Mixed Discontinuous Galerkin methods for Darcy flow

F. Brezzi[1,2], T.J.R. Hughes[3], L.D. Marini[1,2], and A. Masud[4]

Abstract

We consider a family of mixed finite element discretizations of the Darcy flow equations using totally discontinuous elements (both for the pressure and the flux variable). Instead of using a jump stabilization as it is usually done for DG methods (see e.g. [3], [13] and the references therein) we use the stabilization introduced in [18], [17]. We show that such stabilization works for discontinuous elements as well, provided both the pressure and the flux are approximated by local polynomials of degree ≥ 1, without any need for additional jump terms. Surprisingly enough, after the elimination of the flux variable, the stabilization of [18], [17] turns out to be in some cases a sort of jump stabilization itself, and in other cases a stable combination of two originally unstable DG methods (namely, Bassi-Rebay [4] and Baumann-Oden [6]).

1 Introduction

Mixed finite element methods based on the Galerkin formulation have become an increasingly popular way to discretize the Darcy flow equations (see, e.g., as kindly suggested by a reviewer, [11], [12], [5]). Within this framework, the proper trial solution spaces for velocity and pressure are the classical Sobolev spaces $H(\mathrm{div})$ and $L^2/\mathbb{R}$, respectively. Finite dimensional subspaces of $H(\mathrm{div})$ and $L^2/\mathbb{R}$ are referred to as conforming. Stability conditions preclude many desirable combinations of interpolations but successful combinations have been derived, namely, the RT (Raviart and Thomas [19]) and BDM (Brezzi, Douglas and Marini [7]) families, which require continuity of the normal component of velocity in combination with specific discontinuous pressure interpolation. In an effort to enlarge the spectrum of possibilities, Masud and Hughes [18] introduced a stabilized finite element formulation in which an appropriately weighted residual of the Darcy law

[1]Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia (Italy)

[2]IMATI del CNR, Via Ferrata 1, 27100 Pavia (Italy)

[3]Institute for Computational Engineering and Sciences, The University of Texas at Austin, 201 East 24th Street, ACES 6.412, Austin, TX 78712-0027 (USA)

[4]Department of Civil & Materials Engineering, University of Illinois at Chicago, Chicago, IL 60607 (USA)

is added to the standard mixed formulation. In [18] it was proved that this stabilized method is convergent for:

(1) All combinations of conforming velocity and continuous pressure.

(2) All combinations of quadratic or higher-order conforming velocity and discontinuous pressure of any order.

Furthermore, although not specifically dealt with in [18], it is also apparent that the method is convergent for:

(3) All combinations of discontinuous velocity and continuous pressure.

Numerical studies performed in [18] provided confirmation of the theoretical results. An interesting feature of the formulation is that the additional stabilization term does not involve a stabilization parameter depending on the type of element, the mesh, or the constitutive coefficients. It is simply a nondimensional, universal constant that may be selected once and for all. (The value 1/2 was utilized and advocated in [18].) Consequently, the formulation is remarkably "clean".

In Hughes-Masud-Wan [17], the method was extended within the Discontinuous Galerkin (DG) framework in order to consider various combinations of discontinuous velocity and pressure. The convergence proof relied upon a commonly used device in DG methodology, an additional stabilization term that improves control of pressure jumps across element interfaces (see, e.g., [3], [4], [13]). The term is a bilinear form in which the arguments are the jumps in pressure trial solution and pressure weighting function. The bilinear form needs to be weighted by an appropriately selected stabilization parameter that depends on element type, the mesh, and constitutive coefficients. Currently, these parameters can only be estimated by dimensional scaling and convergence arguments, which only provide crude estimates, especially in the case of discontinuous constitutive parameters and anisotropic meshes. Consequently, the practical utility of such methods is diminished because of the ambiguity associated with the selection of the pressure jump stabilization parameter. However, anticipating (and hoping) that some combinations of discontinuous interpolations would be convergent without the pressure jump stabilization term, numerical convergence studies were performed with some one- and two-dimensional elements that led us to conjecture that all combinations of discontinuous velocity and pressure, involving linear and/or higher-order polynomials, were convergent. In this paper, we explore this issue mathematically. Our main result is an affirmation of the conjecture. The theoretical approach enlarges the scope, compared to that considered in [17]. Here, we consider two classes of DG methods that, in the second-order form of Darcy flow, may be identified with the Bassi-Rebay/interior penalty method ([4], [2], [15], [20]) properly stabilized, and with a combination of the Bassi-Rebay and Baumann-Oden method ([6]). We also prove convergence for a range of values of the stabilization constant. Our theoretical results are confirmed by some numerical convergence studies. In all cases, the L^2 rates of convergence for velocity, pressure, and pressure derivatives are at least the minimum of $k+1$ and ℓ, where k and ℓ are the polynomial orders of the velocity and pressure approximations, respectively. In cases where adjoint consistency can be invoked, and k is greater than or equal to $\ell - 1$, the pressure is proved to converge at rate $\ell + 1$. Based on our results, we can assert that the mixed stabilized DG formulation of Darcy flow is

mathematically viable, and we also believe it may be practically useful. It is a "clean" formulation that does not involve any ambiguous stabilization parameters, it generalizes and encompasses all the successful elements described in [17] and [18], and it provides many convergent combinations of discontinuous velocity and pressure interpolations that would otherwise be unstable in the classical mixed DG format.

An outline of the remainder of the paper follows: In Section 2, we describe the boundary value problem of Darcy flow that we consider subsequently. In Section 3, we describe the mixed stabilized DG formulation, we prove convergence and obtain error estimates. We conclude in Section 4 with some numerical convergence studies that support the theory.

In the sequel we shall follow the usual notation for Sobolev spaces (see e.g. Ciarlet [10]). We will also denote by $(\cdot\,,\,\cdot)$ the usual $L^2(\Omega)$-inner product of both scalars and vectors.

2 Darcy flow

Consider the problem

$$\begin{cases} \mathbf{u} &= -\kappa\nabla p & \text{in } \Omega, \\ \operatorname{div}\mathbf{u} &= f & \text{in } \Omega, \\ \mathbf{u}\cdot\mathbf{n} &= g & \text{on } \Gamma = \partial\Omega. \end{cases} \tag{1}$$

where $\kappa \in \mathcal{L}^\infty(\Omega)$ is a given permeability coefficient, $f \in L^2(\Omega)$ is a given source term, and $g \in H^{1/2}(\Gamma)$ is a given prescribed flux at the boundary. All the methods that we are going to describe, and all the analysis that we will carry out, hold independently of the number of dimensions. Here however, for simplicity, we will use a two-dimensional *notation*. To go to the three-dimensional case we just have to use *faces* instead of *edges* and so on. Problem (1) can also be written, eliminating $\mathbf{u}$, as

$$\begin{cases} -\operatorname{div}(\kappa\nabla p) &= f & \text{in } \Omega, \\ \kappa(x)\nabla p\cdot\mathbf{n} &= -\,g & \text{on } \Gamma = \partial\Omega. \end{cases} \tag{2}$$

With the usual compatibility assumption that

$$\int_\Omega f\,\mathrm{d}\mathbf{x} = \int_\Gamma g\,\mathrm{d}s, \tag{3}$$

it is well known that problem (2) has a unique solution p in the space $H^1(\Omega)/\mathbb{R}$ made of the functions belonging to $H^1(\Omega)$ having zero mean value in Ω.

Going back to the original formulation (1) we see that $\mathbf{u}$ is unique as well, and it clearly belongs to the space $H(\operatorname{div};\Omega)$, made of vectors of $(L^2(\Omega))^2$ having their divergence in $L^2(\Omega)$. Using De Giorgi-Nash regularity, we also have that p is Hölder continuous, although we are not going to use it.

The regularity of the solution p (as well as the regularity of $\mathbf{u}$) will depend, in general, on the regularity of Ω, f, g, and the permeability κ. See for instance [14] and the references therein. For the sake of simplicity, we will only consider here the case of piecewise constant

κ. However, we point out that in the case of a more general permeability coefficient we can always approximate it by means of a piecewise constant, substituting κ by its average in each element.

3 Mixed stabilized DG formulation

Let $\mathcal{T}_h$ be a regular family of decompositions of Ω into elements T, triangles or quadrilaterals; let h_T denote the diameter of T, and $h = \max_{T\in\mathcal{T}_h} h_T$. In order to write a discontinuous finite element approximation of problem (1) we need first to introduce typical tools such as *jumps* and *averages* of scalar and vector valued functions across the edges of $\mathcal{T}_h$. Following the notation of [8], [9], [3], let e be an interior edge shared by elements T_1 and T_2. Define the unit normal vectors $\mathbf{n}^1$ and $\mathbf{n}^2$ on e pointing exterior to T_1 and T_2, respectively. For a function φ, piecewise smooth on $\mathcal{T}_h$, with $\varphi^i := \varphi|_{T_i}$ we set

$$\{\varphi\} = \frac{1}{2}(\varphi^1 + \varphi^2), \quad [\![\, \varphi \,]\!] = \varphi^1\mathbf{n}^1 + \varphi^2\mathbf{n}^2 \quad \text{on } e \in \mathcal{E}_h^\circ, \tag{4}$$

where $\mathcal{E}_h^\circ$ is the set of interior edges e. For a vector valued function $\mathbf{v}$, piecewise smooth on $\mathcal{T}_h$, we define $\mathbf{v}^1$ and $\mathbf{v}^2$ analogously, and set

$$\{\mathbf{v}\} = \frac{1}{2}(\mathbf{v}^1 + \mathbf{v}^2), \quad [\![\, \mathbf{v} \,]\!] = \mathbf{v}^1\cdot\mathbf{n}^1 + \mathbf{v}^2\cdot\mathbf{n}^2 \quad \text{on } e \in \mathcal{E}_h^\circ. \tag{5}$$

Notice that the jump $[\![\, \varphi \,]\!]$ of the scalar function φ across $e \in \mathcal{E}_h^\circ$ is a vector parallel to the normal to e, and the jump $[\![\, \mathbf{v} \,]\!]$ of the vector function $\mathbf{v}$ is a scalar quantity. The advantage of these definitions is that they do not depend on assigning an ordering to the elements T_i. For $e \in \mathcal{E}_h^\partial$, the set of boundary edges, we set

$$\{\mathbf{v}\} = \mathbf{v}, \quad [\![\, \varphi \,]\!] = \varphi\mathbf{n} \quad \text{on } e \in \mathcal{E}_h^\partial. \tag{6}$$

We do not require either of the quantities $\{\varphi\}$ or $[\![\, \mathbf{v} \,]\!]$ on boundary edges, and leave them undefined.

Next, with any integer $k \geq 1$ we associate the discontinuous finite element space for vector valued functions:

$$V_h^k = \{\mathbf{v} \in [L^2(\Omega)]^2 : \ \mathbf{v}_{|T} \in [P_k(T)]^2 \quad \forall T \in \mathcal{T}_h\}, \tag{7}$$

where, as usual, P_k is the space of polynomials of degree $\leq k$. Similarly, with any integer $\ell \geq 1$ we associate the space for scalars:

$$Q_h^\ell = \{q \in L^2(\Omega)/\mathbb{R} : \ q_{|T} \in P_\ell(T) \quad \forall T \in \mathcal{T}_h\}. \tag{8}$$

In practical computations the discrete pressure is usually set to be equal to zero at some given point (that is, the zero-mean value condition is never enforced as such). Setting

$$\begin{aligned}
a_h(\mathbf{u}_h, \mathbf{v}_h) &= \sum_{T\in\mathcal{T}_h} \int_T \kappa^{-1}\mathbf{u}_h\cdot\mathbf{v}_h \, \mathrm{d}\mathbf{x} \\
b_h(p_h, \mathbf{v}_h) &= \sum_{T\in\mathcal{T}_h} \int_T p_h \operatorname{div}\mathbf{v}_h \, \mathrm{d}\mathbf{x} - \sum_{e\in\mathcal{E}_h^\circ} \int_e \{p_h\}[\![\, \mathbf{v}_h \,]\!] \, \mathrm{d}s - \sum_{e\in\mathcal{E}_h^\partial} \int_e \{\mathbf{v}_h\}\cdot[\![\, p_h \,]\!] \, \mathrm{d}s \\
(f, q_h) &= \int_\Omega f q_h \, \mathrm{d}\mathbf{x}, \quad \langle g, q_h\rangle = \sum_{e\in\mathcal{E}_h^\partial} \int_e g\, q_h \, \mathrm{d}s,
\end{aligned} \tag{9}$$

the discrete problem can be written as:

$$\begin{cases} \text{find } \mathbf{u}_h \in V_h^k,\ p_h \in Q_h^\ell \quad \text{such that :} \\ a_h(\mathbf{u}_h, \mathbf{v}_h) - b_h(p_h, \mathbf{v}_h) = 0 & \forall \mathbf{v}_h \in V_h^k, \\ b_h(q_h, \mathbf{u}_h) = (f, q_h) - \langle g, q_h \rangle & \forall q_h \in Q_h^\ell. \end{cases} \tag{10}$$

Elimination of the flux variable u

In order to eliminate the flux variable, we first recall a useful identity (see [9], [3]), that holds for vectors $\mathbf{v}$ and scalars φ piecewise smooth on $\mathcal{T}_h$:

$$\sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathbf{v}\cdot\mathbf{n}\,\varphi\,\mathrm{d}s = \sum_{e\in\mathcal{E}_h} \int_e \{\mathbf{v}\}\cdot[\![\,\varphi\,]\!]\,\mathrm{d}s + \sum_{e\in\mathcal{E}_h^\circ} \int_e [\![\,\mathbf{v}\,]\!]\{\varphi\}\,\mathrm{d}s, \tag{11}$$

where $\mathcal{E}_h = \mathcal{E}_h^\circ \cup \mathcal{E}_h^\partial$ is the set of all the edges. Using (11) we have, for all $q_h \in Q_h^\ell$ and $\mathbf{u}_h \in V_h^k$,

$$\sum_{T\in\mathcal{T}_h} \int_T (\mathrm{div}\mathbf{u}_h\, q_h + \mathbf{u}_h\cdot\nabla q_h)\,\mathrm{d}\mathbf{x} = \sum_{e\in\mathcal{E}_h} \int_e \{\mathbf{u}_h\}\cdot[\![\,q_h\,]\!]\,\mathrm{d}s + \sum_{e\in\mathcal{E}_h^\circ} \int_e [\![\,\mathbf{u}_h\,]\!]\{q_h\}\,\mathrm{d}s. \tag{12}$$

Substituting (12) in the first equation of (10) we obtain:

$$\sum_{T\in\mathcal{T}_h} \int_T (\kappa^{-1}\mathbf{u}_h + \nabla p_h)\cdot\mathbf{v}_h\,\mathrm{d}\mathbf{x} - \sum_{e\in\mathcal{E}_h^\circ} \int_e [\![\,p_h\,]\!]\cdot\{\mathbf{v}_h\}\,\mathrm{d}s = 0 \quad \forall \mathbf{v}_h \in V_h^k. \tag{13}$$

Introducing the lifting operator $R : L^1(\cup\partial T) \to V_h^k$ defined by

$$\int_\Omega R([\![\,q\,]\!])\cdot\mathbf{v}_h\,\mathrm{d}\mathbf{x} = -\sum_{e\in\mathcal{E}_h^\circ} \int_e [\![\,q\,]\!]\cdot\{\mathbf{v}_h\}\,\mathrm{d}s \quad \forall \mathbf{v}_h \in V_h^k, \tag{14}$$

we can substitute (14) in (13), thus obtaining:

$$\sum_{T\in\mathcal{T}_h} \int_T (\kappa^{-1}\mathbf{u}_h + \nabla p_h + R([\![\,p_h\,]\!]))\cdot\mathbf{v}_h\,\mathrm{d}\mathbf{x} = 0 \quad \forall \mathbf{v}_h \in V_h^k. \tag{15}$$

We also introduce the operator π_V from, say, $(L^2(\Omega))^2$ to V_h^k defined for all $\mathbf{w} \in (L^2(\Omega))^2$ as

$$(\pi_V\mathbf{w}, \mathbf{v}_h) = (\mathbf{w}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h^k. \tag{16}$$

In other words, π_V is the L^2-projection onto V_h^k. As we assumed that κ is piecewise constant, equation (15) gives now:

$$\kappa^{-1}\mathbf{u}_h = -(\pi_V\nabla_h p_h + R([\![\,p_h\,]\!])), \tag{17}$$

where ∇_h denotes the gradient element by element. It is clear that whenever

$$\nabla_h Q_h^\ell \subset V_h^k, \tag{18}$$

we have $\pi_V \nabla_h q_h \equiv \nabla_h q_h$ for all $q_h \in Q_h^\ell$, but this will not be true in general. In what follows we shall write $(\nabla_h q_h, \mathbf{v}_h)$ or $(\pi_V \nabla_h q_h, \mathbf{v}_h)$, as they are always equal. On the other hand we point out that, for instance, $(\pi_V \nabla_h p_h, \pi_V \nabla_h q_h)$ and $(\nabla_h p_h, \nabla_h q_h)$ will *not* be equal, unless (18) holds true.
Using once more (12) and the lifting operator R defined in (14) we have, for every $\mathbf{v}_h \in V_h^k$ and for every $q_h \in Q_h^\ell$, the following identity:

$$\begin{aligned} b_h(q_h, \mathbf{v}_h) &\equiv -\sum_{T\in\mathcal{T}_h} \int_T \mathbf{v}_h \cdot \nabla q_h \, \mathrm{d}\mathbf{x} + \sum_{e\in\mathcal{E}_h^\circ} \int_e [\![q_h]\!] \cdot \{\mathbf{v}_h\} \, \mathrm{d}s \\ &= -\sum_{T\in\mathcal{T}_h} \int_T \mathbf{v}_h \cdot (\nabla q_h + R([\![q_h]\!])) \, \mathrm{d}\mathbf{x} \quad \forall q_h \in Q_h^\ell. \end{aligned} \tag{19}$$

Replacing (19) with $\mathbf{v}_h = \mathbf{u}_h$ in the second equation of (10) and using (17) we finally obtain:

$$\sum_{T\in\mathcal{T}_h} \int_T \kappa(\pi_V \nabla p_h + R([\![p_h]\!])) \cdot (\pi_V \nabla q_h + R([\![q_h]\!])) \, \mathrm{d}\mathbf{x} = (f, q_h) - \langle g, q_h \rangle \quad \forall q_h \in Q_h^\ell. \tag{20}$$

When assumption (18) holds, defining

$$\mathcal{A}_{BR}(p_h, q_h) := (\kappa(\nabla_h p_h + R([\![p_h]\!])), \nabla_h q_h + R([\![q_h]\!])), \tag{21}$$

formulation (20) can be written as

$$\mathcal{A}_{BR}(p_h, q_h) = (f, q_h) - \langle g, q_h \rangle \quad \forall q_h \in Q_h^\ell, \tag{22}$$

that coincides with the Bassi-Rebay [4] formulation of problem (2), which is known to suffer from stability problems (see e.g.[9], [3]).
When assumption (18) does not hold, formulation (20) can be seen as a "generalized" Bassi-Rebay formulation, and can be written as:

$$\mathcal{A}^\pi_{BR}(p_h, q_h) = (f, q_h) - \langle g, q_h \rangle \quad \forall q_h \in Q_h^\ell, \tag{23}$$

with

$$\mathcal{A}^\pi_{BR}(p_h, q_h) := (\kappa(\pi_V \nabla_h p_h + R([\![p_h]\!])), \pi_V \nabla_h q_h + R([\![q_h]\!])), \quad p_h,\, q_h \in Q_h^\ell. \tag{24}$$

Remark. We have to underline an important point concerning the use of the operator R. We point out that, in writing the bilinear forms, the operator R, as defined in (14), should be used only when tested on elements of the space V_h^k. In order to have a simpler notation, we will actually use it everywhere. However, terms of the type

$$(\kappa \nabla_h \varphi, R([\![q_h]\!])) \equiv \int_\Omega \kappa \nabla_h \varphi \cdot R([\![q_h]\!]) \, \mathrm{d}\mathbf{x} \tag{25}$$

with $\varphi \in H^2(\mathcal{T}_h)$ and $q_h \in Q_h^\ell$, *must* be interpreted as a shorter (although somewhat imprecise) way to write

$$-\sum_{e\in\mathcal{E}_h^\circ} \int_e [\![q_h]\!] \cdot \{\kappa \nabla_h \varphi\} \, \mathrm{d}s. \tag{26}$$

Indeed, although the two expressions (25) and (26) actually coincide whenever $\nabla_h \varphi \in V_h^k$ (as can be seen from the definition (14)), they do not coincide in more general cases. In these cases, the form (26) has to be intended as the right one. ■

Stabilization of formulation (10)

The usual way to stabilize it (see e.g. [3] or [13] and the references therein) is to introduce penalty terms on the jumps of p and/or on the jumps of $\mathbf{u}$. Here instead we are going to consider a different type of stabilization (introduced for these problems in [18] for conforming finite elements, and extended in [17] to DG methods) that is more in the line of the general strategy of [16]. The basic idea of [17] is better described in terms of bilinear forms. We therefore write first (10) in the equivalent form

$$\begin{cases} \text{find } (\mathbf{u}_h,\ p_h) \in V_h^k \times Q_h^\ell \text{ such that :} \\ a_h(\mathbf{u}_h, \mathbf{v}_h) - b_h(p_h, \mathbf{v}_h) - b_h(q_h, \mathbf{u}_h) + \\ (f, q_h) - \langle g, q_h \rangle = 0 \quad \forall\, (\mathbf{v}_h, q_h) \in V_h^k \times Q_h^\ell. \end{cases} \tag{27}$$

The original stabilization of [17], following what had been done in [18] for the conforming case, amounts to change (27) into

$$\begin{cases} \text{find } (\mathbf{u}_h,\ p_h) \in V_h^k \times Q_h^\ell \text{ such that :} \\ a_h(\mathbf{u}_h, \mathbf{v}_h) - b_h(p_h, \mathbf{v}_h) - b_h(q_h, \mathbf{u}_h) - \frac{1}{2}(\kappa \nabla p_h + \mathbf{u}_h, \delta \nabla q_h + \kappa^{-1} \mathbf{v}_h) \\ (f, q_h) - \langle g, q_h \rangle = 0 \quad \forall\, (\mathbf{v}_h, q_h) \in V_h^k \times Q_h^\ell. \end{cases} \tag{28}$$

where δ could assume either the value $+1$ or the value -1 (giving rise to two different possible stabilizations). In a sense, (28) can be seen as a correction using the residual of the first equation in (10), more in the style, as we said, of [16].

Numerical experiments were performed in [17] on the formulation (28) with very good results. The convergence proof, however, was given only for a further modified form of (28), where some additional penalty in the jump terms (classical for DG methods) was added in order to enhance stability. But on the basis of the numerical experiences, the authors conjectured that the additional jump penalty was not needed.

Here indeed we show that such conjecture is perfectly correct: We consider the form (28) as it is (without any additional jump stabilization), and we allow (as a minor generalization) the coefficient in front of the residual correction to be a more general parameter θ instead of just $1/2$. We shall see that, surprisingly enough, after the elimination of the flux variable, the residual-dependent stabilization (28) turns out to be in some cases a sort of jump stabilization itself, and in other cases a stable combination of two originally unstable DG methods (namely, Bassi-Rebay [4] and Baumann-Oden [6]).

Let therefore θ be a parameter to be chosen, and let $\delta = +1$ or -1. Using the equivalent expressions (15) and (19) for the first and second equation of (10), respectively, we consider the problem:

$$\begin{cases} \text{find } \mathbf{u}_h \in V_h^k,\ p_h \in Q_h^\ell \text{ such that },\forall \mathbf{v}_h \in V_h^k,\ \forall q_h \in Q_h^\ell : \\ \displaystyle\int_\Omega (\kappa^{-1}\mathbf{u}_h + \nabla_h p_h + R(\llbracket p_h \rrbracket)) \cdot \mathbf{v}_h \, \mathrm{d}\mathbf{x} - \theta \int_\Omega (\kappa^{-1}\mathbf{u}_h + \nabla_h p_h) \cdot \mathbf{v}_h \, \mathrm{d}\mathbf{x} = 0 \\ \displaystyle -\int_\Omega \mathbf{u}_h \cdot (\nabla_h q_h + R(\llbracket q_h \rrbracket))\, \mathrm{d}\mathbf{x} + \delta\theta \int_\Omega (\mathbf{u}_h + \kappa \nabla_h p_h) \cdot \nabla_h q_h \, \mathrm{d}\mathbf{x} = (f, q_h) - \langle g, q_h \rangle, \end{cases} \tag{29}$$

that is clearly equivalent to (28) with just θ instead of 1/2. In order to check stability and boundedness, we shall eliminate as before the $\mathbf{u}$-variable and rewrite the problem in terms of the p-variable only. From the first equation in (29) and the notation (16) we then deduce:

$$\kappa^{-1}\mathbf{u}_h = -(\pi_V \nabla_h p_h + \frac{1}{1-\theta} R([\![p_h]\!])). \tag{30}$$

Consider first the case $\delta = 1$.The second equation of (29) then reads:

$$-\int_\Omega ((1-\theta)\mathbf{u}_h \cdot \nabla_h q_h + \mathbf{u}_h \cdot R([\![q_h]\!]) - \theta\kappa\nabla_h p_h \cdot \nabla_h q_h)\ \mathrm{d}\mathbf{x} = (f, q_h) - \langle g, q_h\rangle \quad \forall q_h \in Q_h^\ell.$$

In the simpler case in which assumption (18) holds (and hence $\pi_V \nabla_h q_h \equiv \nabla q_h$), substituting the expression (30) and adding and subtracting the term $(\kappa R([\![p_h]\!]), R([\![q_h]\!]))$ we obtain:

$$\mathcal{A}_{BR}(p_h, q_h) + \frac{\theta}{1-\theta} \underbrace{\int_\Omega \kappa R([\![p_h]\!]) \cdot R([\![q_h]\!])\ \mathrm{d}\mathbf{x}}_{\mathcal{S}(p_h, q_h)} = (f, q_h) - \langle g, q_h\rangle \quad \forall q_h \in Q_h^\ell. \tag{31}$$

Hence, when $\delta = 1$ and (18) holds, formulation (29) is symmetric, and coincides with the Bassi-Rebay formulation (22) with the addition of the term

$$\frac{\theta}{1-\theta}\mathcal{S}(p_h, q_h) \equiv \frac{\theta}{1-\theta}\int_\Omega \kappa R([\![p_h]\!]) \cdot R([\![q_h]\!])\ \mathrm{d}\mathbf{x}.$$

Equivalently, thanks to the definition (14), (31) can be seen as the interior penalty method (see, e.g., [2]) with the addition of the term

$$\frac{1}{1-\theta}\mathcal{S}(p_h, q_h) \equiv \frac{1}{1-\theta}\int_\Omega \kappa R([\![p_h]\!]) \cdot R([\![q_h]\!])\ \mathrm{d}\mathbf{x}.$$

The situation is formally a little more complicated when (18) does not hold. However it is not difficult to see that proceeding as before, and adding and subtracting the term $(\kappa R([\![p_h]\!]), R([\![q_h]\!]))$ we can reach the form

$$\begin{aligned} \mathcal{A}^\pi_{BR}(p_h, q_h) + \frac{\theta}{1-\theta}\mathcal{S}(p_h, q_h)\ &+\theta(\kappa(I-\pi_V)\nabla_h p_h, \nabla_h q_h) \\ &= (f, q_h) - \langle g, q_h\rangle \quad \forall q_h \in Q_h^\ell, \end{aligned} \tag{32}$$

where $\mathcal{A}^\pi_{BR}(p_h, q_h)$ is defined in (24).

Consider now the case $\delta = -1$. The second equation of (29) for $\delta = -1$ reads:

$$-\sum_{T\in\mathcal{T}_h}\int_T ((1+\theta)\mathbf{u}_h \cdot \nabla q_h + \mathbf{u}_h \cdot R([\![q_h]\!]) + \theta\kappa\nabla p_h \cdot \nabla q_h)\ \mathrm{d}\mathbf{x} = (f, q_h) - \langle g, q_h\rangle \quad \forall q_h \in Q_h^\ell.$$

As before we analyze first the case when (18) holds. Substituting the expression (30) and rearranging terms we obtain:

$$\mathcal{A}_{BR}(p_h,q_h)+\frac{2\theta}{1-\theta}\underbrace{\sum_{T\in\mathcal{T}_h}\int_T R(\llbracket p_h \rrbracket)\cdot\kappa\nabla q_h\ \mathrm{d}\mathbf{x}}_{\mathcal{A}_{NS}(p_h,q_h)}+\frac{\theta}{1-\theta}\underbrace{\int_\Omega \kappa R(\llbracket p_h \rrbracket)\cdot R(\llbracket q_h \rrbracket)\ \mathrm{d}\mathbf{x}}_{\mathcal{S}(p_h,q_h)} \\ =(f,q_h)-\langle g,q_h\rangle \quad \forall q_h\in Q_h^\ell. \tag{33}$$

We remark that formulation (33) can be rewritten as

$$\frac{1}{1-\theta}\mathcal{A}_{BR}(p_h,q_h)-\frac{\theta}{1-\theta}\mathcal{A}_{BO}(p_h,q_h)=(f,q_h)-\langle g,q_h\rangle \quad \forall q_h\in Q_h^\ell, \tag{34}$$

where $\mathcal{A}_{BO}(p_h,q_h)$ denotes the nonsymmetric bilinear form corresponding to the DG formulation of problem (2) introduced by Baumann and Oden [6], and given by:

$$\begin{aligned}&\mathcal{A}_{BO}(p_h,q_h):=\\ &\sum_{T\in\mathcal{T}_h}\int_T\kappa\nabla p_h\cdot\nabla q_h\ \mathrm{d}\mathbf{x}-\sum_{e\in\mathcal{E}_h^\circ}\int_e\{\kappa\nabla_h p_h\}\cdot\llbracket q_h\rrbracket\ \mathrm{d}s+\sum_{e\in\mathcal{E}_h^\circ}\int_e\{\kappa\nabla_h q_h\}\cdot\llbracket p_h\rrbracket\ \mathrm{d}s\\ &\equiv\sum_{T\in\mathcal{T}_h}\int_T(\kappa\nabla p_h\cdot\nabla q_h+\kappa\nabla p_h\cdot R(\llbracket q_h\rrbracket)-R(\llbracket p_h\rrbracket)\cdot\kappa\nabla q_h)\ \mathrm{d}\mathbf{x}\\ &\equiv\int_\Omega\kappa(\nabla_h p_h-R(\llbracket p_h\rrbracket))\cdot(\nabla_h q_h+R(\llbracket q_h\rrbracket))\ \mathrm{d}\mathbf{x}+\int_\Omega\kappa R(\llbracket p_h\rrbracket)\cdot R(\llbracket q_h\rrbracket)\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{35}$$

When (18), instead, does not hold, we can argue as in the case $\delta=1$ and write the resulting scheme as

$$\frac{1}{1-\theta}\mathcal{A}^\pi_{BR}(p_h,q_h)-\frac{\theta}{1-\theta}\mathcal{A}^\pi_{BO}(p_h,q_h)\ -\theta(\kappa(I-\pi_V)\nabla_h p_h,\nabla_h q_h) \\ =(f,q_h)-\langle g,q_h\rangle \quad \forall q_h\in Q_h^\ell, \tag{36}$$

having defined

$$\mathcal{A}^\pi_{BO}(p_h,q_h):=(\kappa(\pi_V\nabla_h p_h-R(\llbracket p_h\rrbracket)),\pi_V\nabla_h q_h+R(\llbracket q_h\rrbracket))+\mathcal{S}(p_h,q_h)\quad p_h,\ q_h\in Q_h^\ell. \tag{37}$$

Boundedness, Stability and Consistency

Following [3], to consider boundedness and stability of formulations (31) or (32), and (34) or (36), we let $Q(h)=Q_h^\ell+H^2(\Omega)/\mathbb{R}\subset H^2(\mathcal{T}_h)/\mathbb{R}$ and define the following norm for $q\in Q(h)$:

$$|||q|||^2=|q|^2_{1,h}+\sum_{T\in\mathcal{T}_h}h_T^2|q|^2_{2,T}+\|R(\llbracket q\rrbracket)\|^2_{0,\Omega}, \tag{38}$$

where $|q|_{1,h}$ denotes the broken H^1-seminorm. We also notice that by a local inverse inequality the norm (38) is equivalent, for $q_h\in Q_h^\ell$, to

$$|q_h|^2_{1,h}+\|R(\llbracket q_h\rrbracket)\|^2_{0,\Omega}. \tag{39}$$

Following [3], to prove that (38) is a norm in $Q(h)$ we only need to show that it is equivalent to the norm

$$|||q|||_*^2 = |q|_{1,h}^2 + \sum_{T\in\mathcal{T}_h} h_T^2 |q|_{2,T}^2 + \sum_{e\in\mathcal{E}_h^\circ} h_e^{-1} \|[\![q]\!]\|_{0,e}^2. \tag{40}$$

To this end, we can state the following results where, from now on, C will denote a generic constant independent of the mesh size h.

Lemma 1 *For $k \geq \ell \geq 1$, there exist two positive constants C_1 and C_2, depending only on the minimum angle of the decomposition and on the polynomial degree ℓ, such that:*

$$\begin{cases} \forall q \text{ double-valued and polynomial of degree } \leq \ell \text{ on each internal edge we have}: \\ C_1 \|R([\![q]\!])\|_{0,\Omega}^2 \leq \sum_{e\in\mathcal{E}_h^\circ} h_e^{-1} \|[\![q]\!]\|_{0,e}^2 \leq C_2 \|R([\![q]\!])\|_{0,\Omega}^2. \end{cases} \tag{41}$$

Proof. By the definition (14), and Cauchy-Schwarz and trace inequalities, we have:

$$\begin{aligned} \|R([\![q]\!])\|_{0,\Omega}^2 &= -\sum_{e\in\mathcal{E}_h^\circ} \int_e [\![q]\!] \cdot \{R([\![q]\!])\} \, \mathrm{d}s = -\sum_{e\in\mathcal{E}_h^\circ} \int_e h_e^{-1/2} [\![q]\!] \cdot h_e^{1/2} \{R([\![q]\!])\} \, \mathrm{d}s \\ &\leq \Big(\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1} \|[\![q]\!]\|_{0,e}^2\Big)^{1/2} \Big(\sum_{e\in\mathcal{E}_h^\circ} h_e \|\{R([\![q]\!])\}\|_{0,e}^2\Big)^{1/2} \\ &\leq C\Big(\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1} \|[\![q]\!]\|_{0,e}^2\Big)^{1/2} \Big(\sum_{T\in\mathcal{T}_h} \|R([\![q]\!])\|_{0,T}^2\Big)^{1/2}, \end{aligned} \tag{42}$$

thus proving the first inequality of (41), which actually holds for any function $q \in Q(h)$, and not only for polynomials. Next, observe that V_h^k contains the space BDM_ℓ, that is, the space of vectors with each component polynomial of degree ℓ and normal component continuous across the edges [7]. Let then $\overline{e}$ be an internal edge of $\mathcal{T}_h$, shared by elements T_1 and T_2; for every $q \in Q_h^\ell$ (that is, double-valued and polynomial of degree $\leq \ell$ on each internal edge) let $\mathbf{v} \in BDM_\ell$ be defined by:

$$\mathbf{v} \cdot \mathbf{n}_{\overline{e}} = [\![q]\!] \cdot \mathbf{n}_{\overline{e}}, \quad \mathbf{v} \cdot \mathbf{n}_e = 0 \quad \forall e \neq \overline{e}. \tag{43}$$

Clearly we have:

$$\int_{\overline{e}} [\![q]\!] \cdot \{\mathbf{v}\} \, \mathrm{d}s = \|[\![q]\!]\|_{0,\overline{e}}^2, \quad \int_e [\![q]\!] \cdot \{\mathbf{v}\} \, \mathrm{d}s = 0 \quad \forall e \neq \overline{e},$$

and

$$\|\mathbf{v}\|_{0,\Omega} \leq C h_{\overline{e}}^{1/2} \|\mathbf{v} \cdot \mathbf{n}\|_{0,\overline{e}} = C h_{\overline{e}}^{1/2} \|[\![q]\!]\|_{0,\overline{e}}.$$

Hence, from definition (14) we have:

$$\int_\Omega R([\![q]\!]) \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} \equiv \int_{T_1\cup T_2} R([\![q]\!]) \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} = -\int_{\overline{e}} [\![q]\!] \cdot \{\mathbf{v}\} \, \mathrm{d}s = -\|[\![q]\!]\|_{0,\overline{e}}^2.$$

Consequently we deduce:

$$\|[\![q]\!]\|_{0,\overline{e}}^2 = -\int_{T_1\cup T_2} R([\![q]\!]) \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} \leq C \, h_{\overline{e}}^{1/2} (\|R([\![q]\!])\|_{0,T_1} + \|R([\![q]\!])\|_{0,T_2}) \|[\![q]\!]\|_{0,\overline{e}},$$

so that

$$\frac{1}{h_{\overline{e}}}\|[\![\, q \,]\!]\|^2_{0,\overline{e}} \ \leq C(\|R([\![\, q \,]\!])\|^2_{0,T_1} + \|R([\![\, q \,]\!])\|^2_{0,T_2}).$$

Summation over all the internal edges gives then

$$\sum_{e\in\mathcal{E}_h^\circ} \frac{1}{h_e}\|[\![\, q \,]\!]\|^2_{0,e} \leq C\|R([\![\, q \,]\!])\|^2_{0,\Omega}, \tag{44}$$

which proves the second inequality of (41). ■

Lemma 2 *For $k \geq 1$ there exist two positive constants C_1 and C_2, depending only on the minimum angle of the decomposition such that:*

$$C_1\|R([\![\, q \,]\!])\|^2_{0,\Omega} \leq \sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![\, q \,]\!]\|^2_{0,e} \leq C_2(\|R([\![\, q \,]\!])\|^2_{0,\Omega} + |q|^2_{1,h}) \quad \forall q \in H^2(\mathcal{T}_h). \tag{45}$$

Proof. The first inequality follows just as in the previous Lemma. Let us see the second one. As a first step we recall a classical inequality (see e.g. [1], [2]). There exists a constant C_{agm} depending only on the minimum angle, such that: for every triangle T, for every edge e of T and for every function $\varphi \in H^1(T)$ we have

$$\|\varphi\|^2_{0,e} \leq C_{agm}(h_T^{-1}\|\varphi\|^2_{0,T} + h_T|\varphi|^2_{1,T}). \tag{46}$$

For every internal edge e, and for each of the two elements T sharing e we denote by $\overline{q}$ the $L^2(e)$-projection of q onto constants, and by q_k the $L^2(e)$-projection of q on the space of polynomials (on e) of degree $\leq k$, and we extend $\overline{q}$ and q_k in T as constants in the direction normal to e. Clearly we have

$$\|[\![\, \overline{q} \,]\!]\|_{0,e} \leq \|[\![\, q_k \,]\!]\|_{0,e} \quad \forall e. \tag{47}$$

Moreover, it is easy to check that the mapping $q \to \overline{q}$ coincides with the identity whenever q is constant. We have therefore, from (46) and usual approximation properties, that

$$h_e^{-1}\|q-\overline{q}\|^2_{0,e} \leq h_e^{-1}C_{agm}(h_T^{-1}\|q-\overline{q}\|^2_{0,T} + h_T|q-\overline{q}|^2_{1,T}) \leq C|q|^2_{1,T}, \tag{48}$$

for each of the two elements T, that easily implies

$$\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![\, q-\overline{q} \,]\!]\|^2_{0,e} \leq C|q|^2_{1,h} \quad \forall q \in H^2(\mathcal{T}_h). \tag{49}$$

We also note that, from (47) and the previous Lemma with $\ell = k$,

$$\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![\, \overline{q} \,]\!]\|^2_{0,e} \leq \sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![\, q_k \,]\!]\|^2_{0,e} \leq C_2\|R([\![\, q_k \,]\!])\|^2_{0,\Omega} \equiv C_2\|R([\![\, q \,]\!])\|^2_{0,\Omega}, \tag{50}$$

where the last equality follows immediately from the fact that $R([\![q_k]\!]) \equiv R([\![q]\!])$ that, in turn, follows from the definitions of R and q_k. Since $[\![q]\!] - [\![\overline{q}]\!]$ and $[\![\overline{q}]\!]$ are L^2-orthogonal on each edge, we easily have from (49) and (50) that

$$\sum_{e \in \mathcal{E}_h^\circ} h_e^{-1} \|[\![q]\!]\|_{0,e}^2 = \sum_{e \in \mathcal{E}_h^\circ} (h_e^{-1} \|[\![q - \overline{q}]\!]\|_{0,e}^2 + h_e^{-1} \|[\![\overline{q}]\!]\|_{0,e}^2) \leq C(|q|_{1,h}^2 + \|R([\![q]\!])\|_{0,\Omega}^2) \quad (51)$$

for all $q \in H^2(\mathcal{T}_h)$, and the proof is completed. ■

Boundedness

When (18) holds, boundedness of the bilinear forms in (31) and (34) follows directly from the boundedness of the bilinear forms $\mathcal{A}_{BR}$ and $\mathcal{A}_{BO}$, as proved in [3], thanks to the equivalence of the norms (41) and (45). When (18) does not hold, boundedness still follows from the inspection of the corresponding bilinear forms (32) and (36) using the boundedness of the projection operator π_V, the arguments of [3], and the equivalence of norms (41) and (45). We explicitely point out that in dealing with terms of the type (25) containing the operator R, we must use the form (26), as we already said. It is actually in the treatment of these terms that the part

$$\sum_{T \in \mathcal{T}_h} h_T^2 |q|_{2,T}^2$$

of the norm (38) has to be used. We refer to [3] for these types of details. Hence, to summarize the above discussion, denoting by $B_h(\cdot, \cdot)$ any of the the bilinear forms in (31), (34), (32), or (36)we have:

$$\exists\, C_b > 0 \text{ such that} \quad B_h(p, q) \leq C_b \, |||p||| \, |||q||| \quad \forall p,\, q \in Q(h). \quad (52)$$

Stability

Before proving stability of our different stabilized formulations, we anticipate a trivial algebraic lemma, that we are going to use in the sequel.

Lemma 3 *Let $\mathcal{H}$ be a Hilbert space, and λ and μ positive constants. Then, for every ξ and η in $\mathcal{H}$ we have*

$$\lambda \|\xi + \eta\|_{\mathcal{H}}^2 + \mu \|\eta\|_{\mathcal{H}}^2 \geq \frac{\lambda\mu}{2(\lambda + \mu)} (\|\xi\|_{\mathcal{H}}^2 + \|\eta\|_{\mathcal{H}}^2) \quad (53)$$

Proof. We first remark that a simple algebraic computation shows that

$$\lambda \|\xi + \eta\|_{\mathcal{H}}^2 + \mu \|\eta\|_{\mathcal{H}}^2 = \|(\lambda + \mu)^{1/2} \eta + \frac{\lambda \xi}{(\lambda + \mu)^{1/2}}\|_{\mathcal{H}}^2 + \frac{\lambda\mu}{\lambda + \mu} \|\xi\|_{\mathcal{H}}^2. \quad (54)$$

Setting $L := \lambda \|\xi + \eta\|_{\mathcal{H}}^2 + \mu \|\eta\|_{\mathcal{H}}^2$ we have obviously $L \geq \mu \|\eta\|_{\mathcal{H}}^2$ and from (54) we also have $L \geq \dfrac{\lambda\mu}{\lambda + \mu} \|\xi\|_{\mathcal{H}}^2$. Hence taking the sum of the two, and using the obvious fact that $1 \geq \lambda/(\lambda + \mu)$, we get

$$2L \geq \mu \|\eta\|_{\mathcal{H}}^2 + \frac{\lambda\mu}{\lambda + \mu} \|\xi\|_{\mathcal{H}}^2 \geq \frac{\lambda\mu}{\lambda + \mu} (\|\xi\|_{\mathcal{H}}^2 + \|\eta\|_{\mathcal{H}}^2), \quad (55)$$

and the proof is completed. ■

Remark. We note that the constant in (53) is not optimal. With a little eigenvalue analysis it is not difficult to see that the best constant is $\lambda + (\mu/2) - \sqrt{\lambda^2 + (\mu/2)^2}$. However we are not going to need it, and we stick to (53) that is simpler and sufficient for our purposes. ■

Lemma 4 *For $\delta = 1$, $k \geq 1$, and $\ell \geq 1$ problem (29) is stable for all $\theta \in]0, 1[$.*

Proof. We consider first the case when (18) holds. In this case, taking $p_h = q_h$ in the left-hand side of (31), we have:

$$B_h(q_h, q_h) = \|\kappa^{1/2}(\nabla_h q_h + R([\![q_h]\!]))\|^2_{0,\Omega} + \frac{\theta}{1-\theta}\|\kappa^{1/2} R([\![q_h]\!])\|^2_{0,\Omega} \tag{56}$$

and the stability in the norm (39) follows from Lemma 3.

We consider now the case in which (18) does not hold. Considering the formulation (32), we take again $p_h = q_h$ and we have

$$\begin{aligned} B_h(q_h, q_h) = \|\kappa^{1/2}(\pi_V \nabla_h q_h + R([\![q_h]\!]))\|^2_{0,\Omega} - \theta\|\kappa^{1/2}\pi_V \nabla_h q_h\|^2_{0,\Omega} \\ +\theta\|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega} + \frac{\theta}{1-\theta}\|\kappa^{1/2} R([\![q_h]\!])\|^2_{0,\Omega}. \end{aligned} \tag{57}$$

We observe that the sum of the second and third term is always nonnegative. Applying Lemma (3) with $\lambda = 1,\ \mu = \theta/1 - \theta$ we have then

$$2B_h(q_h, q_h) \geq \theta(\|\kappa^{1/2}\pi_V \nabla_h q_h\|^2_{0,\Omega} + \|\kappa^{1/2} R([\![q_h]\!])\|^2_{0,\Omega}). \tag{58}$$

On the other hand, from (57) we also deduce

$$B_h(q_h, q_h) \geq -\theta\|\kappa^{1/2}\pi_V \nabla_h q_h\|^2_{0,\Omega} + \theta\|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega}. \tag{59}$$

Hence,

$$3B_h(q_h, q_h) \geq \theta(\|\kappa^{1/2} R([\![q_h]\!])\|^2_{0,\Omega} + \|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega}), \tag{60}$$

and the result follows. ■

Lemma 5 *For $\delta = -1$, $k \geq 1$, and $\ell \geq 1$ problem (29) is stable for all $\theta < 0$.*

Proof. Once more we start from the case when (18) holds. Taking $p_h = q_h$ in (34) and using the definitions of $\mathcal{A}_{BR}$ and $\mathcal{A}_{BO}$ as given in (21) and (35) respectively, we have:

$$B_h(q_h, q_h) = \frac{1}{1-\theta}\|\kappa^{1/2}(\nabla_h q_h + R([\![q_h]\!]))\|^2_{0,\Omega} - \frac{\theta}{1-\theta}\|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega}$$

and since $\theta < 0$ the result follows again from Lemma 3 .

We consider now the case when (18) does not hold. We take the formulation (36) with $p_h = q_h$, and for the sake of simplicity we set $\alpha = -\theta$. We have

$$\begin{aligned} B_h(q_h, q_h) = \frac{1}{1+\alpha}\|\kappa^{1/2}(\pi_V \nabla_h q_h + R([\![\, q_h \,]\!]))\|^2_{0,\Omega} \\ -\frac{\alpha^2}{1+\alpha}\|\kappa^{1/2}\pi_V \nabla_h q_h\|^2_{0,\Omega} + \alpha\|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega}. \end{aligned} \tag{61}$$

Splitting α as $\alpha = \dfrac{\alpha}{\alpha+1} + \dfrac{\alpha^2}{\alpha+1}$ and using $\|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega} \geq \|\kappa^{1/2}\pi_V \nabla_h q_h\|^2_{0,\Omega}$ and Lemma 3 we deduce

$$\begin{aligned} B_h(q_h, q_h) &\geq \frac{1}{1+\alpha}\|\kappa^{1/2}(\pi_V \nabla_h q_h + R([\![\, q_h \,]\!]))\|^2_{0,\Omega} + \frac{\alpha}{1+\alpha}\|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega} \\ &\geq C(\|\kappa^{1/2}(\pi_V \nabla_h q_h\|^2_{0,\Omega} + \|R([\![\, q_h \,]\!]))\|^2_{0,\Omega}) \end{aligned} \tag{62}$$

with $C = \alpha/2(1+\alpha)^2$. On the other hand, (62) also implies

$$B_h(q_h, q_h) \geq \frac{\alpha}{1+\alpha}\|\kappa^{1/2}\nabla_h q_h\|^2_{0,\Omega}, \tag{63}$$

and the result follows immediately. ■

To summarize, for all the bilinear forms in (31), (32), (34), or (36) we have:

$$\exists\, C_s > 0 \text{ such that} \quad B_h(q_h, q_h) \geq C_s\, |||q_h|||^2 \qquad \forall q_h \in Q_h^\ell, \tag{64}$$

where (64) clearly holds for every $\theta \in]0,1[$ for the symmetric cases ((31), (32)), and for every $\theta < 0$ for the nonsymmetric cases ((34), (36)).

Consistency

The consistency of both the original mixed formulation and of all the stabilized ones is obvious. This is not the case when we deal with the reduced formulations that we obtain after the elimination of $\mathbf{u}_h$. Indeed, as we shall see, consistency does not hold in every case.

We start by considering all the bilinear forms $B_h(p_h, q_h)$ obtained in (31), (34), (32) and (36), and we substitute the exact solution p in place of the approximate solution p_h, taking into account the fact that p is continuous and hence its jumps are zero.

We note that when (18) holds, that is in the cases (31) and (34), $B_h(p, q_h)$ can be written in the form

$$B_h(p, q_h) = (\kappa\nabla p, \nabla_h q_h + R([\![\, q_h \,]\!])). \tag{65}$$

At this point we remember that we must use the form (26) in dealing with the term containing $R([\![\, q_h \,]\!])$, so that (65) must actually be written as

$$(\kappa\nabla p, \nabla_h q_h + R([\![\, q_h \,]\!])) \to \sum_{T\in\mathcal{T}_h}\int_T \kappa\nabla p\cdot\nabla q_h \,\mathrm{d}\mathbf{x} - \sum_{e\in\mathcal{E}_h^\circ}\int_e [\![\, q_h \,]\!]\{\kappa\nabla p\}\,\mathrm{d}s. \tag{66}$$

After integration by parts and using formula (11) we obtain

$$\begin{aligned}
&\sum_{T\in\mathcal{T}_h}\int_T \kappa\nabla p\cdot\nabla q_h\,\mathrm{d}\mathbf{x} - \sum_{e\in\mathcal{E}_h^\circ}\int_e [\![\, q_h \,]\!]\{\kappa\nabla p\}\,\mathrm{d}s \\
&= -\sum_{T\in\mathcal{T}_h}\int_T \mathrm{div}(\kappa\nabla p)\, q_h\,\mathrm{d}\mathbf{x} + \sum_{T\in\mathcal{T}_h}\int_{\partial T} q_h\kappa\nabla p\cdot\mathbf{n}_T\,\mathrm{d}s - \sum_{e\in\mathcal{E}_h^\circ}\int_e [\![\, q_h \,]\!]\{\kappa\nabla p\}\,\mathrm{d}s \\
&= (f, q_h) - \langle g, q_h\rangle,
\end{aligned} \tag{67}$$

proving the consistency of both formulations (31) and (34).

The situation is different when we deal with the cases where (18) does not hold, that is the formulations (32) and (36). In these cases it is not difficult to see that both cases share the general form

$$B_h(p, q_h) = (\kappa\pi_V\nabla p, \nabla_h q_h + R([\![\, q_h \,]\!])) + \lambda(\kappa(I-\pi_V)\nabla p, \nabla_h q_h), \tag{68}$$

where λ depends on the method ($\lambda=\theta$ for (32), $\lambda=-\theta$ for (36)), and the term including R should be interpreted as in (66). By adding and subtracting $(\kappa\nabla p, \nabla_h q_h + R([\![\, q_h \,]\!]))$ in (68) we have:

$$\begin{aligned}
B_h(p, q_h) = &(\kappa(\pi_V - I)\nabla p, \nabla_h q_h + R([\![\, q_h \,]\!])) + \lambda(\kappa(I-\pi_V)\nabla p, \nabla_h q_h) \\
&+(\kappa\nabla p, \nabla_h q_h + R([\![\, q_h \,]\!])).
\end{aligned} \tag{69}$$

For the last term we can use (67), and for the others we use boundedness, to reach

$$\begin{aligned}
B_h(p, q_h) - (f, q_h) + \langle g, q_h\rangle &= (1+\lambda)(\kappa(\pi_V - I)\nabla p, \nabla_h q_h) + (\kappa(\pi_V - I)\nabla p, R([\![\, q_h \,]\!])) \\
&\le C_c\|\pi_V\nabla p - \nabla p\|_{0,\Omega}\,|||q_h|||.
\end{aligned} \tag{70}$$

Error estimates

To perform the analysis, we need to bound the approximation error $|||p - p_I|||$ when $p_I \in Q_h^\ell$ is a suitable interpolant of the exact solution p. If p_I is chosen to be the usual *continuous* interpolant of p, then $R([\![\, p - p_I \,]\!]) \equiv 0$, so that (38) immediately gives

$$|||p - p_I|||^2 = |p - p_I|_{1,h}^2 + \sum_{T\in\mathcal{T}_h} h_T^2 |p - p_I|_{2,T}^2 \le C_a^2\, h^{2\ell}|p|_{\ell+1,\Omega}^2. \tag{71}$$

If p_I is discontinuous, we just need the local approximation property:

$$|p - p_I|_{s,T} \le C\, h_T^{\ell+1-s}|p|_{\ell+1,T} \quad \forall T\in\mathcal{T}_h, \quad s = 0, 1, 2. \tag{72}$$

Then, using (41) we have:

$$|||p - p_I|||^2 \le |p - p_I|_{1,h}^2 + \sum_{T\in\mathcal{T}_h} h_T^2 |p - p_I|_{2,T}^2 + C_1^{-1}\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![\, p - p_I \,]\!]\|_{0,e}^2. \tag{73}$$

Using (46) in (73) we have:

$$|||p - p_I|||^2 \leq C\,(|p - p_I|_{1,h}^2 + \sum_{T\in\mathcal{T}_h} h_T^2 |p - p_I|_{2,T}^2 + \sum_{T\in\mathcal{T}_h} h_T^{-2}\|p - p_I\|_{0,T}^2), \tag{74}$$

and from (72) and (74) we have again

$$|||p - p_I||| \leq C_a\, h^\ell |p|_{\ell+1,\Omega}. \tag{75}$$

For the cases when (18) holds (that is when $k \geq \ell - 1$, as in formulations 31), (34)), we can use stability (64), consistency (67), boundedness (52), and the approximation property (75) to obtain:

$$\begin{aligned} C_s|||p_I - p_h|||^2 &\leq B_h(p_I - p_h, p_I - p_h) = B_h(p_I - p, p_I - p_h) \\ &\leq C_b|||p - p_I|||\;|||p_I - p_h||| \leq Ch^\ell |p|_{\ell+1,\Omega}|||p_I - p_h|||. \end{aligned} \tag{76}$$

Hence, the triangle inequality gives the optimal estimate

$$|||p - p_h||| \leq C\, h^\ell |p|_{\ell+1,\Omega}.$$

On the other hand, when dealing with the case when (18) does not hold, that is with formulations (32) and (36) (corresponding to $k < \ell - 1$), we cannot use the consistency property (67) but we have to use the weaker form (70). We have then

$$\begin{aligned} C_s|||p_I - p_h|||^2 &\leq B_h(p_I - p_h, p_I - p_h) = B_h(p_I - p, p_I - p_h) + B_h(p - p_h, p_I - p_h) \\ &\leq C_b|||p - p_I|||\;|||p_I - p_h||| + C_c\|\nabla p - \pi_V \nabla p\|_{0,\Omega}\,|||p_I - p_h||| \\ &\leq Ch^{k+1}|p|_{k+2,\Omega}|||p_I - p_h|||, \end{aligned} \tag{77}$$

where, in the last line, we took into account the fact that $k < \ell - 1$. Hence, in this case the triangle inequality gives the estimate

$$|||p - p_h||| \leq C\, h^{k+1}|p|_{k+2,\Omega}.$$

We can summarize the results of both cases in the estimate

$$|||p - p_h||| \leq C\, h^s |p|_{s+1,\Omega} \quad \text{with } s := \min\{k+1, \ell\}. \tag{78}$$

We finally point out that, in both cases, it is immediate to derive an estimate for $\mathbf{u}_h$ in terms of $\nabla_h p_h$ and $R(\llbracket p_h \rrbracket)$. Indeed, recalling (30):

$$\mathbf{u}_h = -\kappa(\pi_V \nabla_h p_h + \frac{1}{1-\theta} R(\llbracket p_h \rrbracket)) \tag{79}$$

we easily deduce, using (78),

$$\|\mathbf{u}_h - \mathbf{u}\|_{0,\Omega} \leq C\,|||p_h - p||| \leq C\, h^s |p|_{s+1,\Omega} \quad \text{with } s := \min\{k+1, \ell\}.$$

Remark. For the symmetric schemes (31) and (32) the adjoint consistency property (see [3]) holds true. Consequently, whenever the problem (2) has H^2-regularity, we also have the optimal estimate in L^2:

$$\|p - p_h\|_{0,\Omega} \leq C\, h^{s+1}|p|_{s+1,\Omega} \quad \text{with } s := \min\{k+1, \ell\}. \tag{80}$$

■

4 Numerical results

We consider discontinuous velocity and pressure triangles and quadrilaterals. The domain $\Omega = [0,1] \times [0,1]$ and the exact pressure field is $p = (\sin 2\pi x)(\sin 2\pi y)$. The velocity is computed from the pressure using Darcy's law with $\kappa = 1$, f is computed from the divergence of the velocity, and g is taken to be its normal component on the boundary. The boundary value problem is specified by setting f over Ω and g weakly over the boundary. The parameter θ is set to $1/2$ and δ is chosen as $+1$. This is the case studied previously in [18] and [17]. We consider linear and quadratic interpolations on triangles, and bilinear and biquadratic interpolations on quadrilaterals, in all combinations. The quadrilateral meshes are uniform and the triangular meshes are the same with each quadrilateral split into two triangles with all diagonals running in the same direction.

Convergence results for the equal-order elements are presented in Figures 1-4. Note that the bilinear quadrilaterals and linear triangles attain optimal second -order convergence in L^2 for the pressure field. The pressure results follow from adjoint consistency. The velocity convergence is one order less, which is consistent with the theoretical predictions. We also obtain optimal third-order L^2 pressure convergence for the quadratic triangle and biquadratic quadrilateral, but the L^2 velocity results are again one order lower, consistent with our theoretical predictions.

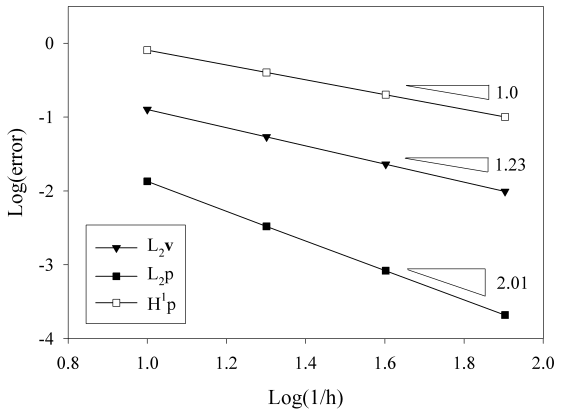

FIGURE 1
Equal-order bilinear quadrilaterals

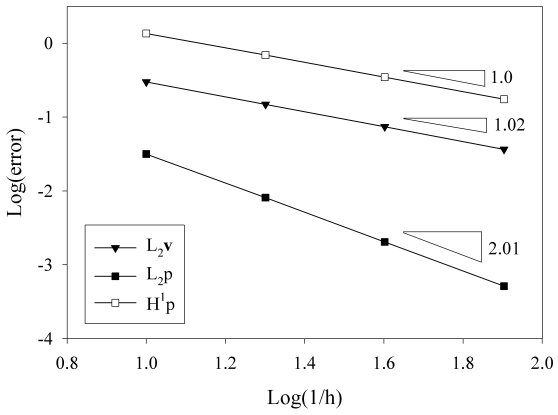

FIGURE 2
Equal-order linear triangles

Results for unequal-order elements are presented in Figures 5-8. For cases in which the pressure interpolation is the higher-order field, we see optimal L^2 rates of convergence for velocity and pressure, second- and third-order, respectively. (See Figures 5 and 6.) For the cases in which the velocity interpolation is the higher-order field, we see optimal second-order L^2 convergence for the pressure, while the order of convergence for velocity is one, as predicted by theory. (See Figures 7 and 8.)

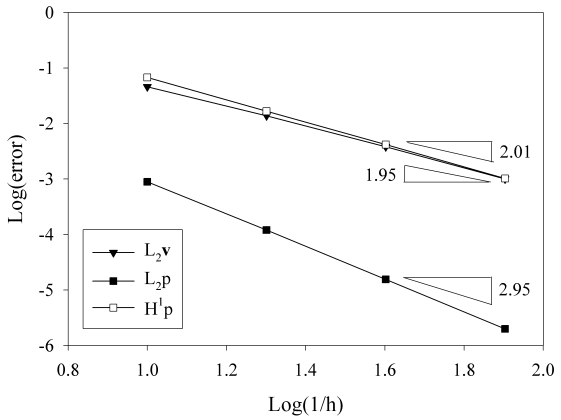

FIGURE 3
Equal-order biquadratic quadrilaterals

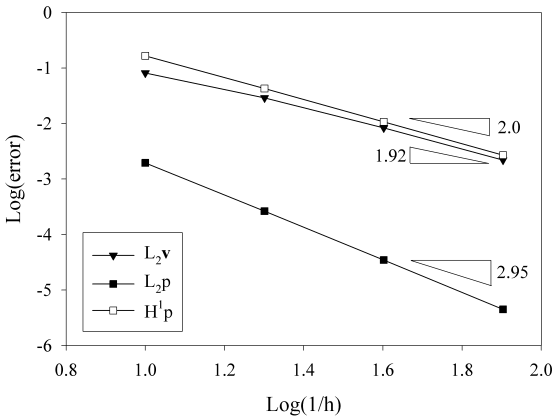

FIGURE 4
Equal-order quadratic triangles

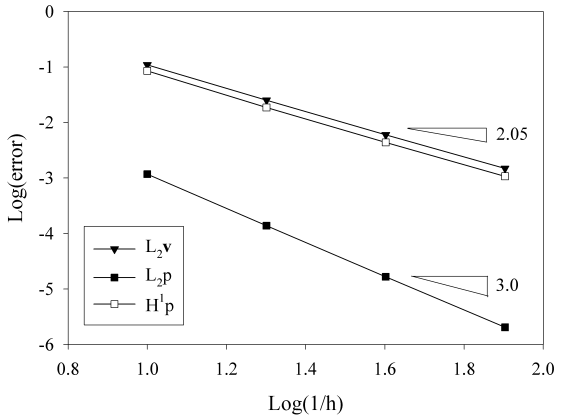

FIGURE 5
Bilinear-velocity biquadratic-pressure quadrilaterals

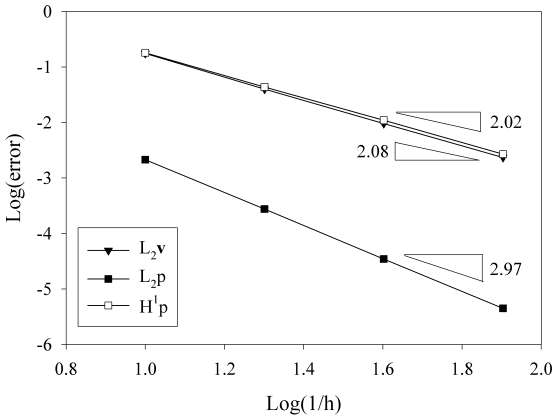

FIGURE 6
Linear-velocity quadratic-pressure triangles

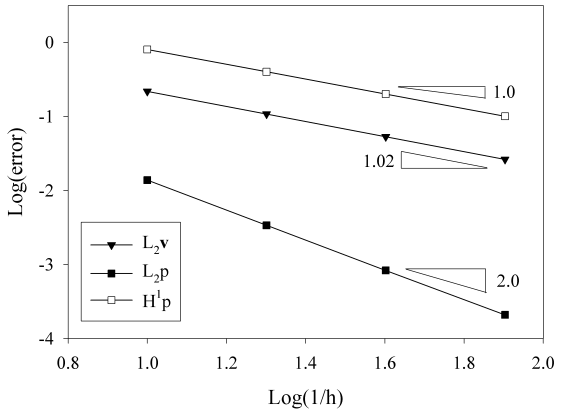

FIGURE 7
Biquadratic-velocity bilinear-pressure quadrilaterals

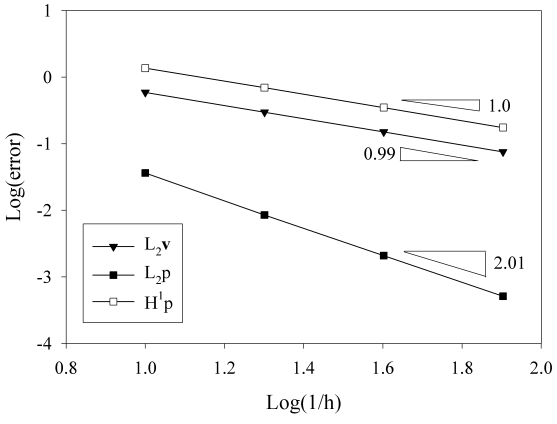

FIGURE 8
Quadratic-velocity linear pressure triangles

References

[1] S. Agmon, *Lectures on elliptic boundary value problems*, Van Nostrand Mathematical Studies, Princeton, NJ, 1965.

[2] D. N. Arnold, *An interior penalty finite element method with discontinuous element*, SIAM J. Numer. Anal. **19** (1982), 742–760.

[3] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini, *Unified Analysis of Discontinuous Galerkin Methods for Elliptic Problems*, SIAM J. Numer. Anal. **39** (2002), 1749–1779.

[4] F. Bassi and S. Rebay, *Discontinuous finite element high order accurate numerical solution of the compressible Navier-Stokes equations*, in Proceedings of the Conference "Numerical methods for fluid dynamics V", K. W. Morton et al., ed., Oxford: Clarendon Press, April 3–6 1995, 295–302.

[5] P. Bastian and B. Rivière, *Superconvergence and H(div) Projection for Discontinuous Galerkin Methods*, Int. J. Numer. Meth. Fluids **42** (2003), 1043–1057.

[6] C.E. Baumann and J.T. Oden, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg. **175** (1999), 311–341.

[7] F. Brezzi, J. Douglas jr., and L.D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math. **47** (1985), 217–235.

[8] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo, *Discontinuous finite elements for diffusion problems*, Atti Convegno in onore di F. Brioschi (Milano 1997), Istituto Lombardo, Accademia di Scienze e Lettere, 1999, 197–217.

[9] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo, *Discontinuous Galerkin approximations for elliptic problems*, Numerical Methods for Partial Differential Equations **16** (2000), 365–378.

[10] P.G. Ciarlet *The finite element methods for elliptic problems*, North Holland, 1978.

[11] B. Cockburn and C. Dawson, *Some extension of the local discontinuous Galerkin method for convection- diffusion equations in multidimensions*, Proc. MAFELAP X, J.R. Whiteman Ed., (2000), 225–238.

[12] B. Cockburn and C. Dawson, *Approximation of the velocity by coupling discontinuous Galerkin and mixed finite element methods for flow problems*, Computational Geosciences **6** (2002), 502–522.

[13] B. Cockburn, G. E. Karniadakis, and C.-W. Shu eds., *Discontinuous Galerkin Methods: Theory, Computation and Applications*, Lecture Notes in Computational Science and Engineering **11**, Springer-Verlag, 2000.

[14] M. Dauge, *Elliptic boundary value problems on corner domains. Smoothness and asymptotics of solutions,* Lecture Notes in Mathematics **1341**, Springer-Verlag, Berlin, 1988.

[15] J. Douglas, Jr. and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, Lecture Notes in Physics, vol. 58, Springer-Verlag, Berlin, 1976.

[16] T.J.R. Hughes, L.P. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics. V. Circumventing the Babuška-Brezzi condition: a stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Methods Appl. Mech. Engrg. **59** (1986), 85–99.

[17] T.J.R. Hughes, A. Masud, and J. Wan, *A stabilized mixed discontinuous Galerkin method for Darcy flow* (in preparation).

[18] A. Masud and T.J.R. Hughes, *A stabilized mixed finite element method for Darcy flow*, Comput. Methods Appl. Mech. Engrg. **191** (2002), 4341–4370.

[19] P.A. Raviart and J.M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method (I. Galligani, E. Magenes, eds.), Lecture Notes in Math., Springer-Verlag, New York, **606** (1977), 292–315.

[20] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal. **15** (1978), 152–161.