d$, with $r = q(d-1)/d > 1$. This is proved by a duality argument based upon Sobolev's injections. This allows us to formalize the meaning of "inflow boundary" as $u \cdot n$ is defined a. e. on Γ. Specifically, we assume

$$u \cdot n < 0 \quad \text{a. e. on } \Gamma^-, \quad \text{and } u \cdot n \geq 0 \quad \text{a. e. on } \Gamma^+.$$

We define the space

$$V = \{v \in H^1(\Omega) / v_{|\Gamma^-} = 0\},$$

and consider the variational formulation of problem (1),

$$\text{Obtain } \rho \in G + V \text{ such that } \quad a(\rho, v) = \langle f, v \rangle \ \ \forall\, v \in V, \tag{2}$$

where $G \in H^1(\Omega)$ is some lifting of g and $a : H^1(\Omega) \times H^1(\Omega) \mapsto \mathbb{R}$ is the bilinear form

$$a(w, v) = \int_\Omega (u \cdot \nabla w)\, v + \nu \int_\Omega \nabla w \cdot \nabla v. \tag{3}$$

Under the above hypotheses on u, problem (2) admits a unique solution in $H^1(\Omega)$ if $g \in H^{1/2}(\Gamma^-)$ and $f \in V'$, by Nečas Lemma. (See Ern & Guermond [16].)

We may assume $g = 0$ up to an additive changement of the source term f, and we shall assume it so, without loss of generality. In this case the solution ρ belongs to V.

To approximate problem (2), let us assume Ω to be a poligonal domain. Consider a triangulation $\mathcal{T}_h$ of Ω by triangles in 2D and tetrahedra in 3D. As usual we assume that h denotes the largest diameter of the elements of $\mathcal{T}_h$. Consider the finite dimensional spaces of piecewise affine finite elements built on $\mathcal{T}_h$:

$$V_h^* = \{v_h \in \mathcal{C}^0(\bar{\Omega}) / \ v_{h_{|T}} \in \mathbb{P}_1 \ \forall\, T \in \mathcal{T}_h\}, \quad V_h = \{v_h \in V_h / \ v_h = 0 \quad \text{on } \Gamma^-\}. \tag{4}$$

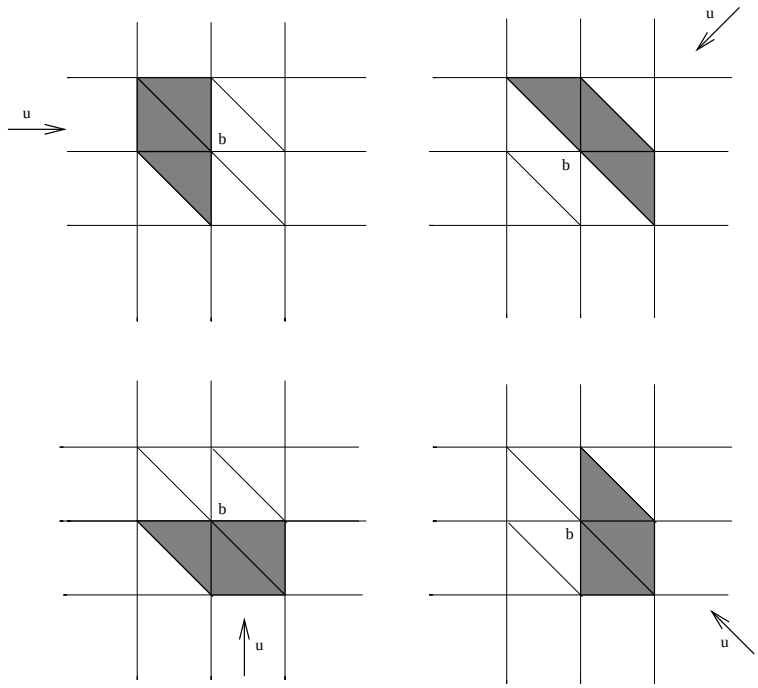

Figure 1: Typical supports of the basis functions $\lambda_j(s_h)$.

Denote by $\{b_j\}_{j=1}^{M}$ the nodes of the mesh located on $\overline{\Omega} \setminus \Gamma^-$ and by $\{b_j\}_{j=M+1}^{N}$ those located on Γ^-. We consider the nodal basis functions of V_h^*, $\{\varphi_i\}_{i=1}^{N}$ defined by

$$\varphi_i(b_j) = \delta_{ij}, \quad 1 \leq i, j \leq N.$$

The nodal base of V_h is then $\{\varphi_i\}_{i=1}^{M}$

We also associate to $\mathcal{T}_h$ and a given element s_h of V_h^* a discrete space of piecewise constant functions, denoted by $W_h^*(s_h)$. This space is defined through its *nodal* basis functions $\lambda_1, \lambda_2, \cdots, \lambda_N$ (also depending on s_h), that we assume known for the time being:

$$W_h^*(s_h) = span\{\lambda_1(s_h), \lambda_2(s_h), \cdots, \lambda_N(s_h)\}. \tag{5}$$

For simplicity of notation, we do not explicit the dependence upon s_h of the λ_i when this is not source of confusion.

We shall use the functions of $W_h^*(s_h)$ to test the advection operator in equation (8). The functions λ_j have supports that look for "upwind" information with respect to the velocity field u (See Fig. 1).

The dependence upon s_h is due to the non-linear nature of the PSI method, which appears under this formulation as the dependency of the constant values that the basis functions λ_i take on each triangle T of $\mathcal{T}_h$: $\lambda_{i|_T} = \lambda_{i|_T}(s_h)$. In Section 4 we give the definition of the λ_j for PSI and actually other fluctuation splitting methods, although for our analysis we shall only consider some abstract properties that we describe next.

We construct an associated interpolation operator taking values on W_h^*, denoted by Π_{s_h}, as follows:

$$\begin{aligned} \Pi_{s_h} : \mathcal{C}^0(\overline{\Omega}) &\longrightarrow W_h^* \\ z &\longrightarrow \Pi_{s_h} z = \sum_{i=1}^{N} z(b_i)\lambda_i. \end{aligned} \tag{6}$$

Note that if $z \in V_h$, then

$$\Pi_{s_h} z = \sum_{i=1}^{M} z(b_i)\lambda_i \ \in W_h = span\{\lambda_1, \lambda_2, \cdots, \lambda_M\}.$$

Note also that $\lambda_j = \Pi_{s_h}\varphi_j,\ j = 1, 2, \cdots, N$. Thus, Π_{s_h} is bijective from V_h^* onto W_h^* and also from V_h onto W_h.

We shall refer to Π_{s_h} as the Distributed Interpolation operator generated by function s_h. We may characterize each actual Fluctuation Splitting method by its associated Distributed Interpolation operator, through the definition of the basis functions λ_j.

We define the bilinear form $a_h : V_h \times V_h \mapsto \mathbb{R}$ as

$$a_h(\rho_h, v_h) = \int_\Omega (u \cdot \nabla\rho_h)\, \Pi_{\rho_h} v_h + \nu \int_\Omega \nabla\rho_h \cdot \nabla v_h. \tag{7}$$

We may now formulate our discrete variational approximation of the advection-diffusion problem (1), as follows:

$$\begin{cases} \text{Find } \rho_h \in V_h \text{ such that} \\ a_h(\rho_h, v_h) = \langle f,\, v_h\rangle \qquad \forall\ v_h \in V_h. \end{cases} \tag{8}$$

To obtain a more accurate scheme for advection-dominated flows, we may also upwind the source term. This is obtained in a natural way, as follows:

$$\begin{cases} \text{Find } \rho_h \in V_h \text{ such that} \\ a_h(\rho_h, v_h) = \langle f,\, \Pi_{\rho_h} v_h\rangle \quad \forall\ v_h \in V_h. \end{cases} \tag{9}$$

The term $\langle f,\, \Pi_{\rho_h} v_h\rangle$ makes sense for smoother f than $f \in V'$, for instance $f \in L^1(\Omega)$.

3 Hypotheses

We next state the general hypotheses that we assume about our approximation, to perform our error analysis. We prove in Section 4 that the PSI method, and some other non-linear FS schemes, when applied to the solution of the advection-diffusion (1), may be cast as Petrov-Galerkin methods (8) or (9), verifying these hypotheses.

We at first assume the hypotheses related to the approximation properties of the interpolation operator Π_{s_h}:

Hypothesis 1 *For any element $T \in \mathcal{T}_h$,*

1. $\lambda_{i_T}^T \geq 0, \quad i = 1, \cdots, d+1,$

2. $\displaystyle\sum_{i=1}^{d+1} \lambda_{i_T}^T = 1,$ $\square$

where i_T is the global index corresponding to the local index i , $i = 1, \cdots, d+1$ on element T, and $\lambda_{i_T}^T$ is the restriction of λ_{i_T} to element T.

These properties yield the stability of the Distribution Interpolation operator (See Section 5).

We next consider additional hypotheses required to obtain the maximum principle. For the discrete problems (8) and (9), we understand this principle in the following sense:

$$\textit{Maximum Principle:} \qquad If \;\; f \geq 0 \quad in \;\; \Omega, \quad then \;\; \rho_h \geq 0 \quad in \;\; \overline{\Omega}.$$

Several authors have performed an extensive analysis about sufficient conditions on the grid that yield the maximum principle for piecewise affine finite element discretizations of elliptic equations, in two and three space dimensions (See Drăgănescu, Dupont and Scott [15] and references therein, Varga [33]). These conditions ensure that the matrix associated to the discrete problem is a monotone matrix, i. e., it is non-singular and its inverse has non-negative entries.

Notice that the matrix $A_\nu = A_\nu(s_h)$ associated to problems (8) and (9) is defined as:

$$A_\nu(s_h) = C(s_h) + \nu\, L,$$

where $C(s_h)$ and L, respectively, are the matrices associated to the discretization of the advection and the Laplacian operators. Their coefficients are given by

$$C_{ij}(s_h) = \int_\Omega (u \cdot \nabla \varphi_j)\, \Pi_{s_h}\, \varphi_i = \int_\Omega (u \cdot \nabla \varphi_j) \lambda_i(s_h),$$

$$L_{ij} = \int_\Omega \nabla \varphi_j \cdot \nabla \varphi_i \quad i, j = 1, \cdots, M.$$

Our analysis of the maximum principle is based on a property stronger than monotonicity: We assume $A_\nu(\rho_h)$ to be an M-matrix. We recall that A is an M-matrix if

1.
$$A_{ii} > 0, \quad \forall\, 1 \leq i \leq M. \tag{10}$$

2.
$$A_{ij} \leq 0, \quad \forall\, 1 \leq i, j \leq M,\, i \neq j. \tag{11}$$

3.
$$A_{ii} \geq \sum_{j \neq i} |A_{ij}|, \quad \forall\, 1 \leq i \leq M \quad \text{i.e., A has diagonal dominance.} \tag{12}$$

We shall assume the following hypothesis:

Hypothesis 2
$A_\nu(s_h)$ is a M-matrix, for any $s_h \in V_h$ and any $\nu > 0$, □

To ensure that the discrete Laplacian L is an M-matrix it is enough to ask that all angles between sides (d=2) or all dyhedric angles between faces (d=3) of all elements of the grid are acute or, at most, of $\pi/2$ degrees. This is a classical result that may be found for instance in [15].

However, building a advection-diffusion matrix A which is a M-matrix for any value of the diffusion coefficient ν is an achievement of FS methods. We shall prove this in Section 5.

We finally state a technical hypothesis on the dependence of the behaviour of the convective matrix with respect to its argument s_h:

Hypothesis 3
The convective matrix $C(s_h)$ is a continuous function from V_h onto the space of square real matrices of dimension $M \times M$. □

This Hypothesis is mainly related to the boundedness of the basis functions λ_j. It is verified by the PSI scheme and other non-linear FS schemes, but not by positive first-order FS methods, in particular by the N-scheme. This agrees with the interpretation of the N-scheme as a finite volume scheme by Perthame in [23].

4 Relationship to Fluctuation Splitting methods

This Section is devoted to prove that methods (8) and (9) are abstract formulations for PSI and other non-linear FS methods applied to the numerical solution of the advection-diffusion problem (1), and that under this characterization, Hypotheses 1, 2 and 3 of Section 3 are verified.

Let us consider that the convective flow of the tracer ρ transported by the velocity field u is

$$q_{conv} = u\rho.$$

The total balance of convective flow ("fluctuation") through ∂T for some $T \in \mathcal{T}_h$ is then

$$\Phi^T = \int_{\partial T} q_{conv} \cdot n = \int_T \nabla \cdot (u\rho) = \int_T u \cdot \nabla \rho,$$

where we have used that $\nabla \cdot u = 0$.

The basic idea of FS methods is to split the fluctuation Φ^T between the vertex neighbouring the element T. This is made by means of some "flux distribution coefficients" $\{\beta_j^T\}_{j=1}^N$. The fluctuation contribution to vertex b_j is

$$\Phi_j = \sum_{T \in \mathcal{T}_h} \beta_j^T \, \Phi^T. \tag{13}$$

To increase the compactness of the numerical scheme, usually the fluctuation generated in an element T is only sent to the vertex of T:

$$\beta_j^T = 0 \quad \text{if } b_j \quad \text{is not a vertex of } T.$$

This yields the following discretization for the steady advection equation :

$$\sum_{T \in E_j} \beta_j^T \int_T u \cdot \nabla \rho = \sum_{T \in E_j} \beta_j^T \int_T f, \tag{14}$$

where E_j is the set of elements of $\mathcal{T}$ that share the vertex b_j.

For a given vertex b_j of the triangulation $\mathcal{T}_h$, we define the asociated piecewise constant basis function λ_j by

$$\lambda_j : \overline{\Omega} \mapsto \mathbb{R}, \quad \text{with} \quad \lambda_j = \beta_j^T \quad \text{on element } T, \quad \forall\, T \in \mathcal{T}_h$$

Then, the l.h.s. of (14) is re-written as

$$\sum_{T \in E_j} \beta_j^T \int_T u \cdot \nabla \rho = \int_\Omega u \cdot \nabla \rho \, \lambda_j,$$

and we recover our Petrov-Galerkin discretization of the advection operator, for the Distribution Interpolation operator associated to these actual λ_j.

The discretization of the source term in (14) is also re-written as

$$\sum_{T \in E_j} \beta_j^T \int_T f = \int_\Omega f \, \lambda_j.$$

To obtain a flux-conservative and a L^∞-stable method, the following properties are usually assumed (Cf. [2]):

$$\sum_{j=1}^{d+1} \beta_{jT}^T = 1, \quad \beta_{jT}^T \geq 0, \text{ for any node } b_j, \text{ for any triangle } T \in \mathcal{T}_h. \tag{15}$$

This trivially yields our first hypothesis:

Lemma 4.1 *Assume the distribution coefficients verify properties (15). Then, Hypothesis 1 holds.*

□

Our proof that Hypotheses 2 and 3 hold for PSI scheme is based upon the construction of this scheme introduced in Abgrall and Mezine [2]. In its turn, this construction is based upon that of N scheme, which is introduced via the fluctuations

$$\Phi_i^T(s_h) = \beta_i^T \, \Phi^T(s_h), \quad \text{for } s_h \in V_h$$

sent to the local nodes. To define the Φ_i^T let us introduce the inward normal vector to the boundary of T, opposite to node b_i,

$$n_i^T = d\,|T|\,\nabla\varphi_{iT}, \tag{16}$$

and the values

$$K_i^T = \frac{1}{d}\,\bar{u}^T \cdot n_i^T, \quad \text{with} \quad \bar{u}^T = \frac{1}{|T|}\int_T u.$$

Then, the N scheme, when $d = 2$ for simplicity, is given by

$$\Phi_i^T(s_h) = \sum_{i=1}^{3} c_{ij}^T\,(s_i^T - s_j^T), \quad c_{ij} = (K_i^T)^+\,M^T(K_j^T)^-, \tag{17}$$

$$\text{where } (K_i^T)^+ = \max\{K_i^T, 0\}, \quad (K_j^T)^- = \min\{K_j^T, 0\}, \quad M^T = \sum_{j=1}^{3}(K_j^T)^-,$$

and s_1^T, s_2^T, s_3^T are the values of s_h at the vertex of T (we use local notation for the index to simplify the notation).

Notice also that M^T is non-zero (in fact, it is strictly negative) if $\overline{u} \neq 0$ as $\displaystyle\sum_{i=1}^{3} K_i^T = 0$. Also, that $c_{ij} \geq 0$. This property (called monotonicity) yields the L^∞ stability of the N scheme for evolution systems of conservation laws, under a CFL condition (Cf. [2]).

Note that the sign of K_i^T indicates wether the node b_i is upflow with respect to u in triangle T ($K_i^T \leq 0$). The nodes located upflow cannot receive any fluctuation contribution for FS schemes to be stable.

In some cases, the fluctuation $\Phi^T(s_h)$ could vanish while the partial fluctuation $\Phi_i^T(s_h)$ remains finite, and then the coefficient β_i^T is not defined. For this reason, the N scheme does not enter in the framework of our Petrov-Garlekin formulation.

The PSI scheme is constructed in [2] by looking at new fluctuations Φ_i^* (we drop the superscript T and the explicit dependence upon s_h for simplicity) such that

$$\Phi_i^* = (1 - \mu_i)\,\Phi_i, \text{ for some coefficients } 0 \leq \mu_i \leq 1,\ i = 1, \cdots, d+1 \tag{18}$$

$$\sum_{i=1}^{d+1} \Phi_i^* = \Phi. \tag{19}$$

$$\text{The coefficients } \beta_i^* = \frac{\Phi_i^*}{\Phi} \quad \text{are bounded independently of the mesh.} \tag{20}$$

This problem is solved in [2] by means of a constructive solution. It is proved *that the μ_i are continuous functions of the Φ_i, and that the $\beta_i^* \in [0, 1]$.*

Note that the β_i^* are functions of s_h, as both Φ^* and Φ depend on s_h. Moreover, from (19) and the property $\beta_i^* \in [0, 1]$ we recover property (15):

$$\sum_{i=1}^{3} \beta_i^* = 1, \quad \beta_i^* \geq 0.$$

We do not detail the actual construction of the PSI scheme as the solution of problem (18)-(20) made in [2], as it is rather lengthy, and is not essential for our analysis.

The PSI scheme is now cast as a non-linear Petrov-Galerkin scheme with the structure (8) or (9) , following the derivation of the first part of this Section. We next prove that it satisfies Hypothesis 2 as follows:

Lemma 4.2 *Assume that all angles between sides (when $d = 2$) or that all dihedral angles between faces (when $d = 3$) of all elements of the grid are of, at most, $\pi/2$ degrees.*

Then the PSI scheme satisfies Hypothesis 2 for any $\nu > 0$.

Proof: Under the above hypotheses, the discrete Laplace matrix L is a M-matrix (Cf., for instance, [15]).

To prove that the discrete advection-diffusion matrix $A^*_\nu(s_h) = C^*(s_h) + \nu\, L$ associated to the PSI method is an M-matrix for any $\nu > 0$ for some $s_h \in V_h$, it is enough to prove that the associated discrete advection matrix C^* verifies (10), (11) and (12) but with non-strict inequalities.

By (18), the elementary coefficients of C^* on a triangle T of $\mathcal{T}_h$ are obtained from those of the N scheme c^T_{ij} by

$$(C^T_{ij})^* = -(1-\mu^T_i)\, c^T_{ij}, \quad \text{if} \quad i \neq j,\ (C^T_{ii})^* = (1-\mu^T_i)\, c^T_{ii}.$$

As $c^T_{ij} \geq 0$ and $0 \leq \mu^T_i \leq 1$ the conclusion follows. □

Remark 4.1 *It does not seem clear that the discrete convection matrix $C^*(s_h)$ also is positive-defined. The global diagonal element for the PSI scheme is given by*

$$\begin{aligned} C^*_{kk} &= \sum_{T\in E_k} (C^T_{kk})^* = \sum_{T\in E_k} (1-\mu^T_k)\, c^T_{kk} \\ &= \sum_{T\in E_k,\, K^T_k \leq 0} (1-\mu^T_k)\, c^T_{kk} + \sum_{T\in E_k,\, K^T_k > 0} (1-\mu^T_k)\, c^T_{kk}. \end{aligned}$$

If b_k is an upstream node for an element T (i. e., if $K^T_k = \frac{1}{2}\,\bar{u}^T \cdot n^T_k \leq 0$), then $c^T_{kk} = \beta^T_k\, K^T_k = 0$, and so

$$C^*_{kk} = \sum_{T\in E_k,\, K^T_k > 0} (1-\mu^T_k)\, c^T_{kk}.$$

Given a triangle $T \in E_k$, denote by Γ_k the side or face of ∂T opposite to the node b_k, and by η^T_k the unit normal to ∂T on Γ_k, outward to T. Using the vector function $\sigma(x) = x - b_k$ we deduce the relation

$$n^T_k = -\eta^T_k\, |\Gamma_k|,$$

where n^T_k is the inward normal defined in (16).

As $\nabla \cdot u = 0$, then

$$0 = \int_{\bigcup_{T\in E_k}} u \cdot \eta = \sum_{T\in E_k} \int_{\Gamma_k} u \cdot \eta^T_k = \sum_{T\in E_k} \hat{u}^T \cdot n^T_k = d \sum_{T\in E_k} \hat{K}^T_k$$

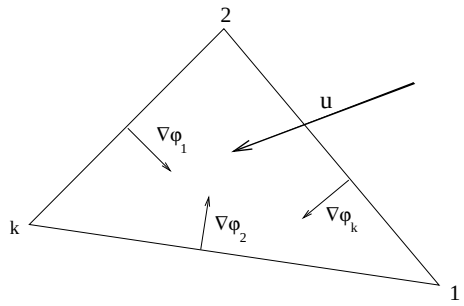

Figure 2: One target triangle: $\nabla\varphi_i \cdot u \leq 0$, $i = 1, 2$, $\nabla\varphi_k \cdot u > 0$.

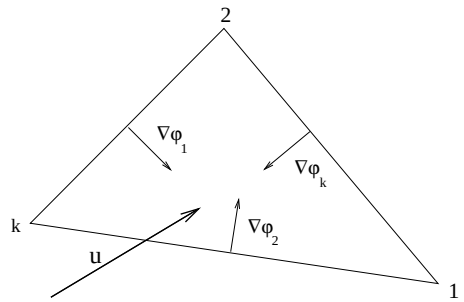

Figure 3: Two target triangle: $\nabla\varphi_i \cdot u > 0$, $i = 1, 2$, $\nabla\varphi_k \cdot u \leq 0$.

where $\hat{u}^T = \dfrac{1}{|\Gamma_k|} \displaystyle\int_{\Gamma_k} u$, *and* $\hat{K}_k^T = \dfrac{1}{d}\, \hat{u}^T \cdot n_k^T$.

Then (excepting very particular grids that can be avoided), for some $T \in E_k$ *we have* $\hat{K}_k^{\tilde{T}} > 0$. *If* u *is smooth and* h *is small enough, this implies* $K_k^{\tilde{T}} > 0$, *and then* $c_{kk}^{\tilde{T}} > 0$.

However, we could have $\mu_k^{\tilde{T}} = 1$ *if all the flux is sent to node* i. *This occurs with PSI scheme in the case of "two target triangle" (or three if* $d = 3$*), as it sends all the flow to one of the targeted vertex (See Figures 2 and 3).*

Therefore, we could have $C_{kk}^* = 0$ *if the node* b_k *is downstream just for one triangle of* E_k, *for which it is the target of a two or more target element.*

We also have:

Lemma 4.3 *The PSI scheme satisfies Hypothesis 3*

Proof: The construction of [2] of PSI method yields continuous elementary coefficients μ_i^T as functions of the fluxes Φ_i^T. As these are continuous functions of s_h by (17), we conclude that Hypothesis 3 holds. □

Remark 4.2 *The construction of [2] allows to build a scheme satisfying properties (18), (19) and (20) from any first-order monotone scheme, and not only from the N scheme.*

In particular, in [2] there are reported the L-Rusanov and the L-upwind, respectively constructed from the Rusanov and the one-dimensional upwind schemes. Thus, the same proofs above prove that these schemes also satisfy Hypotheses 1, 2 and 3.

Consequently, all results of our analysis also apply to the the L-Rusanov and the L-upwind schemes and to any other that could be built by Abgrall and Mezine's technique from a first-order scheme.

5 Tools

Our analysis is based upon some functional properties of the Distribution Interpolation operator and the discrete advection-diffusion operator appearing in our Petrov-Galerkin methods (8) and (9), that we state in this Section. These properties are deduced from Hypotheses 1 and 2 stated above.

We denote by $\|\cdot\|_p$ the $L^p(\Omega)$ norm.

Lemma 5.1 *There exists a constant $C_p > 0$ such that*

$$\|\Pi_{r_h} v_h\|_p \leq C_p \|v_h\|_p \qquad \forall\, r_h\, v_h \in V_h,\, \forall\, 1 \leq p < +\infty.$$

Proof:

For simplicity of notation, we drop the subindex h of v_h.

Step 1. Consider a function $v \in V_h$. By definition, we have:

$$\|\Pi_{r_h} v\|_p^p = \int_\Omega \left| \sum_{i=1}^{N} v(b_i)\lambda_i \right|^p = \sum_{T\in\mathcal{T}_h} \int_T \left| \sum_{i=1}^{d+1} v(b_{i_T})\lambda_{i_T}^T \right|^p,$$

where $\lambda_i = \lambda_i(r_h)$. We may write the norm in $L^p(\Omega)$ of v as:

$$\|v\|_p^p = \sum_{T\in\mathcal{T}_h} \int_T \left| \sum_{i=1}^{d+1} v(b_{i_T}) B_i(x) \right|^p dx,$$

where $B_i(x)$, $1 \leq i \leq d+1$ are the baricentric coordinates of x in the element T.

First, we are going to prove that there exists a constant C_p^* such that:

$$\left(\int_T \sum_{i=1}^{d+1} |v(b_{i_T})|^p dx \right)^{1/p} \leq C_p^* \left(\int_T \left| \sum_{i=1}^{d+1} v(b_{i_T}) B_i(x) \right|^p dx \right)^{1/p}. \tag{21}$$

To see it, consider a change of variables from the reference element $\hat{T}$ into T. If we take $\hat{x} \in \hat{T}$, we define $x \in T$ such that $x = A_T\hat{x} + b_T$, then $dx = |\det A_T| d\hat{x}$.

Furthermore, we can write $|T| = \displaystyle\int_T dx = |\det A_T||\hat{T}|$, so $|\det A_T| = \dfrac{|T|}{|\hat{T}|}$.

Define the seminorms $|\cdot|_1$ and $|\cdot|_2$ on the space $\mathcal{C}^0(T)$ as follows:

$$|v|_1 = \left(\int_T \left| \sum_{i=1}^{d+1} v(b_{i_T}) B_i(x) \right|^p dx \right)^{1/p};$$

$$|v|_2 = \left(\int_T \sum_{i=1}^{d+1} |v(b_{i_T})|^p dx \right)^{1/p} = |T|^{1/p} \left(\sum_{i=1}^{d+1} |v(b_{i_T})|^p \right)^{1/p}.$$

If we apply the change of variables in the definition of $|v|_1$ we have:

$$|v|_1 = \left(\int_{\hat{T}} |\det A_T| \left| \sum_{i=1}^{d+1} v(b_{i_T}) \hat{B}_i(\hat{x}) \right|^p d\hat{x} \right)^{1/p} = \frac{|T|^{1/p}}{|\hat{T}|^{1/p}} \left(\int_{\hat{T}} \left| \sum_{i=1}^{d+1} v(b_{i_T}) \hat{B}_i(\hat{x}) \right|^p d\hat{x} \right)^{1/p}.$$

We next define two norms on $\mathbb{R}^{d+1}$. Given $z = (z_1, \ldots, z_{d+1}) \in \mathbb{R}^{d+1}$ we define:

$$|z|_1^* = \frac{1}{|\hat{T}|^{1/p}} \left(\int_{\hat{T}} \left| \sum_{i=1}^{d+1} z_i \hat{B}_i(\hat{x}) \right|^p d\hat{x} \right)^{1/p};$$

$$|z|_2^* = \left(\sum_{i=1}^{d+1} |z_i|^p \right)^{1/p}.$$

As we have a finite dimensional space, the norms are equivalent, so there exists a constant C_p^* such that

$$|z|_2^* \leq C_p^* |z|_1^*, \quad \forall\, z \in \mathbb{R}^{d+1}.$$

If we consider now $v_T = (v(b_{1_T}), \ldots, v(b_{(d+1)_T}))^t$ as a vector of $\mathbb{R}^{d+1}$, and we use the definitions of $|v_T|_1$ and $|v_T|_2$, we deduce (21).

Step 2. Let us conclude that $\|\Pi_h v\|_p \leq C_p^* \|v\|_p$.

$$\|\Pi_{r_h} v\|_p^p = \sum_{T \in \mathcal{T}_h} \int_T \left| \sum_{i=1}^{d+1} v(b_{i_T}) \lambda_{i_T}^T \right|^p \leq \sum_{T \in \mathcal{T}_h} \int_T \left| \sum_{i=1}^{d+1} |v(b_{i_T})| \lambda_{i_T}^T \right|^p.$$

But by *Hypothesis 1* , we have $\lambda_{i_T}^T \geq 0$, $\forall\ 1 \leq i \leq d+1$ and $\sum_{i=1}^{d+1} \lambda_{i_T}^T = 1$. As the function $g(x) = x^p$ with $p \geq 1$ is convex, then

$$\|\Pi_{r_h} v\|_p^p \leq \sum_{T \in \mathcal{T}_h} \int_T \sum_{i=1}^{d+1} |v(b_{i_T})|^p \lambda_{i_T}^T \leq \sum_{T \in \mathcal{T}_h} \int_T \sum_{i=1}^{d+1} |v(b_{i_T})|^p,$$

because each $\lambda_{i_T}^T \leq 1$.

Thus $\|\Pi_{r_h} v\|_p^p \leq \sum_{T\in\mathcal{T}_h} |v_T|_2^p$ and using (21):

$$\|\Pi_{r_h} v\|_p^p \leq (C_p^*)^p \sum_{T\in\mathcal{T}_h} |v_{|T}|_1^p = (C_p^*)^p \sum_{T\in\mathcal{T}_h} \int_T \left| \sum_{i=1}^{d+1} v(b_{i_T}) B_i(x) \right|^p dx = (C_p^*)^p \|v\|_p^p.$$

This yields the conclusion.

□

Lemma 5.2 *Consider a function $\sigma_h \in V_h$. Then,*

$$\|\Pi_{r_h}\sigma_h - \sigma_h\|_p \leq h\,\|\nabla\sigma_h\|_p \qquad \forall\, p \in [1,+\infty], \qquad \forall\, r_h \in V_h.$$

Proof:

By the definition of Π_{r_h} and Hypothesis 1,

$$\|\Pi_{r_h}\sigma_h - \sigma_h\|_p^p = \sum_{T\in\mathcal{T}_h} \int_T |(\Pi_{r_h}\sigma_h)(x) - \sigma_h(x)|^p dx =$$

$$= \sum_{T\in\mathcal{T}_h} \int_T \left| \sum_{i=1}^{d+1} \sigma_h(b_{i_T})\lambda_{i_T}^T - \sum_{i=1}^{d+1} \sigma_h(x)\lambda_{i_T}^T \right|^p dx.$$

Using the convexity of function $g(x) = x^p$ for $p \geq 1$,

$$\|\Pi_{r_h}\sigma_h - \sigma_h\|_p^p \leq \sum_{T\in\mathcal{T}_h} \int_T \left[\sum_{i=1}^{d+1} \lambda_{i_T}^T |\sigma_h(b_{i_T}) - \sigma_h(x)| \right]^p dx \leq$$

$$\leq \sum_{T\in\mathcal{T}_h} \int_T \sum_{i=1}^{d+1} \lambda_{i_T}^T |\sigma_h(b_{i_T}) - \sigma_h(x)|^p dx.$$

As σ_h is linear, $\sigma_h(b_{i_T}) - \sigma_h(x) = \nabla\sigma_{h_{|T}} \cdot (b_{i_T} - x)$. So,

$$|\sigma_h(b_{i_T}) - \sigma_h(x)| \leq |\nabla\sigma_{h_{|T}}| h_T, \tag{22}$$

where h_T is the diameter of the element T. Thus, by Hypothesis 1,

$$\|\Pi_{r_h}\sigma_h - \sigma_h\|_p^p \leq \sum_{T\in\mathcal{T}_h} \int_T \sum_{i=1}^{d+1} \lambda_{i_T}^T\, h_T^p |\nabla\sigma_{h_{|T}}|^p dx \leq h^p \sum_{T\in\mathcal{T}_h} \int_T |\nabla\sigma_{h_{|T}}|^p dx \leq h^p \|\nabla\sigma_h\|_{L^p(\Omega)}^p.$$

If $p = +\infty$, the result also follows from (22).

□

Note that this result gives an error estimate for the distribution interpolation of piecewise affine finite elements. This estimate will be crucial to handle the transport and up-winded source terms in our error analysis.

We next give a technical lemma to handle the convergence of our discrete approximations (8) and (9).

We shall denote

$$\|v\|_V = \|\nabla v\|_2, \quad \forall v \in V.$$

This is a norm on V equivalent to the $H^1(\Omega)$ norm.

Lemma 5.3 *Assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is regular. Then the following holds*

a) *Given $v \in V$, for any sequence $\{r_h\}_{h>0} \subset V$ with $r_h \in V_h$, $\forall h > 0$, there exists a sequence $\{v_h\}_{h>0} \subset V$ such that $v_h \in V_h$, $\forall h > 0$, satisfying*

$$\lim_{h\to 0} \|\Pi_{r_h} v_h - v\|_r = 0, \quad \textit{for any } r \in [1, r_{max}), \tag{23}$$

where $r_{max} = +\infty$ if $d = 2$ and $r_{max} = 6$ if $d = 3$.

b) *Given $v \in V$ consider a sequence $\{v_h\}_{h>0} \subset V$ with $v_h \in V_h$, $\forall h > 0$, that verifies*

$$\lim_{h\to 0} \|v_h - v\|_r = 0, \quad \textit{for any } r \in [1, r_{max}).$$

Then,

$$\lim_{h\to 0} \|\Pi_{v_h} v_h - v\|_r = 0. \tag{24}$$

Proof:

a) As the family of triangulations is regular, there exits a sequence $\{v_h\}_{h>0} \subset V$ such that $v_h \in V_h$, $\forall h > 0$, strongly convergent to v in V (Cf [6]). As $H^1(\Omega)$ is compactly embedded in $L^r(\Omega)$ if $1 \le r < r_{max}$, we may assume that this sequence is strongly convergent to v in $L^r(\Omega)$. Using Lemma 5.2,

$$\|\Pi_{r_h} v_h - v\|_r \le \|\Pi_{r_h} v_h - v_h\|_r + \|v_h - v\|_r \le h\, \|\nabla v_h\|_r + \|v_h - v\|_r$$

For $1 \le r \le 2$,

$$\|\nabla v_h\|_r \le \|\nabla v_h\|_2.$$

For $r \in (2, +\infty)$, we use an inverse inequality for finite elements (See [7], for instance): Because of the regularity of the mesh, there exists a constant $C_1 > 0$ such that:

$$\|\nabla v_h\|_r \le C_1\, h^{d(\frac{1}{r}-\frac{1}{2})} \|\nabla v_h\|_2 \quad \forall\, r > 2, \quad \forall v_h \in V_h. \tag{25}$$

Then,

$$\|\Pi_{r_h} v_h - v\|_r \leq C_1\, h^\beta\, \|\nabla v_h\|_2 + \|v_h - v\|_r$$

with $\beta = 1 - d(1/2 - 1/r) > 0$ if $r \in (2, r_{max})$ and $\beta = 1$ if $r \in [1, 2]$. This proves a)

b) Consider $v \in V$. Again, there exists a sequence $\{z_h\}_{h>0} \subset V$ such that $z_h \in V_h$, $\forall h > 0$, strongly convergent to v in $V \cap L^r(\Omega)$. Using Lemmas 5.1 and 5.2,

$$\begin{aligned}
\|\Pi_{v_h} v_h - v\|_r &\leq \|\Pi_{v_h} v_h - \Pi_{v_h} z_h\|_r + \|\Pi_{v_h} z_h - z_h\|_r + \|z_h - v\|_r \leq \\
&\leq C_p\, \|v_h - z_h\|_r + h\, \|\nabla z_h\|_r + \|z_h - v\|_r \\
&\leq C_p\, \|v_h - v\|_r + (C_p + 1)\, \|z_h - v\|_r + h\, \|\nabla z_h\|_r.
\end{aligned}$$

Proceeding as in the proof of a),

$$h\, \|\nabla z_h\|_r \leq C_1\, h^\beta\, \|\nabla z_h\|_0.$$

with the same values for β. Then, b) follows.

□

Lemma 5.4 *Under* Hypothesis 1*, for any $r_h \in V_h$, the form $b(r_h) : V_h \times V_h \mapsto \mathbb{R}$ defined by*

$$b(r_h; v_h, w_h) = \int_\Omega (u \cdot \nabla v_h) \Pi_{r_h} w_h$$

is bilinear and bounded, satisfying

$$b(r_h; v_h, w_h) \leq C_q \|u\|_q \|v_h\|_V \|w_h\|_V,$$

where C_q is a positive constant.

Proof: This is a direct consequence of Lemma 5.1 and the Sobolev embedding of $H^1(\Omega)$ in $L^r(\Omega)$, if $1 \leq r \leq r_{max}$ (excluding the case $r = +\infty$).

□

Note that all the above properties of operator Π_{r_h} rely only on Hypothesis 1. We next state some properties of the discrete advection operator, that shall also rely on Hypothesis 2.

Lemma 5.5 *Under Hypotheses 2, for any $r_h \in V_h$, the form $b(r_h)$ is semi-positive, in the sense that it satisfies*

$$b(r_h; v_h, v_h) \geq 0 \quad \forall\, v_h \in V_h.$$

PROOF: Consider an element $z_h = \sum_{i=1}^{M} z_i \varphi_i \in V_h$. Then,

$$\begin{aligned} b(r_h; z_h, z_h) &= \int_\Omega (u \cdot \nabla z_h)\, \Pi_{r_h} z_h = \sum_{i,j=1}^{M} \int_\Omega (u \cdot \nabla \varphi_j)\, \Pi_{r_h} \varphi_i\, z_i z_j = \\ &= \sum_{i,j=1}^{M} z_i\, C_{ij}(r_h)\, z_j = Z^t C(r_h) Z, \end{aligned}$$

where $Z = (z_1 \dots z_M)^t$. Then,

$$\int_\Omega (u \cdot \nabla z_h)\, \Pi_{r_h} z_h + \nu \int_\Omega \nabla z_h \cdot \nabla z_h = Z^t (C(r_h) + \nu\, L) Z = Z^t A_\nu(r_h) Z.$$

By Hypothesis 2 $A_\nu(r_h)$ is a M-matrix, then it is semi-positive defined, and

$$b(r_h; z_h, z_h) = \lim_{\nu \to 0^+} Z^t A_\nu(r_h) Z \geq 0.$$

□

Note that this result proves that Hypothesis 2 could alternativaly be formulated as

Hypothesis 2'

Matrix $C(s_h)$ is semi-positive defined for any $s_h \in V_h$, and matrix L is an M-matrix.

It is doubtful that matrix L is positive defined for PSI scheme (See Remark 4.1).

6 Existence and quasi-uniqueness

In this section we prove the existence of solution for the discrete problems (8) and (9). Also, we obtain an estimate for the difference between two solutions of the same problem. The proof of uniqueness is an open question due to the non-linear nature of the method we are considering.

To prove the existence of solution for the problem above, we shall use a particular form of Brouwer's Fixed Point Theorem that we state in the Lemma 6.1 (cf. [17, 31]). The semi-positiveness of the discrete convective operator (Lemma 5.5) plays a crucial role in this proof.

Lemma 6.1 *Let X be a finite dimensional Hilbert space with scalar product $[\cdot,\cdot]$ and norm $[\cdot]$. Let P be a continuous mapping from X into itself such that:*

$$[P(\xi_0), \xi_0] > 0 \quad \textit{for } [\xi_0] = k > 0.$$

Then $\exists \xi \in X$, with $[\xi] \leq k$, such that

$$P(\xi) = 0.$$

The existence of solutions of Problem (9) is stated by

Theorem 6.1 *Assume that $u \in (L^q(\Omega))^d$ for some $q > d$, $f \in L^p(\Omega)$ for $p > 1$ if $d = 2$ and $p > 6/5$ if $d = 3$. Then problem (9) admits at least one solution $\rho_h \in V_h$ that satisfies the estimate*

$$\|\rho_h\|_V \leq \frac{C}{\nu}\|f\|_p \tag{26}$$

where the constant $C > 0$ only depends on d, p and Ω.

Proof:

We define the mapping $P : V_h \to V_h$ as follows

$$[P(v_h), w_h] = b(v_h; v_h, w_h) + \nu(\nabla v_h, \nabla w_h) - (f, \Pi_{v_h} w_h), \quad \forall\ v_h, w_h \in V_h.$$

Due to Hypothesis 3, P is a continuous mapping on V_h. Also,

$$[P(v_h), v_h] = b(v_h; v_h, v_h) + \nu\|\nabla v_h\|_2^2 - (f, \Pi_{v_h} v_h).$$

Using Lemma 5.1 and the Sobolev embedding of $H^1(\Omega)$ in $L^r(\Omega)$, we obtain

$$|(f, \Pi_{v_h} v_h)| \leq C_{p'}\, \|f\|_p\, \|v_h\|_{p'} \leq C_{p'}\, C\, \|f\|_p\, \|\nabla v_h\|_2$$

As $b(v_h)$ is positive, we may write:

$$[P(v_h), v_h] \geq \nu\|\nabla v_h\|_2^2 - C\, \|f\|_p \|\nabla v_h\|_2 = \|\nabla v_h\|_2(\nu\|\nabla v_h\|_2 - C\, \|f\|_p).$$

Then, for $\|\nabla v_h\|_2 = k$, $[P(v_h), v_h] > 0$, for $k > \dfrac{C\, \|f\|_p}{\nu}$. Therefore, by Lemma 6.1, $\exists \rho_h \in V_h$ such that $P(\rho_h) = 0$. Then ρ_h is a solution of (9).

Furthermore,

$$0 = [P(\rho_h), \rho_h] \geq \|\nabla \rho_h\|_2(\nu\|\nabla \rho_h\|_2 - \|f\|_p).$$

So (26) holds.

□

Also, the existence of solution of Problem (8) is stated as

Theorem 6.2 *Assume that $u \in (L^q(\Omega))^d$ for some $q > d$, $f \in V'$. Then problem (8) admits at least one solution $\rho_h \in V_h$ that satisfies the estimate*

$$\|\rho_h\|_V \leq \frac{1}{\nu}\|f\|_{V'}. \tag{27}$$

Proof: Proceeding as in the proof of Theorem 6.1, we can write

$$[P(v_h), v_h] \geq \nu\|\nabla v_h\|_2^2 - \|f\|_{V'}\|\nabla v_h\|_2 = \|\nabla v_h\|_2(\nu\|\nabla v_h\|_2 - \|f\|_{V'}).$$

By Lemma 6.1, $\exists \rho_h \in V_h$ such that $P(\rho_h) = 0$. Then ρ_h is solution of (8) and so (27) holds.

□

The proximity between two possible solutions of problem (9) is stated in the following result.

Lemma 6.2 *Assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is regular. Assume that $f \in L^p(\Omega)$ for some $p > 1$. Let ρ_{1h} and ρ_{2h} be two solutions of problem (9). Then there exists a constant $C > 0$ such that*

$$\|\nabla(\rho_{1h} - \rho_{2h})\|_2 \leq \frac{C}{\nu^2}\|u\|_q \, \|f\|_p \, h^\alpha + \frac{C}{\nu}\|f\|_p h^\beta,$$

where $\alpha = 1 - \frac{d}{q}$ *and* $\beta = \begin{cases} 1 & if\ 2 \leq p < +\infty, \\ 1 + d\left(\frac{1}{2} - \frac{1}{p}\right) & \begin{cases} if\ 1 < p < 2\ \ for\ \ d = 2, \\ if\ 6/5 < p < 2\ \ for\ \ d = 3. \end{cases} \end{cases}$

Proof:

Using the previous notation, we can write

$$\begin{aligned} 0 &= [P(\rho_{1h}) - P(\rho_{2h}), w_h] = [b(\rho_{1h}; \rho_{1h}, w_h) - (u \cdot \nabla\rho_{1h}, w_h)] + \\ &+ [(u \cdot \nabla\rho_{1h}, w_h) - (u \cdot \nabla\rho_{2h}, w_h)] + [(u \cdot \nabla\rho_{2h}, w_h) - b(\rho_{2h}; \rho_{2h}, w_h)] + \\ &+ \nu(\nabla(\rho_{1h} - \rho_{2h}), \nabla w_h) + (f, \Pi_{\rho_{2h}} w_h - \Pi_{\rho_{1h}} w_h). \end{aligned} \tag{28}$$

By definition of b,

$$\begin{aligned} |b(\rho_{ih}; \rho_{ih}, w_h) - (u \cdot \nabla\rho_{ih}, w_h)| &= |(u \cdot \nabla\rho_{ih}, \Pi_{\rho_{ih}} w_h - w_h)| \leq \\ &\leq \|u\|_q \|\nabla\rho_{ih}\|_2 \|\Pi_{\rho_{ih}} w_h - w_h\|_{\hat{q}} \leq h \, \|u\|_q \|\nabla\rho_{ih}\|_2 \, \|\nabla w_h\|_{\hat{q}}, \end{aligned}$$

with $\dfrac{1}{q} + \dfrac{1}{\hat{q}} = \dfrac{1}{2}$, where we have used Lemma 5.2 in the last estimate. We next use the inverse inequality (25) that yields

$$\|\nabla w_h\|_{\hat{q}} \leq C_1 \ h^{d(\frac{1}{\hat{q}} - \frac{1}{2})} \|\nabla w_h\|_2.$$

So we have

$$|b(\rho_{ih}; \rho_{ih}, w_h) - (u \cdot \nabla\rho_{ih}, w_h)| \leq \frac{C_2}{\nu} \, h^{1-\frac{d}{q}} \, \|u\|_q \|f\|_p \, \|\nabla w_h\|_2,$$

where we have used estimate (26).

By another hand, if $w_h = \rho_{1h} - \rho_{2h}$, the second summand in the r.h.s. of (28) is

$$(u \cdot \nabla(\rho_{1h} - \rho_{2h}), \rho_{1h} - \rho_{2h}) = \frac{1}{2} \int_{\Gamma^+} u \cdot n \, (\rho_{1h} - \rho_{2h})^2 d\sigma \geq 0.$$

We now estimate the last summand of (28)

$$|(f, \Pi_{\rho_{2h}} w_h - \Pi_{\rho_{2h}} w_h)| \leq \|f\|_p \|\Pi_{\rho_{2h}} w_h - w_h\|_{p'} + \|f\|_p \|\Pi_{\rho_{1h}} w_h - w_h\|_{p'} \leq 2C_3 \, h \|f\|_p \|\nabla w_h\|_{p'}$$

where we use Lemma 5.2.
If $p \geq 2$ then $\|\nabla w_h\|_{p'} \leq \|\nabla w_h\|_2$ and if $p < 2$, we can used the inverse estimate (25). So,

$$|(f, \Pi_{\rho_{2h}} w_h - \Pi_{\rho_{2h}} w_h)| \leq 2C_3\, h^{\beta} \|f\|_p \|\nabla w_h\|_2 \tag{29}$$

with $\beta = \begin{cases} 1 & if\ 2 \leq p < +\infty, \\ 1 + d\left(\frac{1}{2} - \frac{1}{p}\right) & if\ p < 2. \end{cases}$

Thus, setting $w_h = \rho_{1h} - \rho_{2h}$, we have from (28),

$$\nu \|\nabla(\rho_{1h} - \rho_{2h})\|_2^2 \leq 2\frac{C_2}{\nu}\, h^{1-\frac{d}{q}}\, \|u\|_q \|f\|_p\, \|\nabla(\rho_{1h} - \rho_{2h})\|_2 + 2C\, h^{\beta} \|f\|_p \|\nabla(\rho_{1h} - \rho_{2h})\|_2,$$

that yields the result.

□

In the same way, we can prove the following result for problem (8)

Lemma 6.3 *Assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$is regular. Let ρ_{1h} and ρ_{2h} be two solutions of problem (8). Then there exists a constant $C > 0$ such that*

$$\|\nabla(\rho_{1h} - \rho_{2h})\|_2 \leq \frac{C}{\nu^2} \|u\|_q\, \|f\|_{V'}\, h^{1-\frac{d}{q}}.$$

7 Convergence analysis and error estimates

We next prove the strong convergence in H^1-norm of a solution of the discretizations (8) and (9) to the solution of the continuous problem (1) and error estimates in H^1-norm. We prove the result for discretization (9), as this is the more involved technically, and just state it for discretization (8).

Theorem 7.1 *Let Ω be a bounded polygonal domain of $\mathbb{R}^d$. Assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is regular. Assume that data of the advection-diffusion problem (1) verify $u \in (L^q(\Omega))^d$ for some $q > d$, $f \in L^p(\Omega)$ for $p > 1$ if $d = 2$ and $p > 6/5$ if $d = 3$.*

Then the sequence of solutions $\{\rho_h\}_{h>0}$ of the discretization (9) is strongly convergent in V to the solution ρ of problem (1) as h goes to zero.

Proof: **a)** By (26), the sequence $\{\rho_h\}_{h>0}$ is bounded in V. This is a closed sub-space of $H^1(\Omega)$. Then there exists a subsequence $\{\rho_{h'}\}_{h'>0}$ weakly convergent in V to some ρ^*.

Consider $v \in V$, by Lemma 5.3there exists a sequence $\{v_h\}_{h>0} \subset V$ such that $v_h \in V_h,\ \forall h > 0$, satisfying

$$\lim_{h \to 0} \|\Pi_{\rho_h} v_h - v\|_r = 0, \quad r \in [1, r_{max})$$

If $u \in (L^q(\Omega))^d$ with $q > d$, $\hat{q} \in [1, r_{max})$ where $\hat{q}$ is such that $\frac{1}{q} + \frac{1}{\hat{q}} = \frac{1}{2}$. Then,

$$\lim_{h' \to 0} \int_\Omega (u \cdot \nabla \rho_{h'})\, \Pi_{\rho'_h} v_h = \int_\Omega (u \cdot \nabla \rho^*).\, v;$$

Also, as $f \in L^p(\Omega)$ with $p > 1$ if $d = 2$ and $p > 6/5$ if $d = 3$, then $p' \in [1, r_{max})$ and so,

$$\lim_{h' \to 0} \int_\Omega f\, \Pi_{\rho'_h} v_h = \int_\Omega f\, v.$$

Consequently, ρ^* satisfies

$$\int_\Omega (u \cdot \nabla \rho^*)\, v + \nu \int_\Omega \nabla \rho^* \cdot \nabla v = \int_\Omega f\, v, \quad \forall v \in V.$$

We deduce that ρ^* is the weak solution of problem (1), that we have denoted ρ. As this solution is unique, the whole sequence $\{\rho_h\}_{h>0}$ converges weakly to ρ in V.

b) We prove the convergence of the $H^1 0(\Omega)$ semi-norm of ρ_h to the $H^1(\Omega)$ semi-norm of ρ. This proves the strong convergence as this is a norm on V equivalent to the norm of $H^1(\Omega)$. We follow the standard procedure. Respectively take $v = \rho$ and $v_h = \rho_h$ as test functions in the continuous and discrete problems (2) and (9). Then we have

$$\int_\Omega (u \cdot \nabla \rho)\, \rho + \nu \int_\Omega |\nabla \rho|^2 = \int_\Omega f\, \rho\,; \tag{30}$$

and

$$\int_\Omega (u \cdot \nabla \rho_h)\, \Pi_{\rho_h} \rho_h + \nu \int_\Omega |\nabla \rho_h|^2 = \int_\Omega f\, \Pi_{\rho_h} \rho_h\,; \tag{31}$$

By Lemma 5.3 b), $\Pi_{\rho_h} \rho_h$ strongly converges to ρ in $L^{\hat{q}}(\Omega)$. Then,

$$\lim_{h \to 0} \int_\Omega (u \cdot \nabla \rho_h)\, \Pi_{\rho_h} \rho_h = \int_\Omega (u \cdot \nabla \rho)\, \rho.$$

Also, as f belongs to $L^p(\Omega)$,

$$\lim_{h \to 0} \int_\Omega f\, \Pi_{\rho_h} \rho_h = \int_\Omega f\, \rho.$$

So we may pass to the limit in (31) to deduce using (30) that

$$\lim_{h \to 0} \int_\Omega |\nabla \rho_h|^2 = \int_\Omega |\nabla \rho|^2.$$

This concludes the proof.

□

The convergence of approximation (8) requires less regularity for the source term. It is stated as follows

Theorem 7.2 *Let Ω be a bounded polygonal domain of $\mathbb{R}^d$. Assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is regular. Assume that the data of the advection-diffusion problem (1) verify $u \in (L^q(\Omega))^d$ for some $q > d$, $f \in V'$.*

Then the sequence of solutions $\{\rho_h\}_{h>0}$ of the discretization (8) is strongly convergent in V to the solution ρ of problem (1) as h goes to zero.

We next state our main error estimates result.

Theorem 7.3 *Under the hypotheses of Theorem 7.1, the following error estimates for the solutions of the discrete problems (9) hold:*

$$\|\nabla(\rho - \rho_h)\|_2 \leq (2 + \frac{C}{\nu}\|u\|_q)\, d_1(\rho, V_h) + \frac{C}{\nu^2}\|u\|_q\|f\|_p\, h^\alpha + \frac{C}{\nu}\,\|f\|_p\, h^\beta, \tag{32}$$

where $d_1(\rho, V_h) = \inf_{v_h \in V_h} \|\nabla(\rho - v_h)\|_2$, *$C > 0$ is a constant independent of h and ν,*

$$\alpha = 1 - \frac{d}{q} > 0, \quad \beta = \begin{cases} 1 & if\ 2 \leq p < +\infty \\ 1 + d\left(\frac{1}{2} - \frac{1}{p}\right) & \begin{cases} if\ 1 < p < 2\ \ for\ \ d = 2 \\ if\ 6/5 < p < 2\ \ for\ \ d = 3. \end{cases} \end{cases}$$

Proof:

Denote by P_h the elliptic projection of ρ onto V_h and define the truncation error $\epsilon_h \in V'$ as:

$$< \epsilon_h, v >= (u \cdot \nabla P_h, v) + \nu(\nabla P_h, \nabla v) - (f, v) \quad \forall\ v \in V. \tag{33}$$

Take $v = v_h \in V_h$ in this expression and substract it to (9):

$$\begin{aligned} (u \cdot \nabla \rho_h, \Pi_{\rho_h} v_h) + \nu(\nabla \rho_h, \nabla v_h) - (u \cdot \nabla P_h, v_h) - \nu(\nabla P_h, \nabla v_h) & \\ -(f, \Pi_{\rho_h} v_h - v_h) \ &= \ - < \epsilon_h, v_h > . \end{aligned}$$

Interpolation error estimate.

Define the interpolation error as $e_h = \rho_h - P_h \in V_h$. Then:

$$(u \cdot \nabla \rho_h, \Pi_{\rho_h} v_h) - (u \cdot \nabla P_h, v_h) + \nu(\nabla e_h, \nabla v_h) - (f, \Pi_{\rho_h} v_h - v_h) = - < \epsilon_h, v_h > .$$

Summing and subtracting $(u \cdot \nabla P_h, \Pi_{\rho_h} v_h)$ in the left-hand side,

$$\begin{aligned} (u \cdot \nabla e_h, \Pi_{\rho_h} v_h) + (u \cdot \nabla P_h, \Pi_{\rho_h} v_h - v_h) + \nu(\nabla e_h, \nabla v_h) & \\ -(f, \Pi_{\rho_h} v_h - v_h) \ &= \ - < \epsilon_h, v_h > . \end{aligned}$$

Taking $v_h = e_h$,

$$\begin{aligned} (u \cdot \nabla e_h, \Pi_{\rho_h} e_h) + (u \cdot \nabla P_h, \Pi_{\rho_h} e_h - e_h) + \nu\|\nabla e_h\|_2^2 & \\ -(f, \Pi_{\rho_h} e_h - e_h) \ &= \ - < \epsilon_h, e_h > . \end{aligned}$$

Thus, thanks to the semi-positiveness of the scheme,

$$\begin{aligned}\nu\|\nabla e_h\|_2^2 &\leq -<\epsilon_h, e_h> -(u\cdot\nabla P_h, \Pi_{\rho_h}e_h - e_h) - (f, \Pi_{\rho_h}e_h - e_h).\\ &\leq -<\epsilon_h, e_h> +\|u\|_q\|\nabla P_h\|_2\|\Pi_{\rho_h}e_h - e_h\|_{\hat{q}}\\ &+ \|f\|_p\|\Pi_{\rho_h}e_h - e_h\|_{p'},\end{aligned} \tag{34}$$

where $\frac{1}{q}+\frac{1}{\hat{q}}=\frac{1}{2}$, $\frac{1}{p}+\frac{1}{p'}=1$.

As P_h is the orthogonal projection of ρ on V_h,

$$\|\nabla P_h\|_2 \leq \|\nabla\rho\|_2 \leq \frac{C}{\nu}\|f\|_p. \tag{35}$$

By another hand, proceeding as in the obtention of estimate (29) in Lemma 6.2, we have

$$\|\Pi_{\rho_h}e_h - e_h\|_{\hat{q}} \leq C\,h^{\alpha}\|\nabla e_h\|_2 \tag{36}$$

with $\alpha = 1-\frac{d}{q}$, and

$$\|\Pi_{\rho_h}e_h - e_h\|_{p'} \leq C\,h^{\beta}\|\nabla e_h\|_2 \tag{37}$$

where $\beta = \begin{cases} 1 & if\ 2\leq p<+\infty,\\ 1+d\left(\frac{1}{2}-\frac{1}{p}\right) & \begin{cases} if\ 1<p<2\ \ for\ \ d=2,\\ if\ 6/5<p<2\ \ for\ \ d=3.\end{cases}\end{cases}$

Inserting estimates (35), (36), and (37), into (34) we obtain

$$\nu\|\nabla e_h\|_2^2 \leq -<\epsilon_h, e_h> + \frac{C}{\nu}\,h^{\alpha}\,\|u\|_q\|f\|_p\|\nabla e_h\|_2 + C\,h^{\beta}\,\|f\|_p\,\|\nabla e_h\|_2.$$

Thus,

$$\nu\|\nabla e_h\|_2 \leq \|\epsilon_h\|_{V'} + \frac{C}{\nu}\,h^{\alpha}\,\|u\|_q\|f\|_p + C\,h^{\beta}\,\|f\|_p. \tag{38}$$

Truncation error estimate.

To obtain the estimate for the truncation error, we subtract (2) to (33):

$$(u\cdot\nabla\rho, v) + \nu(\nabla\rho,\nabla v) - (u\cdot\nabla P_h, v) - \nu(\nabla P_h,\nabla v) = -<\epsilon_h, v>.$$

Thus,

$$<\epsilon_h, v> = (u\cdot\nabla(P_h-\rho), v) + \nu(\nabla(P_h-\rho),\nabla v).$$

We bound this expression using Sobolev injections. We use that as $q>d$, then $\hat{q}<r_{max}$:

$$\begin{aligned}<\epsilon_h, v> &\leq (\|u\|_q\,\|v\|_{\hat{q}} + \nu\,\|\nabla v\|_2)\|\nabla(P_h-\rho)\|_2\\ &\leq (C\,\|u\|_q + \nu)\|\nabla(P_h-\rho)\|_2\|\nabla v\|_2;\end{aligned}$$

and

$$\|\epsilon_h\|_{V'} \leq (C\,\|u\|_q + \nu)\,d_1(\rho, V_h), \tag{39}$$

where we have used $d_1(\rho, V_h) = \|\nabla(P_h-\rho)\|_2$.

Conclusion.

Inserting estimate (39) into (38), we obtain estimate (32).

□

The obtention of error estimates for Problem (8) may be treated as a sub-case of the preceding analysis. In the case of Problem (8), the error term due to the upwinding of the source term does not appear. We may prove the following error estimate result:

Theorem 7.4 *Under the hypotheses of Theorem 7.2, the following error estimates for the solutions of the discrete problems (8) hold:*

$$\|\nabla(\rho-\rho_h)\|_2 \leq (2+\frac{C}{\nu}\|u\|_q)\, d_1(\rho, V_h) + \frac{C}{\nu^2}\|u\|_q\|f\|_{V'}\, h^{(1-\frac{d}{q})}, \tag{40}$$

Remark 7.1 *The error estimate (32) is of optimal order when $u \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$. Indeed, in this case the error term due to the distribution interpolate of the test function is of order $\mathcal{O}(h)$. This is the same order as the interpolation error on V_h, as $d_1(\rho, V_h)$ is $\mathcal{O}(h)$ if $\rho \in H^2(\Omega)$. In the case of problem (8), the error estimate (40) is of optimal order when $u \in L^\infty(\Omega)$.*

8 Maximum principle and L^r-estimates

In this section we prove that the maximum principle is satisfied for the discrete problems, and that we obtain L^r estimates for the discrete solutions for convex polygonal domains.

We only prove the L^r estimates for $d = 3$ since for $d = 2$ are immediate due to Sobolev embeddings.

Theorem 8.1 *Under the conditions of Theorem 6.1 or Theorem 6.2, respectively, assume $f \geq 0$. Then, any solution ρ_h of the discrete problem (9) or (8), respectively, is non-negative.*

Proof:

We have approximated the advection-diffusion problem by a positive method defined by a matrix $A_\nu(\rho_h) = C(\rho_h) + \nu\, L$.

Let us denote by $R = (\rho(b_1), \cdots, \rho(b_M))^t \in \mathbb{R}^M$ the vector of unknowns. In matrix form, R is solution of an algebraic system with the structure

$$\mathcal{A}(R)\, R = F, \text{ with } F = (f_1, \cdots, f_M)^t\,,$$

$$f_i = \int_\Omega f\, \Pi_{\rho_h}\varphi_i \geq 0,\ i = 1, \cdots, M, \text{ for problem (9).}$$

$$f_i = \int_\Omega f\, \varphi_i \geq 0,\ i = 1, \cdots, M, \text{ for problem (8).}$$

Once we know that this non-linear system has a solution, we deduce $R = \mathcal{A}^{-1}(R)\,F$, and we reproduce the classical argument: As $\mathcal{A}(R)$ is an M-matrix, then it is monotone so $\mathcal{A}^{-1}(R)$ has non-negative entries and consequently $R_i = \rho(b_i) \geq 0$, $i = 1, \cdots, M$. As ρ_h is piecewise linear, then $\rho_h \geq 0$ in $\overline{\Omega}$.

□

The obtention of L^r estimates, follow a duality argument developed in Brenner & Scott [7] that generalizes the classical one by Nitsche.

Theorem 8.2 *Assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is regular. Assume also that $f \in V'$ and $u \in (L^q(\Omega))^d$ with $q > 12$. Assume finally that the advection-diffusion problem (1) is regular in the sense that its solution satisfies*

$$\|\rho\|_{2,+\infty} \leq C\,\|f\|_{V'}. \tag{41}$$

Then, the sequence of solutions of problem (8) $\{\rho_h\}$ is bounded in $L^r(\Omega)$, for all $1 \leq r < +\infty$. More specifically, there exists a constant $C > 0$ such that

$$\|\rho_h\|_r \leq C\,\|f\|_{V'}.$$

Proof:

The proof for $1 \leq r < 6$ is a consequence of the theorem of Sobolev and the bound (26).

Consider the adjoint problem: For a given $F \in V'$,

$$\begin{cases} \text{Find } \sigma \in V \text{ such that} \\ a(v, \sigma) =< F, v > \quad \forall\, v \in V, \end{cases} \tag{42}$$

where $a(\cdot,\cdot)$ is the bilinear form associated to the continuous problem (1), defined by (3).

Consider also the approximate dual problem on V_h, by the standard Galerkin Finite Element method :

$$\begin{cases} \text{Find } \sigma_h \in V_h \text{ such that} \\ a(v_h, \sigma_h) =< F, v_h > \quad \forall\, v_h \in V_h. \end{cases} \tag{43}$$

Under hypothesis (41), the following estimates hold for problems (1) (with $g = 0$), (42) and (43) (Cf. [7]):

$$\|\sigma\|_{1,q} \leq C_q\,\|F\|_q \quad \text{with } 1 < q \leq \infty, \tag{44}$$

$$\|\sigma_h\|_{1,r} \leq C_r \|\sigma\|_{W^{1,r}(\Omega)} \quad \text{with } 1 < r \leq \infty, \tag{45}$$

for some positive constants C_q, C_r.

We consider problem (43) with $F = sign(\rho_h)|\rho_h|^{r-1}$, that belongs to $L^{r'}$ with $\dfrac{1}{r} + \dfrac{1}{r'} = 1$ and

$$\|F\|_{r'} = \|\rho_h\|_r^{r-1}. \tag{46}$$

We set $v_h = \rho_h$ in (43). Then we have

$$\|\rho_h\|_r^r = a(\rho_h, \sigma_h).$$

Also, as ρ is a solution of (1) and ρ_h is a solution of (8),

$$\begin{aligned} a(\rho - \rho_h, \sigma_h) &= a(\rho, \sigma_h) - a_h(\rho_h, \sigma_h) + a_h(\rho_h, \sigma_h) - a(\rho_h, \sigma_h) = \\ &= \big[(u \cdot \nabla \rho_h, \Pi_{\rho_h} \sigma_h) + \nu(\nabla \rho_h, \nabla \sigma_h)\big] \\ &- \big[(u \cdot \nabla \rho_h, \sigma_h) + \nu(\nabla \rho_h, \nabla \sigma_h)\big] = \delta_h, \end{aligned}$$

where $a_h(\cdot, \cdot)$ is the bilinear form defined by (7) and

$$\delta_h = (u \cdot \nabla \rho_h, \Pi_{\rho_h} \sigma_h - \sigma_h) = (u \cdot \nabla(\rho_h - \rho), \Pi_{\rho_h} \sigma_h - \sigma_h) + (u \cdot \nabla \rho, \Pi_{\rho_h} \sigma_h - \sigma_h) = I + II \quad (47)$$

Estimate for I

$$|I| \leq \|u\|_q \|\nabla(\rho_h - \rho)\|_2 \|\Pi_{\rho_h} \sigma_h - \sigma_h\|_{\hat{q}} \leq \|u\|_q \|\nabla(\rho_h - \rho)\|_2 h \|\nabla \sigma_h\|_{\hat{q}}, \quad (48)$$

where $\dfrac{1}{q} + \dfrac{1}{\hat{q}} = 1$, and we have used Lemma 5.2.
We can write the estimate (40) as

$$\|\nabla(\rho_h - \rho)\|_2 \leq C_1 \|f\|_{V'} h^{1 - \frac{3}{q}}$$

As $r \in [6, +\infty)$ and $q > 12$, we have $r' < \hat{q}$ and then an inverse inequality similar to (25) yields

$$\|\nabla \sigma_h\|_{\hat{q}} \leq C_2 \, h^{-3\left(\frac{1}{r'} - \frac{1}{\hat{q}}\right)} \|\nabla \sigma_h\|_{r'}$$

Inserting this last estimate into (48) we obtain

$$|I| \leq C_3 \|u\|_q \|f\|_{V'} h^{\gamma} \|\nabla \sigma_h\|_{r'}$$

with $\gamma = 2 - \dfrac{3}{q} - 3\left(\dfrac{1}{r'} - \dfrac{1}{\hat{q}}\right) = \dfrac{7}{2} - \dfrac{6}{q} - \dfrac{3}{r'} > \dfrac{1}{2} - \dfrac{6}{q} > 0$, as $q > 12$ and $r' > 1$.

Estimate for II

Thanks to the regularity of the continuos solution, we have

$$|II| \leq \|u\|_q \, \|\nabla \rho\|_m \, \|\Pi_{\rho_h} \sigma_h - \sigma_h\|_{r'}$$

with $\dfrac{1}{m} = 1 - \dfrac{1}{q} - \dfrac{1}{r'}$.
Using (41) and Lemma 5.1, we deduce

$$|II| \leq C_4 \, \|u\|_q \, \|f\|_{V'} \, \|\sigma_h\|_{r'}. \quad (49)$$

Thus,

$$|\delta_h| \leq C_5 \, \|u\|_q \|f\|_{V'} (1 + h^\gamma) \|\nabla \sigma_h\|_{r'} \leq C_6 \, \|u\|_q \|f\|_{V'} \|F\|_{r'}, \tag{50}$$

where we have used (44) and (45) in last estimate.
Going back to (47), applying (45) and (44) we obtain

$$\begin{aligned} \|\rho_h\|_r^r = a(\rho_h, \sigma_h) &= a(\rho, \sigma_h) - \delta_h \leq C_7 \, \|\rho\|_{1,r} \|\sigma_h\|_{1,r'} - \delta_h \\ &\leq C_8 \|\rho\|_{1,r} \|F\|_{r'} + \delta_h. \end{aligned} \tag{51}$$

Then, using (46) and (50) we have

$$\|\rho_h\|_r^r \leq C_9 \, \|f\|_{V'} \|F\|_{r'} \leq C_{10} \, \|f\|_{V'} \|\rho_h\|_r^{r-1}$$

So,

$$\|\rho_h\|_r \leq C \|f\|_{V'}.$$

□

Theorem 8.3 *Assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is regular. Assume also that $f \in L^p(\Omega)$ for some $p \geq 1$, and $u \in (L^q(\Omega))^d$ with $q > 12$. Assume finally that the advection-diffusion problem (1) is regular in the sense that its solution satisfies*

$$\|\rho\|_{2,+\infty} \leq \, C \, \|f\|_p. \tag{52}$$

Then, the sequence of solutions of problem (9) $\{\rho_h\}_{h>0}$ is bounded in $L^r(\Omega)$, for $r \in [1, p]$. More specifically, there exists a constant $C > 0$ such that

$$\|\rho_h\|_r \leq \, C \, \|f\|_p.$$

Proof: Proceeding as in Theorem 8.2 we obtain for ρ solution of (1) and ρ_h a solution of (9),

$$\begin{aligned} a(\rho - \rho_h, \sigma_h) &= a(\rho, \sigma_h) - a_h(\rho_h, \sigma_h) + a_h(\rho_h, \sigma_h) - a(\rho_h, \sigma_h) = \\ &= (f, \sigma_h) - (f, \Pi_{\rho_h} \sigma_h) + \big[(u \cdot \nabla \rho_h, \Pi_{\rho_h} \sigma_h) + \nu(\nabla \rho_h, \nabla \sigma_h)\big] \\ &- \big[(u \cdot \nabla \rho_h, \sigma_h) + \nu(\nabla \rho_h, \nabla \sigma_h)\big] = \delta_h, \end{aligned}$$

with

$$\begin{aligned} \delta_h = (u \cdot \nabla \rho_h, \Pi_{\rho_h} \sigma_h - \sigma_h) + (f, \sigma_h - \Pi_{\rho_h} \sigma_h) = \\ (u \cdot \nabla(\rho_h - \rho), \Pi_{\rho_h} \sigma_h - \sigma_h) + (u \cdot \nabla \rho, \Pi_{\rho_h} \sigma_h - \sigma_h) + (f, \sigma_h - \Pi_{\rho_h} \sigma_h) = \\ = I + II + III \end{aligned} \tag{53}$$

Estimate of I

This estimate is the similar as that of I in Theorem 8.2 but we now use the error estimate (32) and then we have

$$|I| \leq C_1 \|u\|_q \|f\|_p h^\gamma \|\nabla \sigma_h\|_{r'}, \tag{54}$$

with $\gamma = 2 - \dfrac{3}{q} - 3\left(\dfrac{1}{r'} - \dfrac{1}{\hat{q}}\right) > 0$.

Estimate of II

Using (52) and Lemma 5.1, we deduce

$$|II| \leq C_2 \|u\|_q \|f\|_p \|\sigma_h\|_{r'}. \tag{55}$$

Estimate of III

$$|III| \leq \|f\|_p \|\Pi_{\rho_h} \sigma_h - \sigma_h\|_{p'}$$

As $p \geq r$, then

$$\|\Pi_{\rho_h} \sigma_h - \sigma_h\|_{p'} \leq \|\Pi_{\rho_h} \sigma_h - \sigma_h\|_{r'}$$

and using Lemma 5.1, we deduce

$$|III| \leq C_3 \|f\|_p \|\sigma_h\|_{r'}. \tag{56}$$

Going back to (53) and using (54), (55) and (56) we obtain

$$|\delta_h| \leq C_1 \|u\|_q \|f\|_p h^\gamma \|\nabla \sigma_h\|_{r'} + C_2 \|u\|_q \|f\|_p \|\sigma_h\|_{r'} + C_3 \|f\|_p \|\sigma_h\|_{r'}$$

From this estimate for δ_h we proceed as in Theorem 8.2 and conclude the proof.

□

Remark 8.1 *The hypotheses (41) on the regularity of the advection-diffusion problem (1) is obtained for instance if $u \in L^\infty(\Omega)$ and Ω is convex. This hypothesis is needed to deduce estimate (44) (Cf. [4], [7]).*

9 Order of accuracy and well-balanced property

A question that naturally arises after the preceding analysis is to determine the order of accuracy of the PSI method in low-order norms, in particular in $L^2(\Omega)$ norm.

We may use the standard Nitsche duality argument to estimate this error, using the re-formulations (8) or (9) of the PSI method as a variational method, as we have done in

h	$\|\rho_h - \rho\|_0$	$p^h_{L^2}$	$\|\rho_h - \rho\|_\infty$	$p^h_{L^\infty}$	$\|\nabla(\rho_h - \rho)\|_0$	$p^h_{H^1}$
1/20	0.00653321	–	0.0161677	-	0.0422375	-
1/40	0.00332201	1.02445	0.00869829	0.894308	0.0218688	0.949653
1/80	0.0017075	0.982763	0.00439685	0.984262	0.0115252	0.924079
1/160	0.000840325	1.03452	0.00217046	1.01847	0.00569981	1.01581

Table 1: Convergence orders for PSI method, Test 1.

the proof of Theorem 8.2. However, this procedure only yields first-order accuracy, due to the first-order error stemming from the Distribution Interpolation of the test functions.

This does not necessarily indicates that the convergence order in $L^2(\Omega)$ norm can not be better than the first order obtained for the $H^1(\Omega)$ norm. To test this question we have approximated the advection-diffusion problem (1) in a simple but meaningful $\mathbb{R}^2$ test, when this has a $C^\infty(\Omega)$ solution, and Ω is the unit square. Specifically, we have considered:

Test 1

$$u = (1,0),\ f = y\, e^{x+y}\big[(y-1)(x^2+x-1) - 2x\,\nu(xy+x+y-3)\big],\ g = 0.$$

The exact solution is

$$\rho(x,y) = xy(x-1)(y-1)e^{x+y}.$$

We have solved this problem with the PSI method (9) on non-structured grids. This ensures the distribution of the flux and the genuine non-linear nature of the PSI method for this test. This would not had been the case if we had used structured grids for this particular velocity $u = (1,0)$.

We have estimated the order of convergence in $L^q(\Omega)$ norm by means of the numerical solutions ρ_h and ρ_{2h} computed on two meshes of sizes h and $2h$, respectively, as

$$p^h_{L^q} \simeq \log_2\left(\frac{\|\rho - \rho_{2h}\|_{L^2(\Omega)}}{\|\rho - \rho_h\|_{L^2(\Omega)}}\right).$$

Similarly, the convergence order in $H^1(\Omega)$ norm has been estimated by

$$p^h_{H^1} \simeq \log_2\left(\frac{\|\nabla(\rho - \rho_{2h})\|_{L^2(\Omega)}}{\|\nabla(\rho - \rho_h)\|_{L^2(\Omega)}}\right).$$

Table 1 shows the estimated convergence orders in $L^2(\Omega)$, $L^\infty(\Omega)$ and $H^1(\Omega)$ norms, for $h = 1/N$ with $N = 40$, 80 and 160, where N denotes the number of subdivisions on each side of Ω. We observe that all orders of convergence are close to 1. Although our grids are non-structured, and then we cannot expect a very smooth behaviour of the estimated convergence order, this value roughly approaches the value 1 as h decreases.

As the solution is very smooth and the domain is convex, then the standard Galerkin Finite Element approximation of problem (1) yields a second order accuracy in $L^2(\Omega)$ norm. However, this does not seem the case for the PSI method in view of our results.

Nevertheless, our Petrov-Galerkin formulation allows to prove another remarkable property of the upwinded formulation (9): Its well-balanced character for advection-dominated regimes. To give a rigourous definition of this property, let us consider a solution of problem (1), and define the consistency error $\epsilon(\rho) \in V_h'$

$$\langle \epsilon(\rho), v_h \rangle = \int_\Omega (u \cdot \nabla \rho)\, \Pi_{r_h} v_h + \nu \int_\Omega \nabla \rho \cdot \nabla v_h - \int_\Omega f\, \Pi_{r_h} v_h,$$

where $r_h \in V_h$ is some interpolate of ρ.

Definition 9.1 *(Well-balanced scheme)*
Consider a smooth solution ρ of the advection-diffusion problem (1).

- *We say that the numerical scheme (9) is well-balanced for the solution ρ if $\epsilon(\rho) = 0$.*
- *We say that the numerical scheme (9) is well-balanced for the solution ρ up to order $p > 0$ if $\|\epsilon(\rho)\|_{V_h'} = O(h^p)$.*

For $p = 2$, this property is the adaptation to our context of the "Second-order accuracy at steady state" property stated in [3]. This property is cited as a basic desgin principle for conservative schemes to solve hyperbolic systems of conservation laws.

In the context of numerical solution of Shallow Water equations, for instance, this property helps to design accurate schemes. It ensures, in particular, that water at rest is solved with high accuracy if $p \geq 2$ (See Bermúdez & Vázquez-Cendón [5], Chacón et al. [10]).

Let us remember the definition of the Péclet and grid Péclet numbers,

$$Pe = \frac{U\,L}{\nu}, \quad Pe_h = \frac{U\,h}{\nu},$$

where U and L respectively are a characteristic speed and lenght of the flow. The Péclet number measures the relative balance between the convective and the diffusive terms in equation (1). The advection-diffusion process is said to occur in *advection dominated* regime –with respect to the current grid– if $Pe_h \geq 1$.

Lemma 9.1 *The upwinded PSI scheme (9) is*

- *Exactly well-balanced for the advection equation*

$$\begin{cases} u \cdot \nabla \rho = f & \text{in } \Omega \\ \rho = g & \text{on } \Gamma^-, \end{cases} \tag{57}$$

  and

- *Well-balanced up to second order for the advection-diffusion equation (1) when the flow takes place in advection-dominated regime.*

Proof:

- **Transport equation.** This is a direct consequence of the upwinded structure of scheme (9). Indeed, the consistency error acts as

$$\langle \epsilon(\rho), v_h \rangle = \int_\Omega (u \cdot \nabla \rho - f)\, \Pi_{r_h} v_h = 0, \ \forall\, v_h \in V_h.$$

- **Advection-diffusion equation.** Let us assume for simplicity that equation (1) is written in adimensional quantities. In this case, $Pe = 1/\nu$ and $Pe_h = h/\nu$.

  Assume $\rho \in H^2(\Omega)$. Using integration by parts, $\forall\, v_h \in V_h$,

$$\langle \epsilon(\rho), v_h \rangle = \int_\Omega (u \cdot \nabla \rho - \nu\, \Delta\rho - f)\, \Pi_{r_h} v_h + \nu \int_\Omega \Delta\rho\, (\Pi_{r_h} v_h - v_h).$$

  Then, by Lemma 5.2

$$|\, \langle \epsilon(\rho), v_h \rangle \,| \leq \nu\, \|\Delta\rho\|_2 \|\Pi_{r_h} v_h - v_h\|_2 \leq \nu\, h\, \|\Delta\rho\|_2 \|\nabla v_h\|_2.$$

  So $\|\epsilon(\rho)\|_{V_h'} \leq \nu\, h\, \|\Delta\rho\|_2$. Using that $Pe_h \geq 1$,

$$\|\epsilon(\rho)\|_{V_h'} \leq \nu\, h\, Pe_h\, \|\Delta\rho\|_2 \leq \ \|\Delta\rho\|_2\, h^2.$$

□

Similarly, it may be proved that the scheme (8) is well-balanced up to second order for flows in diffusion-dominated regime, when $\nu \geq C\, U\, h^{-1}$.

Acknowledgements

The research of T. Chacón Rebollo and of M. Gómez Mármol to carry on this work was partially supported by Spanish Ministerio de Ciencia y Tecnología and European Community FEDER Grant BFM2003-07530-C02-01. The research of T. Chacón Rebollo was also partially supported by a Marie Curie Intra-European Fellowship within the 6th European Community Framework Programme.

References

[1] Abgrall, R., Mezine, M: *Construction of second-order accurate monotone and stable residual distribution schemes for unsteady flow problems*, Journal of Computational Physics 188, pp. 16-55 (2003).

[2] Abgrall, R., Mezine, M: *Construction of second-order accurate monotone and stable residual distribution schemes for steady problems*, Journal of Computational Physics 195, pp. 474-507 (2004).

[3] Abgrall, R.: *Toward the ultimate conservative scheme: Following the quest*, Journal of Computational Physics 167, pp. 277-315 (2001).

[4] Amrouche, C.; Girault, V.: *Propiétés fonctionnelles d'opérateurs. Application au problème de Stokes en dimensiones quelconque*, Publications du Laboratorie d'Analyse Numérique P. et M. Curie, R90025 (1990).

[5] A. Bermúdez, M. E. Vázquez Cendón, *Upwind Methods for Hyperbolic Conservation Laws with Source Terms.* Computers & Fluids 23(8), pp. 1049-1071 (1994).

[6] Bernardi, C., Maday, Y., Rapetti, F., *Discrétisations variationnelles de problèmes aux limites elliptiques.* Mathématiques & Applications, 45. Springer-Verlag, Berlin (2004).

[7] Brenner, S.; Ridgway Scott, L.: *The Mathematical Theory of Finite Element Methods*, Springer (2002).

[8] Brézis, H.: *Análisis funcional. Teoría y aplicaciones*, Alianza Editorial, Madrid (1984).

[9] Brezzi, F.; Rappaz, J.; Raviart, P.A.: *Finite dimensional approximation of nonlinear problems 1. Branches of singular solutions*, Numer. Math. 36, pp. 1-25 (1980).

[10] T. Chacón Rebollo, A. Domínguez Delgado, E. D. Fernández Nieto, *Asymptotically balanced schemes for non-homogeneous hyperbolic systems. Application to Shallow Water Equations.* C.R. Acad. Sci. Paris, Sr. I 338, pp. 85-90 (2004).

[11] Cockburn, B. : *An introduction to the discontinuous Galerkin method for advection-dominated problems.* Advanced numerical approximation of nonlinear hyperbolic equations (Cetraro, 1997), 151–268, Lecture Notes in Math., 1697, Springer, Berlin (1998).

[12] Cockburn, B; Shu, C.-W. : *Nonlinearly stable compact schemes for shock calculations.* SIAM J. Numer. Anal. 31, no. 3, 607–627 (1994).

[13] Deconinck, H. Struijs, R., Bourgeois, P, ; Roe, P.L.: *Compact advection schemes on unstructured meshes*, VKI Lecture Series 1993-04 Comput. Fluid Dynamics (1993).

[14] Deconinck, H.; Paillère, H.; Struijs, R.; Roe, P.L.: *Multidimensional upwind schemes based on flucuation-splitting for systems of conservation laws*, Computational Mechanics 11, pp. 323-340 (1993).

[15] Drăgănescu A., Dupont T. and Scott L Ridgway. *Failure of the discrete maximum principle for an elliptic Finite Element problem*, Math. Comp. **74**, No. 249, pp. 1-23 (2004).

[16] Ern, A.; Guermond, J.L.: *Eléments finis: théorie, applications, mise en oeuvre*, Springer, Paris (2001).

[17] Girault, V.; Raviart, P.A.: *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin (1986).

[18] Mizukami, A.; Hughes, T.R.J.: *A Petrov-Galerkin finite element method for advection-dominated flows: an accurate upwinding technique for satisfying the maximum principle*, Computer Methods in Appl. Mech. Engrg., 50, pp.181-193 (1985).

[19] Morton, K.W.: *Numerical Solution of advection-Diffusion Problems*, Chapman & Hall (1996).

[20] Narbona, G.: *Aproximación numérica de algunos flujos de interés en Arquitectura e Ingeniería mediante esquemas positivos en elementos finitos*, Tesis Doctoral de la Universidad de Sevilla (2004).

[21] Paillère, H.; Deconinck, H.; Struijs, R.; Roe, P.L.; Mesaros, L.M.; Müller, J.D.: *Computations of inviscid compressible flows using fluctuation-splitting on triangular meshes*, AIAA 93-3301, Julio 6-9, Orlando, Florida (1993).

[22] Perthame, B.; Qiu, Y.; Stoufflet, B. : *Sur la convergence des schmas "fluctuation-splitting" pour l'advection et leur utilisation en dynamique des gaz.* C. R. Acad. Sci. Paris Sr. I Math. 319, no. 3, 283–288 (1994).

[23] Perthame, B. *Convergence of N-schemes for linear advection equations.* Trends in applications of mathematics to mechanics (Lisbon, 1994), 323–333, Pitman Monogr. Surveys Pure Appl. Math., 77, Longman, Harlow (1995).

[24] Pironneau, O. : *On the transport-diffusion algorithm and its applications to the Navier-Stokes equations.* Numer. Math. 38, no. 3, 309–332 (1981/82).

[25] Roe, P.L.: *Linear advection schemes on triangular meshes*, Technical report, Cranfield Institute of Tecnology, CoA 8720 (1987).

[26] Roe, P.L.: *A Framework for Numerical Evolution Problems*, Numerical Methods for Fluid Dynamics, Academic Press (1982).

[27] Roe, P.L.; Sildikover, D.: *Optimum positive linear schemes for advection in two and three dimensions*, SIAM J. Numer. Anal. 29(6), pp. 1542-1568 (1992).

[28] Struijs, R.; Deconinck, H.; Roe, P.L.: *Fluctuation splitting schemes for the 2D Euler equations*, VKI Lecture Series 1991-01 Comput. Fluid Dynamics (1991).

[29] Struijs, R.; Deconinck, H.; Roe, P.L.; Do Palma, P.; Powell, A.G.: *Progress on multi-dimensional upwind euler solvers for unstructured grids*, AIAA 91-1550 (1991).

[30] Sli, Endre: *Convergence and nonlinear stability of the Lagrange-Galerkin method for the Navier-Stokes equations.* Numer. Math. 53, no. 4, 459–483 (1988).

[31] Temam, R.: *Navier-Stokes equations*, North-Holland Publishing Company, (1977).

[32] Toro, E. F. : *Riemann solvers and numerical methods for fluid dynamics*, Springer Verlag, Berlin (1977).

[33] Varga, R.,: *Matrix iterative analysis*, Prentice-Hall, New-Jersey (1977).