Exact pressure elimination for the Crouzeix-Raviart scheme applied to the Stokes and Navier-Stokes problems

Eric Chénier, Robert Eymard

Exact pressure elimination for the Crouzeix-Raviart scheme applied to the Stokes and Navier-Stokes problems

Eric Chénier[*] and Robert Eymard[†]

January 10, 2021

Abstract

We show that, using the Crouzeix-Raviart scheme, a cheap algebraic transformation, applied to the coupled velocity–pressure linear systems issued from the transient or steady Stokes or Navier-Stokes problems, leads to a linear system only involving as many auxiliary variables as the velocity components. This linear system, which is symmetric positive definite in the case of the transient Stokes problem and symmetric invertible in the case of the steady Stokes problem, with the same stencil as that of the velocity matrix, provides the exact solution of the initial coupled linear system. Numerical results show the increase of performance when applying direct or iterative solvers to the resolution of these linear systems.

Keywords: Navier-Stokes equations, Crouzeix-Raviart scheme, exact pressure elimination, hybridisation

1 Introduction

This paper is focused on the resolution of the coupled velocity-pressure linear systems issued from the discretisation by the Crouzeix-Raviart scheme [7] of the steady (or transient and semi-discretised in time) Stokes and Navier-Stokes problems, considering for the sake of simplicity homogeneous Dirichlet boundary conditions for the velocity. In the case of the Navier-Stokes problem, these linear systems are resulting from the Newton-Raphson method applied to iteratively solve the non-linear equations.

These linear systems are under the form

$$\begin{bmatrix} A & D^t \\ D & 0 \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{1}$$

where A is the rigidity matrix resulting from the use of the $\mathbb{P}^1$ non conforming finite element method for the velocities, completed by the mass matrix in case of the transient case, and by some derivatives issued from the convection term in case of the Navier-Stokes problem, D is the discrete divergence matrix written element by element, U is the vector of all velocity unknowns, P is the vector of all (but one) pressures unknowns and R is the right hand-side resulting from the momentum source terms.

In the case of the steady or transient Stokes problem, the matrix A is symmetric. This is no longer the case for the Navier-Stokes problem. But even in the case of the steady or transient Stokes problem, the matrix of the linear system (1) is not positive definite, due to the fact that there are negative eigenvalues

[*] MSME, Univ. Gustave Eiffel, Univ. Paris Est Créteil, CNRS, F-77454, Marne-la-Vallée, France. eric.chenier@univ-eiffel.fr

[†] LAMA, Univ. Gustave Eiffel, Univ. Paris Est Créteil, CNRS, F-77454, Marne-la-Vallée, France. Robert.Eymard@univ-eiffel.fr

since there are zeros on the main diagonal. This property makes much more complicate the use of iterative solvers based, for example, on conjugate gradient or GMRES [13] methods preconditioned with Incomplete Lower-Upper (ILU) factorization. Note that the implementation of the ILU preconditioners on parallel architectures [6] fails to provide the same preconditioning properties as ILU on only one processor, due to the loss of some sequential computations.

Then many authors are led, on small cases, to use direct solvers (recall the remarkable performances of the direct MUMPS solvers on parallel architectures [3, 2]). But on large matrices, such direct methods can no longer be reasonably applied, and there is a need to use all the same an iterative linear solver.

Another option consists in adding a small diagonal pressure-pressure connection, as performed by the augmented Lagrangian methods. But then, the iterative convergence properties of the solutions for such a modified system to that of the original one may become very slow.

Such difficulties for solving the linear systems issued from a mixed formulation are well-known when solving a simple Laplace problem. In this case, $H_{\rm div}$ conforming finite elements are used for the approximation of the gradient of the unknown (the Raviart-Thomas finite element is often used in the case of simplicial meshes), and piecewise constant elements are used for the unknown. A very clever method is then known for overcoming the difficulty of solving the linear systems issued from this problem: it is the famous hybridisation of the problem, leading to solve a symmetric positive definite linear system on the trace of the unknown on the faces of the mesh [4, 8, 14]. Note that a similar idea is used in [1] in the case of the Stokes problem, discretised by Hybrid High Order methods.

This paper is based on the extension of the same idea for applying an algebraic hybridisation to the case of the coupled linear systems (1). Let us emphasize that the solution of the linear system is not modified by the use of this hybridisation. In order the method to apply to the Navier-Stokes problem, we select an implementation of the non-linear convection term which does not increase the stencil of the Stokes problem [10].

In the transient Stokes problem, we get, after hybridisation, a symmetric positive definite linear system with as many unknowns as the velocities, and the same connection stencil (even in the case of the Stokes problem, the different space components of the auxiliary unknowns are connected, contrarily to the original velocity-velocity matrix).

In the steady case, we are led to introduce a modification in the diagonal blocks to have an invertible block diagonal matrix. Once again, the solution of the linear system is not altered by this modification. After hybridisation, we obtain in the case of the Stokes problem final symmetric linear system to be solved with as many unknowns as the velocities, but the matrix is no longer positive definite.

This paper is organised as follows. We first detail in Section 2 the construction of the scheme, with precising the treatment of the right-hand-side allowing exact numerical solutions in the case where it resumes to the gradient of a scalar field, and with a formulation of the convection term which does not enlarge the stencil. We then show in section 3 how the linear systems issued from this scheme can be algebraically handled for obtaining smaller linear systems with the same sparsity. We finally compare, in Section 4, the numerical efficiency of different linear solvers, applied to the initial coupled linear system and applied to their algebraic transformation.

2 The Crouzeix-Raviart scheme for $d = 2$ or $d = 3$

Let us first give the strong formulation of the Stokes and Navier-Stokes equations in their steady or semi-discrete transient versions:

$$\left\{\begin{aligned} \mu\overline{\boldsymbol{u}} - \nu\Delta\overline{\boldsymbol{u}} + \nabla\overline{p} + \boldsymbol{b}(\overline{\boldsymbol{u}}) &= \overline{\boldsymbol{f}} \quad \text{in } \Omega \\ \mathrm{div}\overline{\boldsymbol{u}} &= 0 \quad \text{in } \Omega \\ \overline{\boldsymbol{u}} &= 0 \quad \text{on } \partial\Omega \\ \textstyle\int_\Omega \overline{p}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0 & \end{aligned}\right. \tag{2}$$

where $\overline{\boldsymbol{u}} = (\overline{u}^{(i)})_{i=1,\ldots,d}$ with $d = 2$ or $d = 3$ represents the velocity field, $\Delta\overline{\boldsymbol{u}} = (\Delta\overline{u}^{(i)})_{i=1,\ldots,d}$, $\overline{p}$ is the pressure, the domain Ω with boundary $\partial\Omega$ is a bounded open set in $\mathbb{R}^d$, $\nu > 0$ is the invert of the Reynolds number, $\overline{\boldsymbol{f}} = (\overline{f}^{(i)})_{i=1,\ldots,d}$ is a given function defined on Ω, $\nabla\overline{p} = \big(\partial_i\overline{p}\big)_{i=1,\ldots,d}$, $\mathrm{div}\overline{\boldsymbol{u}} = \sum_{i=1}^d \partial_i\overline{u}^{(i)}$.

For the steady problem, $\mu = 0$ and in the case where the problem is transient, $\mu > 0$ is the invert of the time step: then $\overline{\boldsymbol{f}}$ includes a term issued from the velocity at the beginning of the time step (and the transient problem is semi-discretised in time).

For the transient or steady Stokes problems, we let

$$\boldsymbol{b}(\overline{\boldsymbol{u}}) = 0, \tag{3}$$

and for the Navier-Stokes problem, we define the non-linear convection term by

$$\boldsymbol{b}(\overline{\boldsymbol{u}}) = (\overline{\boldsymbol{u}} \cdot \nabla)\overline{\boldsymbol{u}} = \big(\sum_{j=1}^{d} \overline{u}^{(j)}\partial_j\overline{u}^{(i)}\big)_{i=1,\ldots,d} \,. \tag{4}$$

The standard weak formulation of Problem (2) is the following mixed one. Defining $L^2_0(\Omega)$ as the set of elements of $L^2(\Omega)$ with null mean value on Ω, this formulation is given by

$$\left\{\begin{array}{lll} \text{Find} & \overline{\boldsymbol{u}} \in H^1_0(\Omega)^d \text{ and } \overline{p} \in L^2_0(\Omega) \text{ such that} & \\ \forall\overline{\boldsymbol{v}} \in H^1_0(\Omega)^d, & \displaystyle\int_\Omega \Big(\mu\overline{\boldsymbol{u}} \cdot \overline{\boldsymbol{v}} + \nu\nabla\overline{\boldsymbol{u}} : \nabla\overline{\boldsymbol{v}} - \overline{p}\mathrm{div}\overline{\boldsymbol{v}} + \boldsymbol{b}(\overline{\boldsymbol{u}}) \cdot \overline{\boldsymbol{v}}\Big)\mathrm{d}\boldsymbol{x} & = \displaystyle\int_\Omega \overline{\boldsymbol{f}} \cdot \overline{\boldsymbol{v}}\mathrm{d}\boldsymbol{x} \\ \forall\overline{q} \in L^2_0(\Omega), & \displaystyle\int_\Omega \mathrm{div}\overline{\boldsymbol{u}}\ \overline{q}\mathrm{d}\boldsymbol{x} & = 0 \end{array}\right. \tag{5}$$

The Crouzeix-Raviart scheme [7] is the translation of the weak formulation (5) into discrete sets and operators applying on simplicial meshes (triangles in 2D, tetrahedra in 3D). It reads

$$\left\{\begin{array}{lll} \text{Find} & \boldsymbol{u} \in (V_h)^d \text{ and } p \in Q_{h,0} \text{ such that} & \\ \forall\boldsymbol{v} \in (V_h)^d, & \displaystyle\int_\Omega \Big(\mu\Pi_h\boldsymbol{u} \cdot \Pi_h\boldsymbol{v} + \nu\nabla_h\boldsymbol{u} : \nabla_h\boldsymbol{v} - p\mathrm{div}_h\boldsymbol{v}\Big)\mathrm{d}\boldsymbol{x} + b_h(\boldsymbol{u}, \boldsymbol{v}) & = \displaystyle\int_\Omega \overline{\boldsymbol{f}} \cdot \widehat{\Pi}_h\boldsymbol{v}\mathrm{d}\boldsymbol{x} \\ \forall q \in Q_{h,0}, & \displaystyle\int_\Omega \mathrm{div}_h\boldsymbol{u}\ q\mathrm{d}\boldsymbol{x} & = 0. \end{array}\right. \tag{6}$$

Let us define each of the discrete objects involved in (6).

1. **The finite dimensional space V_h.**
   Let $\mathcal{M}$ be a simplicial mesh, that is a finite set of disjoint open simplices whose closure recovers Ω. For $K \in \mathcal{M}$, we denote by $\boldsymbol{x}_K$ the centre of gravity of K. Denote by $\mathcal{F}$ the set of all faces (edges in 2D) of the mesh, that is partitioned into $\mathcal{F}_{\text{int}} \cup \mathcal{F}_{\text{ext}}$ (the set of interior and exterior faces), and denote for any $K \in \mathcal{M}$ by $\mathcal{F}_K$ the set of the faces of K. We denote by $\mathcal{F}_{K,\text{int}} = \mathcal{F}_K \cap \mathcal{F}_{\text{int}}$.

   For any $\sigma \in \mathcal{F}_K$, we denote by $\boldsymbol{n}_{K,\sigma}$ the unit vector, normal to σ and outward to K, and we let

   $$\boldsymbol{a}_{K,\sigma} = |\sigma|\boldsymbol{n}_{K,\sigma}.$$

   We assume that there are no hanging nodes, which implies that the cardinal of any $\mathcal{F}_K$ is equal to $d + 1$ (3 in 2D, 4 in 3D). For any face $\sigma \in \mathcal{F}$, we denote by $\mathcal{M}_\sigma$ the set of the simplices $K \in \mathcal{M}$

such that $\sigma \in \mathcal{F}_K$. Then the cardinal of $\mathcal{M}_\sigma$ is 2 for an interior face, 1 for an exterior face. For any $\sigma \in \mathcal{F}$, we denote by $\boldsymbol{x}_\sigma$ the centre of gravity of σ.

We then define, for any $\sigma \in \mathcal{F}_{\rm int}$ with $\mathcal{M}_\sigma = \{K, L\}$, the function $\varphi_\sigma : \Omega \to \mathbb{R}$ whose the restriction $\varphi_{\sigma,K}$ on K (respectively $\varphi_{\sigma,L}$ on L) is an affine function on K (respectively L) and which is null on any other element of the mesh. Moreover, one requests that the mean values of both $\varphi_{\sigma,K}$ and $\varphi_{\sigma,L}$ are equal to 1 on σ and equal to 0 on any $\sigma' \in \mathcal{F}_K \cup \mathcal{F}_L$ different from σ. These conditions are sufficient for defining in an unique way the affine functions $\varphi_{\sigma,K}$ and $\varphi_{\sigma,L}$, on each of which $d+1$ independent conditions have been specified. This definition ensures the continuity of the mean value of these functions on any face of the mesh, as well as the continuity of these functions at the centre of gravity of the faces of the mesh.

Then the space V_h is defined as the space spanned by the family $(\varphi_\sigma)_{\sigma \in \mathcal{F}_{\rm int}}$.

For any $v \in V_h$ and $K \in \mathcal{M}$, we denote by v_K the restriction of v to K (it is therefore an affine function).

For any $\boldsymbol{v} \in (V_h)^d$ and $\sigma \in \mathcal{F}$, we then denote by $\boldsymbol{V}_\sigma$ the vector $\boldsymbol{V}_\sigma = (V_\sigma^{(i)} := v_K^{(i)}(\boldsymbol{x}_\sigma))_{i=1,\ldots,d}$.

2. **The discrete operators ∇_h and div_h.**
The discrete operators ∇_h and div_h are defined as the "broken" ones, that means that there restriction to any element of the mesh are defined as the continuous ones:

$$\forall v \in V_h,\ \forall K \in \mathcal{M},\ (\nabla_h v)_{|K} = \nabla v_K \text{ and } \forall \boldsymbol{v} \in (V_h)^d,\ \forall K \in \mathcal{M},\ (\mathrm{div}_h \boldsymbol{v})_{|K} = \mathrm{div} v_K.$$

3. **The discrete reconstruction operator Π_h.**
The operator Π_h is introduced in order to obtain some mass lumping in the "mass matrix" term, that is in order to get a diagonal mass matrix. If $d = 2$, using the Crouzeix-Raviart basis functions, the matrix

$$M_{\sigma,\sigma'} = \int_K \varphi_\sigma(\boldsymbol{x})\varphi_{\sigma'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

is already diagonal, and then $\Pi_h = \mathrm{Id}$. But this fails if $d = 3$. We then denote by $\Pi_h \varphi_\sigma$ a piecewise constant function, equal to 1 in a domain surrounding σ and 0 elsewhere (this domain is defined as the union of the two triangles (2D) or tetrahedra (3D), the basis of which is σ, and the vertex of which is the centre of gravity of the neighbouring simplices).

4. **The discrete reconstruction operator $\widehat{\Pi}_h$.**
Following [12], the operator $\widehat{\Pi}_h \boldsymbol{v}$ is designed to ensure the following properties: $\widehat{\Pi}_h \boldsymbol{v} \in H_{\rm div}(\Omega)$ (which means a kind of continuity of the normal trace on any internal boundary), $\widehat{\Pi}_h \boldsymbol{v} - \boldsymbol{v}$ tends to 0 as h tends to 0 if $\boldsymbol{v}$ is the interpolation of any regular function, and finally there holds,

$$\forall \boldsymbol{v} \in V_h^d,\ \left(\forall q \in Q_{h,0},\ \int_\Omega \mathrm{div}_h \boldsymbol{v}\ q \mathrm{d}\boldsymbol{x} = 0\right) \Rightarrow \mathrm{div}\widehat{\Pi}_h \boldsymbol{v} = 0 \text{ a.e. in } \Omega. \tag{7}$$

Indeed, if we change $\overline{\boldsymbol{f}}$ into $\overline{\boldsymbol{f}} + \nabla\varphi$, Property (7) implies that the discrete velocity is not modified, only the pressure field is changed by the addition of an interpolation of φ. This property leads to a substantial decrease of the numerical error, in particular in the case where the major part of $\overline{\boldsymbol{f}}$ is constituted by the gradient of a scalar field. To this purpose, we use the Raviart-Thomas basis, which is conforming in $H_{\rm div}(\Omega)$ and defined, for all $K \in \mathcal{M}$, $\sigma \in \mathcal{F}_K$ and $\boldsymbol{x} \in K$, by

$$\boldsymbol{\psi}_{K,\sigma}(\boldsymbol{x}) = \frac{|\sigma|}{d\,|K|}(\boldsymbol{x} - \boldsymbol{s}_\sigma),$$

where $\boldsymbol{s}_\sigma$ is the vertex of K which is not a vertex of σ. Then we define, for any $\boldsymbol{x} \in K$,

$$\widehat{\Pi}_h \boldsymbol{v}(\boldsymbol{x}) = \sum_{\sigma \in \mathcal{F}_K} \boldsymbol{V}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \boldsymbol{\psi}_{K,\sigma}(\boldsymbol{x}) = \sum_{\sigma \in \mathcal{F}_K} \frac{\boldsymbol{V}_\sigma \cdot \boldsymbol{a}_{K,\sigma}}{d\,|K|}(\boldsymbol{x} - \boldsymbol{s}_\sigma).$$

We then approximate $\int_K \overline{\boldsymbol{f}} \cdot \widehat{\Pi}_h \boldsymbol{v} \mathrm{d}\boldsymbol{x}$ by

$$\int_\Omega \overline{\boldsymbol{f}} \cdot \widehat{\Pi}_h \boldsymbol{v} \mathrm{d}\boldsymbol{x} = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \boldsymbol{V}_\sigma \cdot \boldsymbol{a}_{K,\sigma} \Big(\frac{1}{|K|} \int_K \overline{\boldsymbol{f}} \mathrm{d}\boldsymbol{x} \Big) \cdot (\boldsymbol{x}_\sigma - \boldsymbol{x}_K). \tag{8}$$

5. **The finite dimensional space $Q_{h,0}$.**
We define Q_h as the finite dimensional subset of $L^2(\Omega)$ spanned by the characteristic functions ψ_K of all the simplices $K \in \mathcal{M}$ (ψ_K is the piecewise constant function defined on Ω which is equal to one inside K and 0 elsewhere). Since the pressures can be defined up to a constant value, instead of defining a space of functions with null average (which would connect all components of the function together), we select a given element of the mesh $\mathcal{M}$, denoted K_0, and we define the set $Q_{h,0}$ as the set of all elements $p \in Q_h$ vanishing on K_0. Note that, for any $p \in Q_{h,0}$, we retrieve an element of $L^2_0(\Omega)$, considering $p - \frac{1}{|\Omega|}\int_\Omega p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

6. **The non-linear form $b_h(\boldsymbol{u}, \boldsymbol{v})$.**
This non-linear form vanishes for the transient or steady Stokes problems. For the Navier-Stokes problem, the following discretisation for $b_h(\boldsymbol{u}, \boldsymbol{v})$ has been proposed by [10] and is compared to other choices in [9]. Its main advantage is to keep a reduced stencil in the linear systems. All the simplices K are split into co-volumes linked to the faces, as shown by Figure 1. The co-volume associated with a face σ in a simplex K, is defined as the cone $D_{K,\sigma}$ based on σ, whose vertex is the centre of gravity of K (it is then a simplex as well). This sub-mesh leads to the definition of $d(d-1)$ internal faces, each of them being common to $D_{K,\sigma}$ and $D_{K,\sigma'}$, denoted $\tau_{\sigma,\sigma'}$, for any pair $\sigma, \sigma' \in \mathcal{F}_K$. Then the unit normal vector to the face $\tau_{\sigma,\sigma'}$, oriented from $D_{K,\sigma}$ to $D_{K,\sigma'}$, is denoted by $\boldsymbol{n}_{\sigma,\sigma'}$. We then define $b_h(\boldsymbol{u}, \boldsymbol{v})$ by the relation

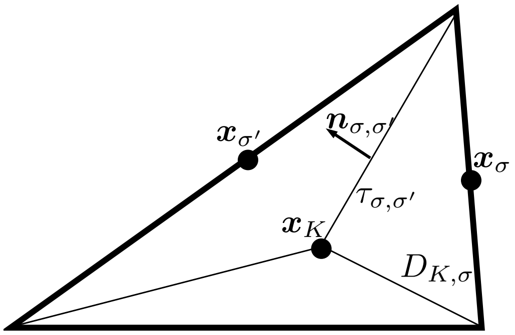

Figure 1: Co-volumes associated with faces

$$b_h(\boldsymbol{u}, \boldsymbol{v}) := \sum_{K \in \mathcal{M}} \sum_{\{\sigma,\sigma'\} \subset \mathcal{F}_K} F_{\sigma,\sigma'}(\boldsymbol{u})(\boldsymbol{U}_{\sigma'} - \boldsymbol{U}_\sigma) \cdot \frac{\boldsymbol{V}_\sigma + \boldsymbol{V}_{\sigma'}}{2}, \tag{9}$$

which also satisfies

$$b_h(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2} \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \boldsymbol{V}_\sigma \cdot \sum_{\sigma' \in \mathcal{F}_K \setminus \{\sigma\}} F_{\sigma,\sigma'}(\boldsymbol{u})(\boldsymbol{U}_{\sigma'} - \boldsymbol{U}_\sigma),$$

where $F_{\sigma,\sigma'}(\boldsymbol{u})$ is defined by

$$F_{\sigma,\sigma'}(\boldsymbol{u}) = \int_{\tau_{\sigma,\sigma'}} \boldsymbol{u}_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{\sigma,\sigma'} \mathrm{d}s(\boldsymbol{x}).$$

We remark that, for $\sigma, \sigma' \in \mathcal{F}_K$, the centre of gravity $\boldsymbol{x}_{\sigma,\sigma'}$ of $\tau_{\sigma,\sigma'}$ is given by

$$\boldsymbol{x}_{\sigma,\sigma'} = \boldsymbol{x}_\sigma + \boldsymbol{x}_{\sigma'} - \boldsymbol{x}_K,$$

and we observe that

$$|\tau_{\sigma,\sigma'}|\boldsymbol{n}_{\sigma,\sigma'} = \frac{1}{d+1}(\boldsymbol{a}_{K,\sigma'} - \boldsymbol{a}_{K,\sigma}).$$

This yields

$$F_{\sigma,\sigma'}(\boldsymbol{u}) = \Big(\boldsymbol{U}_\sigma + \boldsymbol{U}_{\sigma'} - \frac{1}{d+1} \sum_{\sigma'' \in \mathcal{F}_K} \boldsymbol{U}_{\sigma''}\Big) \cdot \frac{1}{d+1}(\boldsymbol{a}_{K,\sigma'} - \boldsymbol{a}_{K,\sigma}).$$

We then check that the relation $\sum_{\sigma' \in \mathcal{F}_K} \boldsymbol{U}_{\sigma'} \cdot \boldsymbol{a}_{K,\sigma'} = 0$ implies that

$$\sum_{\sigma' \in \mathcal{F}_K} F_{\sigma,\sigma'}(\boldsymbol{u}) = -\boldsymbol{U}_\sigma \cdot \boldsymbol{a}_{K,\sigma}.$$

Hence, the above definition is such that, if $\mathrm{div}_h \boldsymbol{u} = 0$, then there holds $b_h(\boldsymbol{u}, \boldsymbol{u}) = 0$ for all $\boldsymbol{u} \in V_h$. Indeed, there holds

$$0 = \int_{D_{K,\sigma}} \mathrm{div} \boldsymbol{u}_K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\sigma' \in \mathcal{F}_K \setminus \{\sigma\}} F_{\sigma,\sigma'}(\boldsymbol{u}) + \int_\sigma \boldsymbol{u}_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x}),$$

which implies that

$$\begin{aligned} b_h(\boldsymbol{u}, \boldsymbol{u}) &= \frac{1}{2} \sum_{K \in \mathcal{M}} \sum_{\{\sigma,\sigma'\} \subset \mathcal{F}_K} F_{\sigma,\sigma'}(\boldsymbol{u})(|\boldsymbol{U}_{\sigma'}|^2 - |\boldsymbol{U}_\sigma|^2) \\ &= -\frac{1}{2} \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\boldsymbol{U}_\sigma|^2 \sum_{\sigma' \in \mathcal{F}_K \setminus \{\sigma\}} F_{\sigma,\sigma'}(\boldsymbol{u}) = \frac{1}{2} \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\boldsymbol{U}_\sigma|^2 \int_\sigma \boldsymbol{u}_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x}), \end{aligned}$$

and this last term vanishes, since if $\sigma \in \mathcal{F}_{\mathrm{ext}}$, $\int_\sigma \boldsymbol{u}_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x}) = 0$, and if $\mathcal{M}_\sigma = \{K, L\}$, then, by definition of V_h from $\widehat{V}_h$, there holds

$$\int_\sigma \boldsymbol{u}_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x}) + \int_\sigma \boldsymbol{u}_L(\boldsymbol{x}) \cdot \boldsymbol{n}_{L,\sigma} \mathrm{d}s(\boldsymbol{x}) = 0.$$

The main advantage of Definition (9) for $b_h(\boldsymbol{u}, \boldsymbol{v})$ is the following: for a given $\boldsymbol{V}_\sigma$, it only involves values $\boldsymbol{U}_{\sigma'}$ with $\sigma' \in \mathcal{F}_K$, which means that, using a Newton-Raphson method, the stencil of the Jacobian matrix issued from the trilinear term is block-diagonal, similarly to the diffusion terms (note that it leads to cross dependencies between all the components of the velocities).

3 Study of the linear systems

3.1 The coupled velocity-pressure linear system

Let us now detail the construction of the linear system which is directly issued from (6) in the case where $b_h = 0$ or issued from the Newton method applied to (6) if $b_h \neq 0$. For any finite set E, we denote by $\#E$ its cardinal.

This system of linear equations is obtained, first selecting $\boldsymbol{v}$ in the first equation of (6) with one component equal to 1 and all the other ones equal to 0, then selecting q in the second equation of (6) with one component equal to 1 and all the other ones equal to 0. Letting $U = ((U_{i,\sigma})_{i=1,\ldots,d,\sigma\in\mathcal{F}_{\rm int}}$ and $P = (P_K)_{K\in\mathcal{M}\setminus\{K_0\}}$, the linear system reads

$$\begin{bmatrix} A & D^t \\ D & 0 \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix}. \tag{10}$$

Let us detail the construction of the matrices A and D, and of the right-hand side R.

For any $K \in \mathcal{M}$, we first define the elementary assembly matrix S_K, whose side is equal to $s_K := \#\mathcal{F}_{K,\rm int}$ (recall that this side is equal to 3 in 2D and 4 in 3D for any interior element K), by

$$(S_K)_{\sigma,\sigma'} = \int_K \Big(\mu \Pi_h \varphi_\sigma \Pi_h \varphi_{\sigma'} + \nu \nabla\varphi_\sigma \cdot \nabla\varphi_{\sigma'}\Big) \mathrm{d}\boldsymbol{x}.$$

Note that the matrix S_K is symmetric positive definite if $\mu > 0$ (transient problems) and only symmetric positive if $\mu = 0$.

We now define the elementary assembly matrix A_K, whose side is equal to ds_K, such that, if $b_h = 0$,

$$(A_K)_{i,\sigma,j,\sigma'} = \begin{cases} (S_K)_{\sigma,\sigma'} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

If $b_h \neq 0$, this matrix is completed with the derivatives of the convection term with respect to the local velocity unknowns $i = 1, \ldots, d$ and $\sigma \in \mathcal{F}_{K,\rm int}$.

We then define, for any element K of the mesh, the rectangular matrix H_K with $d\#\mathcal{F}_{\rm int}$ lines and ds_K columns, such that, at the column associated to the local velocity unknown $(i, \sigma) \in \{1, \ldots, d\} \times \mathcal{F}_{K,\rm int}$, all the components are null except the one that is at the line associated to the global unknown $U_{i,\sigma}$. Then the matrix A in (10) is obtained by assembling the elementary matrices, as follows:

$$A = \sum_{K\in\mathcal{M}} H_K A_K H_K^t.$$

For the line of A associated to the global unknown $U_{i,\sigma}$, non-zero terms may occur at the columns associated to the global unknown $U_{j,\sigma'}$ such that there exists $K \in \mathcal{M}$ with $\sigma, \sigma' \in \mathcal{F}_{K,\rm int}$. If $b_h = 0$, the matrix A is symmetric positive definite; its inverse is a full matrix, so one cannot solve the linear system by eliminating the velocity unknowns.

We define the matrix D_K with ds_K lines and one column (it is then assimilated to a vector), letting for $i \in \{1, \ldots, d\}$ and $\sigma \in \mathcal{F}_{\rm int}$,

$$(D_K)_{i,\sigma} = -\boldsymbol{a}^{(i)}_{K,\sigma}.$$

We then define the rectangular matrix F_K, with $\#\mathcal{M} - 1$ lines and 1 column, by 0 everywhere, except 1 at the line corresponding to the global unknown P_K, for $K \in \mathcal{M} \setminus \{K_0\}$. Then the matrix D in (10) is defined by

$$D = \sum_{K\in\mathcal{M}\setminus\{K_0\}} F_K D_K^t H_K^t.$$

Finally, for any $K \in \mathcal{M}$, let R_K be the elementary right-hand-side issued from (8), under the form of a vector with ds_K components, defined in the case where $b_h = 0$, for all $i \in \{1, \ldots, d\}$ and $\sigma \in \mathcal{F}_{K,\rm int}$ by

$$(R_K)_{i,\sigma} = \boldsymbol{a}^{(i)}_{K,\sigma} \Big(\frac{1}{|K|} \int_K \overline{\boldsymbol{f}} \mathrm{d}\boldsymbol{x}\Big) \cdot (\boldsymbol{x}_\sigma - \boldsymbol{x}_K).$$

In the case where $b_h \neq 0$, R_K is completed by the non-linear terms issued from the Newton method. Then the assembled right hand side in (10) is given by

$$R = \sum_{K\in\mathcal{M}} H_K R_K.$$

As recalled in the introduction, the resolution of (10) is then a difficult problem for large meshes. Direct methods can no longer be used, and iterative methods must be based on efficient preconditioners.

3.2 Hybridisation of the linear system

We construct in this section a linear system, whose the solution directly provides that of (10), and which can be solved in some cases (see the numerical examples) by cheaper methods. As recalled in the introduction of this paper, the method used for constructing this linear system follows the hybridisation method used in [4, 8, 14].
To this purpose, we introduce, for any $K \in \mathcal{M}$, two diagonal matrices E_K and C_K with the same side ds_K, satisfying the following properties:

$$(E_K)_{i,\sigma,i,\sigma} + (E_L)_{i,\sigma,i,\sigma} = 0 \text{ for all } i = 1,\ldots,d, \tag{11}$$

and

$$(C_K)_{i,\sigma,i,\sigma} + (C_L)_{i,\sigma,i,\sigma} = 0 \text{ for all } i = 1,\ldots,d, \tag{12}$$

in the case where $\mathcal{M}_\sigma = \{K, L\}$. The matrix C_k is meant to be invertible (in practice, we let the diagonal terms of C_K be equal to ± 1), whereas, if $\mu > 0$, the choice $E_K = 0$ can be done.

We consider a global vector $\widehat{U}_{K,i,\sigma}$, associated to the component $i \in \{1,\ldots,d\}$ of the velocity defined at the face $\sigma \in \mathcal{F}_{K,\text{int}}$ of $K \in \mathcal{M}$. The number of components of this vector is equal to $\sum_{L\in\mathcal{M}} d\ s_L$; this number is equal to $2d\#\mathcal{F}_{\text{int}}$ since any velocity unknown appears twice at any interior face.
We then define, for any element K of the mesh, in a similar way to the matrix H_K, the rectangular matrix $\widehat{H}_K$ with $\sum_{L\in\mathcal{M}} d\ s_L$ lines and ds_K columns, such that, at the column associated to the local velocity unknown $(i,\sigma) \in \{1,\ldots,d\} \times \mathcal{F}_{K,\text{int}}$, all the components are null except the one that is at the line associated to the global unknown $\widehat{U}_{K,i,\sigma}$.

Let us define the following matrices, using the matrices $E_K, C_K, \widehat{H}_K$ defined in this section and the matrices A_K, H_K, D_K, F_K defined in the previous section:

$$\widehat{A}_K = A_K + E_K \text{ and } \widehat{A} = \sum_{K\in\mathcal{M}} \widehat{H}_K \widehat{A}_K \widehat{H}_K^t,$$

$$\widehat{D} = \sum_{K\in\mathcal{M}\setminus\{K_0\}} F_K D_K^t \widehat{H}_K^t,$$

$$\widehat{C} = \sum_{K\in\mathcal{M}} H_K C_K \widehat{H}_K^t,$$

and the following right-hand side, using the right-hand sides R_K defined in the previous section:

$$\widehat{R} = \sum_{K\in\mathcal{M}} \widehat{H}_K R_K.$$

We consider the following unknown

- $\widehat{U} = (\widehat{U}_{K,i,\sigma})$ for $K \in \mathcal{M}$, component $i \in \{1,\ldots,d\}$ and $\sigma \in \mathcal{F}_{K,\text{int}}$,

- $\widehat{P} = (\widehat{P}_K)$ for $K \in \mathcal{M} \setminus \{K_0\}$,
- $\widehat{W} = (\widehat{W}_{i,\sigma})$ for $i \in \{1, \dots, d\}$ and $\sigma \in \mathcal{F}_{\rm int}$,

solution to the following linear system

$$\begin{bmatrix} \widehat{A} & \widehat{D}^t & \widehat{C}^t \\ \widehat{D} & 0 & 0 \\ \widehat{C} & 0 & 0 \end{bmatrix} \begin{bmatrix} \widehat{U} \\ \widehat{P} \\ \widehat{W} \end{bmatrix} = \begin{bmatrix} \widehat{R} \\ 0 \\ 0 \end{bmatrix}. \tag{13}$$

In the preceding linear system, the equations $\widehat{A}\widehat{U} + \widehat{D}^t\widehat{P} + \widehat{C}^t\widehat{W} = \widehat{R}$ can be seen as the splitting of the equations $(AU + D^tP)_\sigma = R_\sigma$ with $P = \widehat{P}$, which hold for all $\sigma \in \mathcal{F}_{\rm int}$, into two equations, one for K, σ and the other one for L, σ when $\mathcal{M}_\sigma = \{K, L\}$, thanks to the introduction of an additional unknown $\widehat{W}_\sigma$. The velocity unknowns are also doubled, and the equality between the doubled unknowns is ensured by the relation $\widehat{C}\widehat{U} = 0$.
The next paragraphs are providing details on the following points (among others): the elimination of $\widehat{W}$ is done by addition of these two equations (owing to (12)), and then one recovers $(AU + D^tP)_\sigma = R_\sigma$ (owing to (11)); the system (13) is well-posed, and it is possible, under appropriate choices of the matrices E_K, to eliminate $\widehat{U}$ and $\widehat{P}$ in (13), in order to obtain a linear system only on $\widehat{W}$, with the same stencil as the matrix A, and which is symmetric positive definite in some situations.

Indeed, the following properties hold.

1. **Block diagonal property of $\widehat{H}_K$ and $\widehat{A}$.**

   We have the property, for all $K, L \in \mathcal{M}$,

   $$\widehat{H}_K^t \widehat{H}_L = \begin{cases} \mathrm{Id}_K & \text{if } K = L \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

   Moreover, the matrix $\widehat{A}$ has the blocks $\widehat{A}_K$ on the diagonal and is null elsewhere. In the case where all the matrices $(\widehat{A}_K)_{K\in\mathcal{M}}$ are invertible, there holds

   $$\widehat{A}^{-1} = \sum_{K\in\mathcal{M}} \widehat{H}_K \widehat{A}_K^{-1} \widehat{H}_K^t.$$

   This leads to a cheap computation of $\widehat{A}^{-1}$ and fully scalable.

2. **Recovery of the solution to (10).**

   Any solution $(\widehat{U}, \widehat{P}, \widehat{W})$ of (13) must satisfy

   $$\widehat{C}\widehat{U} = \sum_{K\in\mathcal{M}} H_K C_K \widehat{H}_K^t \widehat{U} = 0.$$

   For any $i \in \{1, \dots, d\}$ and $\sigma \in \mathcal{F}_{K,\rm int}$ with $\mathcal{M}_\sigma = \{K, L\}$, this means that

   $$(C_K)_{i,\sigma,i,\sigma} \widehat{U}_{K,i,\sigma} + (C_L)_{i,\sigma,i,\sigma} \widehat{U}_{L,i,\sigma} = 0,$$

   which, together with (12) and the invertibility of C_K and C_L, provides

   $$\widehat{U}_{K,i,\sigma} = \widehat{U}_{L,i,\sigma} := U_{i,\sigma}, \tag{15}$$

denoting by $U_{i,\sigma}$ this common value. Introducing the vector $U = (U_{i,\sigma})_{i=1,\dots,d,\ \sigma\in\mathcal{F}_{\rm int}}$, we then have

$$\widehat{H}_K^t \widehat{U} = H_K^t U \text{ for all } K \in \mathcal{M}. \tag{16}$$

We now multiply by the left the equality $\widehat{A}\widehat{U} + \widehat{D}^t\widehat{P} + \widehat{C}^t\widehat{W} = \widehat{R}$ by the matrix J which is the matricial translation of the addition of the two equations K, i, σ and L, i, σ for $\mathcal{M}_\sigma = \{K, L\}$. This matrix J, which has $d\#\mathcal{F}_{\rm int}$ lines and $2d\#\mathcal{F}_{\rm int}$ columns, is defined by

$$J = \sum_{K\in\mathcal{M}} H_K \widehat{H}_K^t.$$

On each line of J, all the components are null except two of them, equal to 1, which enables the addition of pairs of lines. We then obtain

$$J\widehat{A}\widehat{U} + J\widehat{D}^t\widehat{P} + J\widehat{C}^t\widehat{W} = J\widehat{R}.$$

We then remark that, accounting for (14),

$$J\widehat{A}\widehat{U} = \sum_{K\in\mathcal{M}} H_K \widehat{H}_K^t \sum_{L\in\mathcal{M}} \widehat{H}_L(A_L + E_L)\widehat{H}_L^t\widehat{U} = \sum_{K\in\mathcal{M}} H_K(A_K + E_K)\widehat{H}_K^t\widehat{U}.$$

We apply (16), thus obtaining

$$J\widehat{A}\widehat{U} = \sum_{K\in\mathcal{M}} H_K(A_K + E_K)H_K^t U.$$

Let us now observe that the matrix $\sum_{K\in\mathcal{M}} H_K E_K H_K^t$ vanishes applying (11). We then get

$$J\widehat{A}\widehat{U} = \sum_{K\in\mathcal{M}} H_K A_K H_K^t U = AU.$$

We now compute, again accounting for (14),

$$J\widehat{D}^t\widehat{P} = \sum_{K\in\mathcal{M}} H_K\widehat{H}_K^t \sum_{L\in\mathcal{M}\setminus\{K_0\}} \widehat{H}_K D_L^t F_L^t \widehat{P} = \sum_{K\in\mathcal{M}\setminus\{K_0\}} H_K D_K F_K^t \widehat{P} = D^t\widehat{P},$$

$$J\widehat{R} = \sum_{K\in\mathcal{M}} H_K\widehat{H}_K^t \sum_{L\in\mathcal{M}} \widehat{H}_L R_L = \sum_{K\in\mathcal{M}} H_K R_K = R.$$

The matrix $J\widehat{C}^t$ satisfies

$$J\widehat{C}^t = \sum_{K\in\mathcal{M}} H_K\widehat{H}_K^t \sum_{L\in\mathcal{M}} \widehat{H}_L^t C_L H_L = \sum_{K\in\mathcal{M}} H_K C_K H_K,$$

which vanishes owing to (12). So we get

$$AU + D^t\widehat{P} = R.$$

Turning to the equation $\widehat{D}\widehat{U} = 0$, we get

$$\widehat{D}\widehat{U} = \sum_{K\in\mathcal{M}\setminus\{K_0\}} F_K D_K^t \widehat{H}_K^t\widehat{U} = \sum_{K\in\mathcal{M}\setminus\{K_0\}} F_K D_K^t U = DU = 0,$$

applying (16). So we conclude that $(U, \widehat{P})$ is solution to (10). Since this latter system is invertible, we get $\widehat{P} = P$.

3. **Invertibility of** (13)**.**

   The invertibility of the linear system is proved, if one assumes that the right-hand side is null, this implies that the solution is null too. This is done by assuming that, in (13), we let $\widehat{R} = 0$ (which is obtained if we let $R_K = 0$ for all $K \in \mathcal{M}$). Since this is a particular case of the linear system under study, the conclusions obtained in the preceding paragraphs, that any solution of this linear system is also a solution to (10), are remaining true in this case. Then, the vectors U issued owing to the preceding computations from $\widehat{U}$ and $\widehat{P}$, are solution to (10) with $R = 0$, since R is computed from null R_K. We recall that the linear system (10) is invertible, which implies that $U = 0$ and $\widehat{P} = 0$. From $U = 0$, we deduce by (15) that $\widehat{U} = 0$, which proves from $\widehat{A}\widehat{U} + \widehat{D}^t\widehat{P} + \widehat{C}^t\widehat{W} = \widehat{R}$ that

   $$\widehat{C}^t\widehat{W} = 0.$$

   The preceding relations are equivalent to $(\widehat{C}_K)_{i,\sigma,i,\sigma}\widehat{W}_{i,\sigma} = 0$ and $(\widehat{C}_L)_{i,\sigma,i,\sigma}\widehat{W}_{i,\sigma} = 0$, for any $\sigma \in \#\mathcal{F}_{\text{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$. which shows that $\widehat{W} = 0$ (recall that the matrices $\widehat{C}$ must have a non-zero diagonal).

   The linear system (13) is therefore invertible, and its resolution provides the solution to (10).

4. **Elimination of** $(\widehat{U}, P)$**.**

   Assuming that, for all $K \in \mathcal{M}$, all the eigenvalues of the symmetric matrix $\widehat{A}_K$ are either strictly positive or strictly negative, let us proceed to the elimination of $\widehat{U}$ and P. We first have

   $$\widehat{U} = \widehat{A}^{-1}(-\widehat{D}^tP - \widehat{C}^t\widehat{W} + \widehat{R}).$$

   This yields

   $$\widehat{U} = \sum_{K\in\mathcal{M}} \widehat{H}_K\widehat{A}^{-1}(-\widehat{D}^tP - \widehat{C}^t\widehat{W} + \widehat{R}),$$

   Then we have

   $$\widehat{D}\widehat{A}^{-1}(-\widehat{D}^tP - \widehat{C}^t\widehat{W} + \widehat{R}) = 0.$$

   Let us compute the matrix $B = \widehat{D}\widehat{A}^{-1}\widehat{D}^t$. Using the property

   $$\widehat{H}_K^t\widehat{A}^{-1}\widehat{H}_L = \begin{cases} \widehat{A}_K^{-1} & \text{if } K = L \\ 0 & \text{otherwise,} \end{cases}$$

   we get

   $$B = \sum_{K\in\mathcal{M}\setminus\{K_0\}} F_K D_K^t \widehat{A}_K^{-1} D_K F_K^t.$$

   We then get that B is the diagonal matrix with the values $B_K := D_K^t\widehat{A}_K^{-1}D_K$ on the diagonal. Letting $\underline{\lambda}_K$ be the smaller absolute value of the eigenvalues of $\widehat{A}_K^{-1}$, we get that

   $$|B_K| \geq \underline{\lambda}_K\|D_K\|_2 > 0,$$

   since there exists at least one component of D_K which is different from 0. So the diagonal matrix B is invertible, and we can write

   $$B^{-1} = \sum_{K\in\mathcal{M}\setminus\{K_0\}} \frac{1}{B_K} F_K F_K^t,$$

and

$$P = B^{-1}\widehat{D}\widehat{A}^{-1}(-\widehat{C}^t\widehat{W} + \widehat{R}).$$

We then obtain

$$\widehat{C}\widehat{A}^{-1}(-\widehat{D}^t P - \widehat{C}^t\widehat{W} + \widehat{R}) = 0,$$

which leads, denoting $G = \widehat{C}\Big(\widehat{A}^{-1} - \widehat{A}^{-1}\widehat{D}^t B^{-1}\widehat{D}\widehat{A}^{-1}\Big)\widehat{C}^t$ and $S = \widehat{C}\Big(\widehat{A}^{-1} - \widehat{A}^{-1}\widehat{D}^t B^{-1}\widehat{D}\widehat{A}^{-1}\Big)\widehat{R}$, to

$$G\widehat{W} = S.$$

The matrix G is then invertible, since this resolution process is equivalent to the initial linear system (under the above assumption on $\widehat{A}_K$).

5. **Stencil of G**

   Under the same assumption as previously (for all $K \in \mathcal{M}$, all the eigenvalues of the symmetric matrix $\widehat{A}_K$ are either strictly positive or strictly negative), a simple computation using (14) and $F_K^t F_L = 1$ if $K = L$ and 0 otherwise, leads to

   $$G = \sum_{K\in\mathcal{M}} H_K G_K H_K^t,$$

   with, for all $K \in \mathcal{M} \setminus K_0$,

   $$G_K = C_K\Big(\widehat{A}_K^{-1} - \frac{1}{B_K}\widehat{A}_K^{-1} D_K D_K^t \widehat{A}_K^{-1}\Big)C_K,$$

   and

   $$G_{K_0} = C_{K_0}\widehat{A}_{K_0}^{-1} C_{K_0}.$$

   This shows that the assembling of G leads to the same stencil as that of $A = \sum_{K\in\mathcal{M}} H_K A_K H_K^t$ (in the case where the matrix A_K is full).

6. **Case $\mu > 0$ and $b_h = 0$.**

   In the case $\mu > 0$ and $b_h = 0$, all the matrices A_K are symmetric positive definite and we let $E_K = 0$. Let us show that the resulting matrix G is symmetric positive definite. Indeed, for any vector $\widehat{W}$, let us compute

   $$a = \widehat{W}^t G\widehat{W}.$$

   Denoting by $Z_K = C_K H_K^t \widehat{W}$, and defining the scalar product $\langle X, Y\rangle_K = X^t\widehat{A}_K^{-1}Y$, we get that

   $$a = \sum_{K\in\mathcal{M}\setminus\{K_0\}} \Big(\langle Z_K, Z_K\rangle_K - \frac{(\langle Z_K, D_K\rangle_K)^2}{\langle D_K, D_K\rangle_K}\Big) + \langle Z_{K_0}, Z_{K_0}\rangle_{K_0}.$$

   The Cauchy-Schwarz inequality implying

   $$(\langle Z_K, D_K\rangle_K)^2 \le \langle Z_K, Z_K\rangle_K \langle D_K, D_K\rangle_K,$$

   we get that $a \ge 0$. Since we proved above that, under a weaker hypothesis, the matrix G is invertible, it is then positive symmetric definite.

7. **Computation of E_K in the case $\mu = 0$.**

   Different strategies can be used. One of them consists in partitioning $\mathcal{M}$ in $\mathcal{M}_1 \cup \mathcal{M}_2$, such $\mathcal{M}_2$ is the set of all the neighbours of all $K \in \mathcal{M}_1$. Then for all $K \in \mathcal{M}_1$, we let $E_K = -\lambda\text{Id}$ with λ

larger than all the eigenvalues of A_K. Then, for all $K \in \mathcal{M}_2$ and $\sigma \in \mathcal{F}_{K,\mathrm{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$, if $L \in \mathcal{M}_1$ (such a L exists by construction), we set $(E_K)_{\sigma,\sigma} = \lambda$. If $L \notin \mathcal{M}_1$, we set $(E_K)_{\sigma,\sigma} = 0$.

Then Property (11) holds, as well as the fact that all the matrices $(A_K)_{K\in\mathcal{M}_1}$ are symmetric and have all their eigenvalues strictly negative and all the matrices $(A_K)_{K\in\mathcal{M}_2}$ are symmetric positive definite.

In conclusion of this section, we can use the following method, called the **hybrid method** for solving (10):

1. One computes the matrix G and the right-hand side S as defined above (this leads to cheap computations).

2. One then solves the linear system $G\widehat{W} = S$ by a direct method for the small cases or by an iterative method for the larger ones. Note that, in the case where $\mu > 0$ and $b_h = 0$, a simple preconditioned conjugate gradient solver may be used, and the side of this linear system is smaller than that of (10) with a stencil similar to that of A, which is a part of the matrix of (10).

3. One then recovers P and U by the preceding relations which only leads to cheap and fully scalable computations.

The numerical section provides a few comparisons of this method with the resolution of (10) by a solver with unknowns (U, P).

4 Numerical results

4.1 Numerical convergence of the scheme

Although the Crouzeix-Raviart scheme (6) is highly standard in the transient or steady Stokes case, the implementation for the right hand side through the reconstruction $\widehat{\Pi}_h$ is not completely classical. Note that, if $\overline{f}$ is a constant vector (which means that the velocity is null and that the gradient of the exact pressure is equal to $\overline{f}$), a standard computation of the right-hand side by the integration of $\overline{f}$ against the Crouzeix-Raviart basis functions provides a significant error on the velocity field. On the contrary, owing to the reconstruction $\widehat{\Pi}_h$, we obtain a null numerical velocity and the exact pressure field (at the machine precision).
Let us also observe that the non-linear term (9) introduced by [10] is not so well-known, and that it is therefore interesting to check, on the analytical Green-Taylor solution, the numerical convergence of this scheme, independently of the algebraic method used for solving the linear systems.
First letting $d = 2$, we assume that the analytical solution is given by $\overline{f} = 0$,

$$\overline{\boldsymbol{u}}(\boldsymbol{x}, t) = \mathrm{Re}\begin{pmatrix} -\cos(2\pi(x_1 + \frac{1}{4}))\sin(2\pi(x_2 + \frac{1}{2}))\exp(-8\pi^2 t) \\ \sin(2\pi(x_1 + \frac{1}{4}))\cos(2\pi(x_2 + \frac{1}{2}))\exp(-8\pi^2 t) \end{pmatrix} \tag{17}$$

and

$$\overline{p}(\boldsymbol{x}, t) = -\frac{\mathrm{Re}^2}{4}\Big(\cos(4\pi(x_1 + \frac{1}{4})) + \cos(4\pi(x_2 + \frac{1}{2}))\Big)\exp(-16\pi^2 t). \tag{18}$$

We then implement the values $\overline{\boldsymbol{u}}(\boldsymbol{y}, 0)$ as initial numerical value at all the nodes of the mesh $\boldsymbol{y}$, and the values $\overline{\boldsymbol{u}}(\boldsymbol{y}_b, t^{(n)})$ at all the boundary nodes of the mesh $\boldsymbol{y}_b$ and at the discrete times $t^{(n)} = n\Delta t$. The hybrid method and a direct solver are used for these computations which are not dedicated to observe computing performances. Letting Re = 100 and the final time be equal 0.01, we find the numerical errors given by Table 1 with different meshes and time steps.

Δt	h	errl2U	ratio	errl2P	ratio
1.25e-04	0.2500	0.277E+02	-	0.487E+03	-
3.13e-05	0.1250	0.854E+01	1.70	0.262E+03	0.89
7.81e-06	0.0625	0.315E+01	1.44	0.112E+03	1.23
1.95e-06	0.0312	0.918E+00	1.78	0.351E+02	1.67
4.88e-07	0.0156	0.240E+00	1.94	0.986E+01	1.83
1.22e-07	0.0078	0.608E-01	1.98	0.299E+01	1.72

Table 1: Numerical errors in the case of the 2D Green-Taylor analytical solution of the Navier-Stokes problem.

The meshes are those labelled from 1 to 6 of the triangular family Mesh1 used in the 2D benchmark [11]. The time step Δt and the mesh size h are such that $\Delta t/h^2$ is constant. The numerical errors are computed at the nodes for the velocities, and at the centre of gravity of the triangles for the pressures. In Table 1, the ratios are computed by the formula $\log(E_{i-1}/E_i)/\log(2)$, where E_i is a value taken in the column "errl2U" or "errl2P" and E_{i-1} is the value immediately above in the table.

We observe in Table 1 that the numerical order of convergence tends to 2 for the velocity errors and the finest meshes, and to a value greater than 1 for the pressure errors, as it is currently observed by numerical schemes in this case.
We now turn out to a 3d case ($d = 3$) with Re = 100 and the final time equal to 0.01. In order to ensure that the 3D meshes present the same regularity factor, the tetrahedral mesh is obtained by splitting in 6 tetrahedra each cube of a uniform cubic mesh of the test domain. The common side of all the cubes of the cubic mesh have all the same side h.
The first 3D numerical test concerns a Stokes problem case, where the analytical solution is an extension to the 3D case of the preceding Green-Taylor test. The first two components of the velocity are given by (17) extended for all $x_3 \in [0, 1]$, the third component is equal to 0 on the whole domain as well as the pressure (recall that in the Green-Taylor test, the non-linear term is balanced by the pressure gradient). Using the hybrid method, and a conjugate gradient solver with the "boomer AMG" preconditioners, we obtain the results provided by Table 2.

Δt	h	errl2U	ratio	errl2P	ratio
1.00e-4	1.38e-1	3.44	-	33	-
2.50e-5	6.88e-2	0.85	2.02	17	0.96
6.25e-6	3.44e-2	0.22	1.95	8.1	1.07
1.56e-6	1.72e-2	5.4e-2	2.03	4.0	1.02

Table 2: Numerical errors in the case of the 3D Green-Taylor analytical solution of the Stokes problem.

The convergence orders shown in Table 2 are similar to those observed in Table 1. Turning to a 3D Navier-Stokes case, we again consider the extension to the 3D case of the 2D Green-Taylor test. The first two components of the velocity are again given by (17) for any $x_3 \in [0, 1]$, the third component is again equal to 0 on the whole domain, and the pressure is given by (18) for any $x_3 \in [0, 1]$. Again, applying the same method for solving the linear systems as in the previous test case, we obtain the results provided by Table 3.
The convergence orders shown in Table 3 show a light loss of convergence order in this case, compared to the ones observed in Table 2, although they give a numerical confirmation of the efficiency of the scheme.

These tests validate the use of the Crouzeix-Raviart scheme (6) in association with the trilinear term (9), in 2D and 3D cases. The remaining part of the numerical section is now devoted to 2D and 3D

Δt	h	errl2U	ratio	errl2P	ratio
1.56e-4	6.88e-2	10.7	-	297	-
3.91e-5	3.44e-2	4.10	1.38	143	1.05
9.77e-6	1.72e-2	1.28	1.68	48.5	1.56

Table 3: Numerical errors in the case of the 3D Green-Taylor analytical solution of the Navier-Stokes problem.

comparisons of the computing performances for solving the linear systems, with or without the use of the hybrid method, in association with a variety of linear solvers.

4.2 Comparison of algebraic methods and solvers on the transient Stokes problem

The aim of this section is to assess the interest of the hybrid method in the case of transient Stokes problems (that means that $\mu = 1/\Delta t > 0$ and $b_h = 0$). In this case, as seen above, the hybrid method leads to positive symmetric definite linear systems, compared to the non-hybrid method, which only provides symmetric linear systems which are not positive and larger.

We performed the computation using a direct sequential solver, the only purpose of these tests being to assess the gain of computing time per time step due to smaller linear systems with the hybrid method compared to the linear systems without the hybrid method. We consider the 3D Green-Taylor Stokes problem, with analytical solution given by (17) and $p = 0$. The linear systems are solved with a simple Gaussian elimination with natural ordering, the time step is equal to 5.10^{-4} and various meshes are used (see Table 4). The decrease in the size of the linear systems leads to a clear diminution in the computing time.

Ncv	not hybrid	hybrid
46	2.9e-3	3.7e-3
384	1.7e-1	1.2e-1
3062	2.5e+1	1.4e+1
24576	3.4e+3	2.4e+3

Table 4: Computation time in seconds per linear system solved by Gaussian elimination in the case of the 3D Green-Taylor analytical solution of the Stokes problem. Ncv denotes the number of tetrahedra.

We now turn to the evaluation of the possibility to use parallel solvers with or without the hybrid method. All the tests are done using the HYPRE/Euclid library for the solvers and the preconditioners, on a computer with 16 cpus.

Conjugate gradient with algebraic multi grid preconditioners in 2D.

We study the possibility of using the BoomerAMG preconditioners, which is known to provide an optimal speed-up in the case of the linear systems issued from diffusion operators. The numerical choices are the following:

- The mesh is “Mesh1-7” of the triangular family Mesh1 used in the 2D benchmark [11] (it corresponds to a mesh size equal to $h = 3.90625 \cdot 10^{-3}$, which leads to 917 504 triangles),
- Smoother algorithm : Hybrid symmetric Gauss-Seidel or SSOR

- Parallel coarsening algorithm : one-pass Ruge-Stueben coarsening on each processor, no boundary treatment.

We observe that, without hybridisation, non-convergence is observed in all tested cases.
On the contrary, using hybridisation, the convergence of the method is obtained. In Table 5, we provide the computing times needed for the resolution of one linear system (in this transient Stokes problem with constant time step, all the linear systems have the same matrix) for two different values of the time step.

proc	$\Delta t = 0.0001$ time/iter (s)	speed-up	$\Delta t = 0.0512$ time/iter (s)	speed-up
1	163	-	656	-
2	90	1.81	331	1.98
4	47	1.91	166	1.99
8	25	1.88	89	1.87
16	19	1.32	65	1.37

Table 5: Computation time with hybridisation, using conjugate gradient with boomerAMG

CG, BCGS et GMRES with ILU in 2D.

We now consider the case where we use different linear solvers (we use “Mesh1-7” with $\Delta t = 0.0001$):

- CG : preconditioned conjugate gradient,
- BCGS : Bi-conjugate gradient with stabilization,
- GMRES,

with the Euclid/ ILU preconditioners. Recall that the efficiency of ILU is mainly lost in the case of multi-processor computations, but that it remains in any case much greater that that of boomer AMG. A parameter of ILU is the filling degree (from 1 to 4 in our tests).
We again observe that no convergence is obtained using conjugate gradient without hybridisation. We show in Table 6 the results obtained using conjugate gradient with hybridisation. These results show a lower speed-up compared to the use of boomer AMG, but better absolute performances. Let us finally observe that no results were obtained with increasing the filling degree of the ILU method with more than one processor.

proc	ILU	time/iter (s)
1	1	25
1	2	19
1	3	22
1	4	18
2	1	18
4	1	11
8	1	11
16	1	8.4

Table 6: Computation time with hybridisation, using conjugate gradient with ILU

We also used the BCGS and GMRES methods without hybridisation. We then get no result with more that 2 processors, the best performance being 23 s per iteration with 4th degree of ILU, BCGS and 1 processor.

Numerical results in 3D

We only obtained numerical results using ILU preconditioners and only one processor.

In these conditions, the results without hybridisation with BCGS were better than those with conjugate gradient, whatever be the degree of filling of the ILU method: for example, using degree 2 and BCGS, the time per iteration without hybridisation is 376 s with Ncv $= 1\,572\,864$ and $\Delta t = 0.0005$ for the Green-Taylor problem in Stokes conditions, where it is equal to 498 s with conjugate gradient and hybridisation. Additional tests seem to be necessary for improving this comparison.

4.3 Comparison of linear solvers on the steady lid driven cavity test in 2D

This test is dedicated to the comparison of the efficiency of the different algebraic solvers in the case of the steady lid driven cavity with Re $= 1000$, in 2 space dimensions. We again consider the mesh named "Mesh1-7" of the triangular family Mesh1 used in the 2D benchmark [11] (it corresponds to a mesh size equal to $h = 3.90625 \cdot 10^{-3}$ and Ncv $= 917\,504$).
The non-linear system provided by the scheme is approximated by the Newton method. Since the resulting linear systems are no longer symmetric positive, we cannot use the conjugate gradient method; we use the GMRES method with a convergence threshold equal to 10^{-11} in association with an ILU preconditioners with filling degree 2 to 8. This preconditioners has been shown in several tests to provide a sufficient efficiency, letting the filling degree increase [5]. Unfortunately, this efficiency falls down on parallel architectures, so this test is only considered with one processor.

In order to assess the additional difficulty issued from the non-linear terms, we first consider the Stokes problem (in this case, only one Newton iteration is needed, and the linear system is in fact symmetric, but not positive).
The numerical results presented in Table 7 show that the computation time is largely lower with the hybrid method, compared to the results without hybridisation, and that the comparison shows higher contrasts with low filling degree.
This observation remains true in the Navier-Stokes case. To compare the two methods in the Navier-Stokes case, the GMRES threshold has to be reduced to 10^{-8} to ensure the convergence of the linear solver when the non-hybrid approach is employed. The convergence threshold required for the non-linear iterations is equal to 5.10^{-7}. In this case and starting from a fluid flow at rest, 10 and 9 Newton iterations are needed respectively for the scheme with and without hybridisation. The results of Table 8 show that, despite one additional Newton iteration, the hybrid method converges about twice quicker than the standard approach.

5 Conclusions

In 2D and on different test cases of the Crouzeix-Raviart scheme, the numerical results show an advantage for using the hybridisation method for solving the coupled linear systems issued from the Newton-Raphson method or from the Stokes problem.

In particular, the hybridisation method allow the use of conjugate gradient solvers.

Additional tests must be done in 3D in order to assess the influence of hybridisation on solvers performances.

References

[1] J. Aghili, S. Boyaval, and D. A. Di Pietro. Hybridization of mixed high-order methods on general meshes and application to the Stokes equations. *Comput. Methods Appl. Math.*, 15(2):111–134, 2015.

Table 7: Stokes lid driven cavity

iLU	Hybrid.	time (s)
2	without	2025
	with	547.9
3	without	747.9
	with	477.4
4	without	391.4
	with	231.6
5	without	211.1
	with	152.9
6	without	163.5
	with	89.13
7	without	128.8
	with	65.96
8	without	93.04
	with	69.76

Table 8: Navier-Stokes lid driven cavity

iLU	Hybrid.	time (s)
2	without	24120
	with	9531
3	without	9517
	with	4446
4	without	5420
	with	3169
5	without	3244
	with	1696
6	without	2332
	with	1028
7	without	1703
	with	768.0
8	without	1398
	with	717.1

[2] P. R. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. Mary. Performance and scalability of the block low-rank multifrontal factorization on multicore architectures. *ACM Trans. Math. Software*, 45(1):Art. 2, 26, 2019.

[3] P. R. Amestoy, I. S. Duff, J.-Y. L'Excellent, and J. Koster. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM J. Matrix Anal. Appl.*, 23(1):15–41, 2001.

[4] Z. Chen. Equivalence between and multigrid algorithms for nonconforming and mixed methods for second-order elliptic problems. *East-West J. Numer. Math.*, 4(1):1–33, 1996.

[5] E. Chénier, R. Eymard, R. Herbin, and O. Touazi. Collocated finite volume schemes for the simulation of natural convective flows on unstructured meshes. *Internat. J. Numer. Methods Fluids*, 56(11):2045–2068, 2008.

[6] E. Chow and A. Patel. Fine-grained parallel incomplete LU factorization. *SIAM J. Sci. Comput.*, 37(2):C169–C193, 2015.

[7] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 7(R-3):33–75, 1973.

[8] R.G Durán. Mixed finite elements. In D Boffi and L. Gastaldi, editors, *Mixed finite elements, compatibility conditions, and applications: lectures given at the CIME Summer School held in Cetraro, Italy, June 26-July 1, 2006*, volume 1939 of *Lecture Notes in Mathematics*, pages 1–44. Springer, 2008.

[9] R. Eymard, P. Feron, and C. Guichard. Family of convergent numerical schemes for the incompressible Navier-Stokes equations. *Math. Comput. Simulation*, 144:196–218, 2018.

[10] L. Gastaldo, R. Herbin, and J.-C. Latché. An unconditionally stable finite element-finite volume pressure correction scheme for the drift-flux model. *M2AN Math. Model. Numer. Anal.*, 44(2):251–287, 2010.

[11] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications V*, pages 659–692. ISTE, London, 2008.

[12] A. Linke. On the role of the Helmholtz decomposition in mixed methods for incompressible flows and a new variational crime. *Comput. Methods Appl. Mech. Engrg.*, 268:782–800, 2014.

[13] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[14] M. Vohralík, J. Maryška, and O. Severýn. Mixed and nonconforming finite element methods on a system of polygons. *Appl. Numer. Math.*, 57(2):176–193, 2007.