

## PIECEWISE CONSTANT PRESSURE FOR DARCY LAW

Franco Brezzi\*<sup>1</sup>, Michel Fortin<sup>3</sup>, and L. Donatella Marini<sup>2</sup>

<sup>1</sup> *Dipartimento di Matematica - Università di Pavia  
and IMATI-CNR  
Via Ferrata 1, I-27100 Pavia, Italy  
Email: brezzi@imati.cnr.it  
Web page: <http://www.imati.cnr.it/~brezzi>*

<sup>2</sup> *Dipartimento di Matematica - Università di Pavia  
and IMATI-CNR  
Via Ferrata 1, I-27100 Pavia, Italy  
Email: marini@imati.cnr.it  
Web page: <http://www.imati.cnr.it/~marini>*

<sup>3</sup> *GIREF  
Université Laval  
Québec  
Canada, G1K 7P4  
Email: mfortin@giref.ulaval.ca*

**Abstract.** In this paper we present a short discussion on some finite element formulations for linear elliptic problems. For the sake of simplicity we consider the Poisson equation  $-\Delta p = f$ , taking the notation from Darcy's law. Among the zillions of such methods, we shall concentrate our attention on FEM leading to a final system of linear algebraic equations  $MP = F$  where each unknown  $P_i$  represents the constant value of the approximated pressure  $p_h$  in a single element. It is indeed well known that for some applications there is a certain demand for these types of schemes.

**Key words:** Darcy, Finite Elements, Mixed formulations, Constant pressures, Quadrature formula's, Finite Volumes

## 1 INTRODUCTION

We consider, for simplicity, the model toy-problem

$$-\Delta p = f \quad \text{in } \Omega, \quad (1)$$

$$p = 0 \quad \text{on } \Gamma \equiv \partial\Omega, \quad (2)$$

where  $\Omega$  is a polygon in  $\mathbb{R}^d$ ,  $d = 2$  or  $3$  being the number of dimensions, and  $f$  is a given function, say, in  $L^2(\Omega)$ .

The number of different Finite Element Methods used to deal with this problem, although finite, is practically uncountable. Nevertheless, there are much fewer methods that allow to deal with a piecewise constant approximation of the pressure  $p$ . Most of them are related to the so called *mixed formulation* (see e.g.<sup>4,6</sup>) of (1)-(2) where the velocity  $\mathbf{u} = -\nabla p$  is introduced as an additional unknown. This is for instance the case of the lowest order Raviart-Thomas element ( $RT$ )<sup>13</sup> or the lowest order Brezzi-Douglas-Marini element ( $BDM$ ),<sup>5</sup> that are both based on the mixed variational formulation

$$\int_{\Omega} \mathbf{u} \cdot \mathbf{v} \, d\mathbf{x} = \int_{\Omega} p \operatorname{div} \mathbf{v} \, d\mathbf{x} \quad \forall \mathbf{v}, \quad (3)$$

$$\int_{\Omega} q \operatorname{div} \mathbf{u} \, d\mathbf{x} = \int_{\Omega} f q \, d\mathbf{x} \quad \forall q, \quad (4)$$

where the spaces for the variations of  $\mathbf{v}$  and  $q$  have to be made precise. In particular,  $q$  varies in  $L^2(\Omega)$  and  $\mathbf{v}$  varies over the space  $H(\operatorname{div}; \Omega)$  defined as

$$H(\operatorname{div}; \Omega) := \{\mathbf{v} \in (L^2(\Omega))^d \text{ such that } \operatorname{div} \mathbf{v} \in L^2(\Omega)\}. \quad (5)$$

This last condition, for piecewise smooth vectors, implies the continuity of  $\mathbf{v} \cdot \mathbf{n}$  at interfaces. This makes the elimination of the velocity unknown  $\mathbf{u}$  rather difficult. In this sense, these methods, unless we do some further manipulation, do not give rise to a final system of algebraic equations of the type

$$M P = F \quad (6)$$

where  $P$  represents the values of the discretized pressure inside each element (which is our request, and, here, “the name of the game”).

In what follows we shall overview a few tricks for deriving a final system of the type (6). We shall briefly start from methods that allow to eliminate  $\mathbf{u}$  from classical discretized versions of the mixed formulation (3)-(4), and then consider some other less classical formulations that still allow the elimination of  $\mathbf{u}$ . In doing that, in an almost inevitable way, we will often get close to (or rediscover) some of the Finite Volume formulations of the original problem (1)-(2). However, a review of the (zillions of) different FV approximations of it is well beyond the scopes of this paper, as well as beyond the competence of the authors.

Acknowledgments: This paper has been inspired by the continuous pestering that T.J.R. Hughes operated during his last visit to Pavia, concerning Finite Element formulations of Darcy’s laws using a piecewise constant pressure. If the paper was not to be dedicated to him, we would have contacted him in order to write a joint paper. That would have been fair from us, and we beg his pardon for not having

done so. However, the estimate of the amount of gain or loss that Tom had by *not being* asked to work on this paper is unclear to us, and is left to him and to the reader. But in any case please, Tom, keep pestering us! Your pestering is a never ending source of inspiration.

## 2 CLASSICAL MIXED FORMULATIONS

We start by recalling the two main choices for mixed discretizations of lowest degree. Assume therefore that we are given a regular sequence  $\{\mathcal{T}_h\}_h$  of decompositions of  $\Omega$  into triangles (in two dimensions) or tetrahedra (in three dimensions). Let  $K$  be an element in  $\mathcal{T}_h$ . We define the local Raviart-Thomas and Brezzi-Douglas-Marini spaces as follows:

$$RT(K) := \{\mathbf{v} \mid \mathbf{v} = \mathbf{c} + \gamma \mathbf{x}, \text{ with } \mathbf{c} \in \mathbb{R}^d \text{ and } \gamma \in \mathbb{R}\}, \quad (7)$$

and

$$BDM(K) := \{\mathbf{v} \mid \mathbf{v} \in (P_1)^d\}, \quad (8)$$

where  $P_1$  is as usual the space of polynomials of degree  $\leq 1$ . Starting from the local definitions (7) and (8) we can now define the global subspaces of  $H(\text{div}; \Omega)$

$$V_{RT} = \{\mathbf{v} \mid \mathbf{v} \in H(\text{div}; \Omega), \mathbf{v}|_K \in RT(K) \ \forall K \in \mathcal{T}_h\}, \quad (9)$$

and

$$V_{BDM} = \{\mathbf{v} \mid \mathbf{v} \in H(\text{div}; \Omega), \mathbf{v}|_K \in BDM(K) \ \forall K \in \mathcal{T}_h\}. \quad (10)$$

As we already pointed out, the condition  $\mathbf{v} \in H(\text{div}; \Omega)$  implies that the normal component of  $\mathbf{v}$  must be continuous at the interfaces (edges or faces, according to whether  $d = 2$  or  $3$ ) between one element and the other. For both choices the space for approximating the pressure is the space  $Q_h$  of piecewise constants. The discretized problem for Raviart-Thomas mixed approximation is then

find  $\mathbf{u}_h \in V_{RT}$  and  $p_h \in Q_h$  such that:

$$\int_{\Omega} \mathbf{u}_h \cdot \mathbf{v}_h \, d\mathbf{x} = \int_{\Omega} p_h \, \text{div} \, \mathbf{v}_h \, d\mathbf{x} \quad \forall \mathbf{v}_h \in V_{RT}, \quad (11)$$

$$\int_{\Omega} q_h \, \text{div} \, \mathbf{u}_h \, d\mathbf{x} = \int_{\Omega} f \, q_h \, d\mathbf{x} \quad \forall q_h \in Q_h, \quad (12)$$

while its *BDM* counterpart reads

find  $\mathbf{u}_h \in V_{BDM}$  and  $p_h \in Q_h$  such that:

$$\int_{\Omega} \mathbf{u}_h \cdot \mathbf{v}_h \, d\mathbf{x} = \int_{\Omega} p_h \, \text{div} \, \mathbf{v}_h \, d\mathbf{x} \quad \forall \mathbf{v}_h \in V_{BDM}, \quad (13)$$

$$\int_{\Omega} q_h \, \text{div} \, \mathbf{u}_h \, d\mathbf{x} = \int_{\Omega} f \, q_h \, d\mathbf{x} \quad \forall q_h \in Q_h. \quad (14)$$

It will be convenient to take suitable notation in order to write mixed formulations in a simpler way. For this we introduce, for each  $K \in \mathcal{T}_h$ , the bilinear forms

$$a_K(\mathbf{u}, \mathbf{v}) := \int_K \mathbf{u} \cdot \mathbf{v} \, d\mathbf{x}, \quad (15)$$

and, for  $q|_K = \text{constant}$ ,

$$b_K(q, \mathbf{v}) := \int_K q \operatorname{div} \mathbf{v} \, d\mathbf{x} \equiv \int_{\partial K} q \mathbf{v} \cdot \mathbf{n}_K \, ds, \quad (16)$$

where  $\mathbf{n}_K$  is the outward unit normal to  $\partial K$ . We can also define the bilinear forms on  $\Omega$ . We set

$$a(\mathbf{u}, \mathbf{v}) := \sum_{K \in \mathcal{T}_h} a_K(\mathbf{u}, \mathbf{v}) \quad b(q, \mathbf{v}) := \sum_{K \in \mathcal{T}_h} b_K(q, \mathbf{v}) \quad (f, q) := \int_{\Omega} f q \, d\mathbf{x}. \quad (17)$$

Both formulations (11)-(12) and (13)-(14) can then be written as

find  $\mathbf{u}_h \in V_h$  and  $p_h \in Q_h$  such that:

$$a(\mathbf{u}_h, \mathbf{v}_h) = b(p_h, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_h, \quad (18)$$

$$b(q_h, \mathbf{u}_h) = (f, q_h) \quad \forall q_h \in Q_h, \quad (19)$$

and we choose one method or another by selecting  $V_h = V_{RT}$  or  $V_h = V_{BDM}$ . Problems of the form (18)-(19) give rise, with obvious notation, to linear algebraic systems of the form

$$AU - B^t P = 0 \quad (20)$$

$$BU = F \quad (21)$$

In these cases the main problem, from the computational point of view, is the elimination of the velocity unknowns  $U$ . If we succeed in doing that, we can obtain the double gain of reducing the number of degrees of freedom, and to pass from an indefinite system to one with a symmetric and positive definite matrix. Clearly one could always write

$$B A^{-1} B^t P = F, \quad (22)$$

but  $A^{-1}$  will not be computable (from the practical point of view) in an explicit way. In the use of iterative methods such as Conjugate Gradient (CG) this could be an only minor drawback: the matrix  $A$  is an approximation of the identity, and a few ‘‘inner iterations’’ could be enough to solve a system of the form  $AV = G$  for each CG step. However, this complicates the solver, and one is never sure that the number of inner iterations is sufficient, so that most researchers avoid doing that.

A similar remark can be done for *discontinuous mixed formulations* (see e.g.<sup>12</sup> or<sup>8</sup>) where the presence of interface consistency terms and/or stabilizing jump terms renders an easy local elimination of the velocity field impossible.

We are therefore looking for methods that allow a more *explicit* elimination of the velocity unknowns.

A classical way of doing this is the use of the so-called *hybridization*. The idea goes back to Fraeijns de Veubeke,<sup>10</sup> and consists in starting with a discontinuous version of  $V_{RT}$  (or  $V_{BDM}$ ), and to force back the continuity of the normal component via a suitable Lagrange multiplier, that we assume to be constant for  $RT$  and linear for  $BDM$  on each edge (resp. face for  $d = 3$ ). Being now, a priori, totally discontinuous, the velocity space can then be eliminated at the element level, leading to a system where the only unknowns left are the pressure  $p_h$  and the Lagrange multipliers at the interfaces (that turn out to be themselves approximations of the pressure). The

original unknown  $p_h$  can also be eliminated at the element level, leaving a final system involving the Lagrange multipliers only. The resulting matrix is symmetric and positive definite. The method, introduced in<sup>10</sup> and then analyzed in<sup>2</sup>, is quite efficient, but does not have the required form (6), that would require  $P$  to represent the values of  $p_h$  in each element.

Another possibility to reach a symmetric and positive definite matrix is to use some algebraic trick (rather, coming from Graph Theory and Operational Research) in order to look directly for a  $\mathbf{u}_h$  written as the sum of a particular solution of (12) plus all possible velocity fields in  $V_{RT}$  having zero divergence. The use of classical *tree-cotree* algorithms for doing that (for  $RT$  elements) was advocated in<sup>1</sup>, and reduces the number of unknowns to the number of edges (resp. faces) *minus* the number of elements (which is precisely the dimension of the free-divergence subspace of  $V_{RT}$ ). The resulting matrix is symmetric and positive definite. This is also an interesting trick, but it does not fit the required form (6) either.

In the next sections we shall see some tricks that allow us to reach exactly the form (6), plus some additional formulation that makes the elimination procedure easier. In doing so we shall discuss the two-dimensional and the three-dimensional cases separately. This is because the two-dimensional case is obviously easier, and obviously provides already some insight into the three-dimensional one, but the similarities will not be sufficient to treat them both at the same time in a simple and effective way.

### 3 THE TWO-DIMENSIONAL CASE

Historically, the first successful attempt to eliminate the velocities was made by Baranger-Maitre-Oudin in<sup>3</sup>. Inspired by a previous result of Haugazeau-Lacoste<sup>11</sup> concerning  $H(\text{curl}; \Omega)$  spaces, they decided to look, for every element  $K$ , for a suitable bilinear form  $a_{K,h}(\mathbf{u}, \mathbf{v})$  of the type

$$a_{K,h}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^3 \omega_K^i (\mathbf{u}(M_i) \cdot \mathbf{n}_K^i) (\mathbf{v}(M_i) \cdot \mathbf{n}_K^i). \quad (23)$$

In (23)  $M_i$  represents the midpoint of the  $i$ -th edge, and  $\mathbf{n}_K^i$  is the outward unit normal to that edge ( $i = 1, 2, 3$ ). The weights  $\omega_K^i$  must be looked for in order to have

$$a_{K,h}(\mathbf{u}, \mathbf{v}) = a_K(\mathbf{u}, \mathbf{v}) \quad \forall \mathbf{u}, \mathbf{v} \text{ constant on } K. \quad (24)$$

After some manipulations, one discovers that such a bilinear form, satisfying (24), indeed exists, and that the weights  $\omega_K^i$  can be computed in the following way: let  $C_K$  be the circumcenter of  $K$  (that is, the center of the unique circle that passes through the vertices of  $K$ ), let  $H_K^i$  denote the distance of  $C_K$  from the  $i$ -th edge  $e_i$  ( $i = 1, 2, 3$ ), and let  $|e_i|$  be the length of  $e_i$ . The straight line containing  $e_i$  clearly splits the whole plane into two half-planes. If  $C_K$  belongs to the same half-plane containing  $K$ , then we set  $\omega_K^i = |e_i| H_K^i$ . Otherwise we set  $\omega_K^i = -|e_i| H_K^i$ . It is easy to check that if for instance all the angles of  $K$  are acute, then  $C_K$  falls inside  $K$ , and the choice  $\omega_K^i = |e_i| H_K^i$  will be made for all  $i$ 's. In this case, all the weights come out to be positive. If however the edge  $e_i$  is opposite to an obtuse angle, then  $\omega_K^i$  turns out to be  $-|e_i| H_K^i$ , and it will be negative. Up to a certain extent, this

could be tolerated (see<sup>3</sup> for further details). When  $e_i$  is opposite to a right angle, then  $H_K^i$  is zero, and so is  $\omega_K^i$ .

Coming back to the whole space  $V_{RT}$ , a classical construction of a basis for it is the following. For each edge  $e_k$  in  $\mathcal{T}_h$  we choose a unit vector  $\mathbf{n}^k$  normal to  $e_k$ . We do it for  $k = 1, \dots, NE$ , where  $NE$  is the number of edges in  $\mathcal{T}_h$ . Then, for each  $k$  we define the vector  $\mathbf{v}^k$  as the unique vector in  $V_{RT}$  that satisfies

$$\mathbf{v}^k \cdot \mathbf{n}^k = 1 \text{ on } e_k \quad \text{and} \quad \mathbf{v}^k \cdot \mathbf{n}^r = 0 \text{ on } e_r \quad \forall r \neq k. \quad (25)$$

It is immediate to check that for any vector  $\mathbf{v}$  of the form (7) and for any straight line  $\ell$ , the component of  $\mathbf{v}$  in the direction normal to  $\ell$  is constant on  $\ell$ , so that the definition (25) makes sense. It is also immediate to see that, with respect to this basis, the matrix

$$A_{r,k} := \sum_{K \in \mathcal{T}_h} a_{K,h}(\mathbf{v}^k, \mathbf{v}^r) \quad (26)$$

is diagonal. The idea is then the following one: change the original bilinear form  $a(\cdot, \cdot)$  into

$$a_h(\mathbf{u}_h, \mathbf{v}_h) := \sum_{K \in \mathcal{T}_h} a_{K,h}(\mathbf{u}_h, \mathbf{v}_h), \quad (27)$$

then change the original  $RT$  mixed formulation (11)-(12) into

find  $\mathbf{u}_h \in V_{RT}$  and  $p_h \in Q_h$  such that:

$$a_h(\mathbf{u}_h, \mathbf{v}_h) = b(p_h, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_{RT}, \quad (28)$$

$$b(q_h, \mathbf{u}_h) = (f, q_h) \quad \forall q_h \in Q_h, \quad (29)$$

and remark that this produces a system of the form (20)-(21) where now  $A$  is a diagonal matrix. Then eliminate  $U = A^{-1}B^tP$  to reach the form (22), with  $A^{-1}$  explicitly known. It is proved in<sup>3</sup> that the consistency error originated by the change of the bilinear form  $a(\cdot, \cdot)$  can be kept under control, and hence we still have optimal a priori error estimates. Surprisingly enough, the resulting scheme coincides with a classical Finite Volume scheme for diffusion operators (see e.g.<sup>9</sup>), where the flux on each edge  $e$ , common to the triangles  $K^1$  and  $K^2$ , is defined by dividing the jump  $p_h^1 - p_h^2$  by the distance  $\overline{C_1C_2}$  between the circumcenters of  $K^1$  and  $K^2$ .

There is however another more suitable mixed formulation, different from the BMO (28)-(29), that allows a simpler analysis. It is worth looking at it since, as we shall see in the next section, the BMO interpretation does not hold in three dimensions.

Assume, for simplicity, that all the angles of all the triangles are acute. This is not strictly necessary (in the sense that the condition can be weakened) but makes the exposition much simpler. In this case all the circumcenters will be internal to their respective triangles. Split every triangle  $K$  in three subtriangles using the circumcenters. Every internal edge  $e_k$  will belong to two such subtriangles: take the union of the two, and call it  $L_k$ . For the boundary edges we will have just one subtriangle, that we still call  $L_k$ . The union of all the  $L_k$  ( $k = 1, \dots, NE$ ) obtained in this way is still equal to  $\Omega$ . Consider now the new vector space

$$V_L := \{\mathbf{v} \mid \mathbf{v}|_{L_k} = c\mathbf{n}^k \text{ with } c \in \mathbb{R} \forall k = 1, \dots, NE\}, \quad (30)$$

where, as before,  $\mathbf{n}^k$  is the chosen unit vector normal to  $e_k$ . For vectors  $\mathbf{v} \in V_L$  and scalars  $q \in Q_h$  the bilinear form  $b(q, \mathbf{v})$  still makes sense, provided we use the second form in the right-hand side of the definition of  $b_K$  (16), giving

$$b(q, \mathbf{v}) = \sum_K \int_{\partial K} q \mathbf{v} \cdot \mathbf{n}_K \, ds. \quad (31)$$

Please do not confuse  $\mathbf{n}^k$  (normal to the edge  $e_k$ ) and  $\mathbf{n}_K$  (outward unit normal to  $\partial K$ ). In what follows it will be sometimes convenient to rearrange terms in the sum appearing in (31), making the sum over the edges rather than over the triangles. In order to do so, we first introduce the jumps of a piecewise constant function  $q \in Q_h$  on an edge  $e_k$  in the following way. Let  $K^1$  and  $K^2$  be the two triangles having  $e_k$  as an edge, let  $q^1$  and  $q^2$  be the corresponding values of  $q$ , and let  $\mathbf{n}_{K^1}$  and  $\mathbf{n}_{K^2}$  be the corresponding outward unit normals. If  $e_k$  is a boundary edge, belonging to a single triangle  $K$ , we set  $q^1 = q|_K$ ,  $q^2 = 0$ ,  $\mathbf{n}_{K^1} = \mathbf{n}_K$  and  $\mathbf{n}_{K^2} = -\mathbf{n}_K$ . The jump of  $q$  over  $e_k$  is the *vector*

$$\llbracket q \rrbracket_k := q^1 \mathbf{n}_{K^1} + q^2 \mathbf{n}_{K^2}. \quad (32)$$

Note, for future purposes, that the jump  $\llbracket q \rrbracket_k$  of  $q$  is normal to  $e_k$  and points toward the triangle where the value of  $q$  is lower. It is now easy to see that, whenever convenient, the bilinear form  $b$  given in (31) can be written (for  $\mathbf{v} \in V_L$  and  $q \in Q_h$ ) as

$$b(q, \mathbf{v}) = \sum_{k=1}^{NE} \int_{e_k} \llbracket q \rrbracket_k \cdot \mathbf{v} \, ds. \quad (33)$$

Another way of writing the bilinear form  $b$  can be obtained associating to every piecewise constant function  $q$  its “gradient”  $\mathbf{g}(q)$  defined as

$$\mathbf{g}(q)|_{L_k} = -\llbracket q \rrbracket_k / h_k, \quad (34)$$

with  $h_k$  given by

$$h_k = 2 \frac{\text{meas}(L_k)}{\text{meas}(e_k)}, \quad (35)$$

(that is, for internal edges,  $h_k$  is the distance of the two circumcenters). The minus sign in (34) is natural, since the gradient is expected to point toward the triangle where the value of  $q$  is bigger. From (34)-(35) we immediately see that  $\mathbf{g}(q)$  is the unique element in  $V_L$  such that

$$2 \int_{L_k} \mathbf{g}(q) \, d\mathbf{x} = - \int_{e_k} \llbracket q \rrbracket_k \, ds \quad \forall k = 1, \dots, NE. \quad (36)$$

Consequently, (31) and (36) imply

$$b(q_h, \mathbf{v}) = -2 \sum_{k=1}^{NE} \int_{L_k} \mathbf{g}(q_h) \cdot \mathbf{v} \, d\mathbf{x} \equiv -2a(\mathbf{g}(q_h), \mathbf{v}) \quad \forall \mathbf{v} \in V_L. \quad (37)$$

We finally observe that for  $\mathbf{u}$  and  $\mathbf{v}$  in  $V_L$  we have

$$\begin{aligned} 2a(\mathbf{u}, \mathbf{v}) &\equiv 2 \int_{\Omega} \mathbf{u} \cdot \mathbf{v} \, d\mathbf{x} = 2 \sum_k \int_{L_k} \mathbf{u} \cdot \mathbf{v} \, d\mathbf{x} = \sum_k |e_k| h_k (\mathbf{u} \cdot \mathbf{n}^k)(\mathbf{v} \cdot \mathbf{n}^k) \\ &= \sum_K \sum_{i=1}^3 \omega_K^i (\mathbf{u}(M_i) \cdot \mathbf{n}_K^i)(\mathbf{v}(M_i) \cdot \mathbf{n}_K^i). \end{aligned} \quad (38)$$

We can now consider the mixed formulation

$$\begin{aligned} & \text{find } \mathbf{u}_h \in V_L \text{ and } p_h \in Q_h \text{ such that:} \\ & 2a(\mathbf{u}_h, \mathbf{v}_h) = b(p_h, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_L, \end{aligned} \quad (39)$$

$$b(q_h, \mathbf{u}_h) = (f, q_h) \quad \forall q_h \in Q_h, \quad (40)$$

that, indeed, is just a different (and apparently a little more cumbersome) way of writing the BMO formulation (28)-(29). The analysis, however, can come out simpler. We give just a quick outline of it.

We consider two different approximations of the exact velocity  $\mathbf{u}$ . The first one, that we call  $\mathbf{u}_I$ , is defined as the unique element in  $V_L$  that satisfies

$$\int_{e_k} (\mathbf{u} - \mathbf{u}_I) \cdot \mathbf{n}^k \, ds = 0 \quad \forall k = 1, \dots, NE. \quad (41)$$

Observe that, using (31) and the divergence theorem, (41) gives

$$b(q_h, \mathbf{u}_I) = b(q_h, \mathbf{u}) = (f, q_h) \quad \forall q_h \in Q_h. \quad (42)$$

Hence, from (42) and (40) we immediately get

$$b(q_h, \mathbf{u}_h - \mathbf{u}_I) = 0 \quad \forall q_h \in Q_h. \quad (43)$$

A useful property. The second approximation for  $\mathbf{u}$ , that we call  $\mathbf{u}_I^*$  will be obtained by considering first  $p_I \in Q_h$  as the unique piecewise constant that verifies

$$p_I(C_K) = p(C_K) \quad \text{for } C_K = \text{circumcenter of } K \quad \forall K \in \mathcal{T}_h. \quad (44)$$

We then set

$$\mathbf{u}_I^* := -\mathbf{g}(p_I), \quad (45)$$

where we used the operator  $q \rightarrow \mathbf{g}(q)$  as defined in (36) or (34). Property (37) implies then:

$$2a(\mathbf{u}_I^*, \mathbf{v}) = -2a(\mathbf{g}(p_I), \mathbf{v}) = b(p_I, \mathbf{v}) \quad \forall \mathbf{v} \in V_L. \quad (46)$$

The error estimate now goes easily. Setting  $\mathbf{w} := \mathbf{u}_h - \mathbf{u}_I$ , and using (39), (46), and then (43) we deduce that

$$2a(\mathbf{u}_h - \mathbf{u}_I^*, \mathbf{w}) = b(p_h, \mathbf{w}) - b(p_I, \mathbf{w}) = 0 - 0 = 0. \quad (47)$$

Adding and subtracting  $\mathbf{u}_I^*$  and using the above property we get

$$\|\mathbf{w}\|^2 = a(\mathbf{w}, \mathbf{w}) = a(\mathbf{u}_h - \mathbf{u}_I^*, \mathbf{w}) + a(\mathbf{u}_I^* - \mathbf{u}_I, \mathbf{w}) = a(\mathbf{u}_I^* - \mathbf{u}_I, \mathbf{w}), \quad (48)$$

that easily implies  $\|\mathbf{w}\| \leq \|\mathbf{u}_I^* - \mathbf{u}_I\|$ . The proof ends by remarking that the line joining two circumcenters  $C_1$  and  $C_2$  of two triangles  $K^1$  and  $K^2$  having an edge  $e_k$  in common is perpendicular to  $e_k$ . This implies, using the definition of  $p_I$  (44) and the definition of  $\mathbf{g}$  (34), that  $\mathbf{u}_I^*$  equals the value of the normal part of  $\mathbf{u} \equiv -\nabla p$  on a point of the segment joining  $C_1$  and  $C_2$ . On the other hand, the normal component

of  $\mathbf{u}_I$  equals the value of the normal component of  $\mathbf{u}$  on a point of the edge  $e_k$ , and the difference of the two is easily bounded.

There is yet another way of reaching the linear system (6) that is less conventional. We can consider indeed the *BDM* mixed formulation (13)-(14) and observe that we can have a basis for  $V_{BDM}$  in the following way. In general, a vector  $\mathbf{v} \in V_{BDM}$  is determined by prescribing its (linear) normal component on each edge of  $\mathcal{T}_h$ . Remember that only the normal component of  $\mathbf{v}$  has to be continuous, while the vector itself, in general, is not. In particular, its tangential component will be doubly defined at each interface, and both components will be multiply defined at each vertex of  $\mathcal{T}_h$ . Now for every edge  $e_k$  (with  $k = 1, \dots, NE$ ) we consider the two endpoints (that will be two vertices in  $\mathcal{T}_h$ ), that we call  $S_k^1$  and  $S_k^2$ . We can then define a basis  $\{\mathbf{v}_k^s\}$  (with  $k = 1, \dots, NE$  and  $s = 1, 2$ ) as follows

$$\mathbf{v}_k^s(S_k^s) \cdot \mathbf{n}^k = 1 \quad \text{and} \quad \mathbf{v}_k^s(S_k^{s'}) \cdot \mathbf{n}^k = 0 \quad \text{if } k \neq k' \text{ or } s \neq s'. \quad (49)$$

It is clear that the value of each  $\mathbf{v}_k^s$  at a given vertex  $S$  will be multiply defined. But what we need is that  $\mathbf{v}_k^s$  can be reconstructed in a unique way in each triangle and that its normal component is continuous at the interfaces. The continuity of the vector itself is not to be expected, as the elements of  $V_{BDM}$  are *not*, in general, continuous. On the other hand we point out that the number of elements in our basis equals twice the number of edges, which is indeed the dimension of  $V_{BDM}$ . We can now, for each  $K \in \mathcal{T}_h$ , define another approximate bilinear form  $\tilde{a}_{K,h}$  as follows

$$\tilde{a}_{K,h}(\mathbf{u}, \mathbf{v}) = \frac{|K|}{3} \sum_{r=1}^3 \mathbf{u}(V_r) \cdot \mathbf{v}(V_r), \quad (50)$$

where  $V_1, V_2, V_3$  are the three vertices of  $K$ . We then proceed as before, defining

$$\tilde{a}_h(\mathbf{u}, \mathbf{v}) := \sum_K \tilde{a}_{K,h}(\mathbf{u}, \mathbf{v}), \quad (51)$$

and finally considering the problem

find  $\mathbf{u}_h \in V_{BDM}$  and  $p_h \in Q_h$  such that:

$$\tilde{a}_h(\mathbf{u}_h, \mathbf{v}_h) = b(p_h, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_{BDM}, \quad (52)$$

$$b(q_h, \mathbf{u}_h) = (f, q_h) \quad \forall q_h \in Q_h. \quad (53)$$

It is reasonably clear that when we write the above (52)-(53) as a linear system (20)-(21), the corresponding matrix  $A$  will be *block diagonal*, each block being associated to a vertex  $S$  in  $\mathcal{T}_h$ . The dimension of each block will be equal to the number of triangles having  $S$  in common ( $\equiv$  number of edges having  $S$  in common). Moreover, on each row there will be only three nonzero elements. Indeed, in the row corresponding to the degree of freedom  $\mathbf{v}(S) \cdot \mathbf{n}^k = 1$ , the three nonzeros will be the diagonal element and the two off diagonal elements corresponding to the degrees of freedom  $\mathbf{v}(S) \cdot \mathbf{n}^j = 1$  for the two edges  $e_j$  having  $S$  as an endpoint and sharing a triangle with  $e_k$ .

Hence, although not diagonal anymore (as it was for the BMO formulation), the explicit inversion of  $A$  will be feasible, thus leading to a final system of the required form (6).

We remark that in this case the normal component of  $\mathbf{u}_h$  on each edge will result in a suitably weighted combination of the values of  $p_h$  in all the triangles having a vertex in common with that edge. There must be a Finite Volume analogue of this formula, but we have not been able to find it. The analysis of the method, that goes along the same lines of<sup>3</sup>, will appear in<sup>7</sup>.

#### 4 THE THREE DIMENSIONAL CASE

We consider now the three dimensional problem. The definition of the local spaces  $RT(K)$  and  $BDM(K)$ , as given in (7) and (8) remains unchanged, as well as the definitions of the spaces  $V_{RT}$ ,  $V_{BDM}$ , and the bilinear forms  $a$  and  $b$ . Indeed, the whole Section 2 was dealing with the two-dimensional and the three-dimensional case at the same time.

The first change with respect to the two-dimensional case is that we cannot extend the BMO trick<sup>3</sup> in an immediate way. Actually, we can still write a formula of the type (23), that would diagonalize the approximated bilinear form  $a_h$  (see (27)) in the following way. Assume for simplicity that for every tetrahedron  $K$  the center  $C_K$  of its circumsphere (that is the unique sphere that passes through the four vertices of  $K$ ) lies inside  $K$ . As in the two-dimensional case this assumption can be relaxed, but to the expenses of the simplicity of the presentation. We also point out that this condition is stricter than assuming that the projection of each vertex on the opposite face falls inside the face. In fact this condition is not even satisfied by the usual reference tetrahedron having vertices  $(0, 0, 0)$ ,  $(1, 0, 0)$ ,  $(0, 1, 0)$ , and  $(0, 0, 1)$ .

Using the assumption  $C_K \in K$  we can now split every tetrahedron  $K$  in four subtetrahedra  $K_r$  joining, for each face  $f_r$  of  $K$ , the center  $C_K$  of the circumsphere to the three vertices of  $f_r$ . The projection of  $C_K$  onto the face  $f_r$  is the circumcenter  $C_r$  of  $f_r$ . We could then consider the formula

$$a_{K,h} := 3 \sum_{r=1}^4 |K_r| (\mathbf{u}(C_r) \cdot \mathbf{n}^r) (\mathbf{v}(C_r) \cdot \mathbf{n}^r) \quad (54)$$

where  $|K_r|$  is obviously the volume of  $K_r$ , and  $\mathbf{n}^r$  the unit normal to  $f_r$ . However it can be easily seen that such a formula will not give back  $a_K(\mathbf{u}, \mathbf{v})$  for constant vectors  $\mathbf{u}$  and  $\mathbf{v}$  (as we had in (24) for the two-dimensional case). Hence the analysis of BMO<sup>3</sup> cannot be extended.

We can however extend in a rather easy way the alternative analysis that we performed in the previous section, based on the spaces  $V_L$  and the related mixed formulation (39)-(40).

Keeping the assumption that every  $C_K$  is internal to  $K$ , and splitting again each tetrahedron in four tetrahedra  $K_r$ , we can now attach to each face  $f_k$  a region  $L_k$  as we did for triangles in the previous section. This allows us to define the space  $V_L$ , formally as in (30), and to proceed with the corresponding mixed formulation (39)-(40).

The jump of a piecewise constant  $q$  can still be defined as in (32), and the alternative way of writing  $b$  given in (33) still holds.

It is not difficult to check that the analysis sketched in the previous section works practically with no changes. It can also be seen that this gives back a classical Finite Volume scheme for diffusion operators (see e.g.<sup>9</sup>).

We shall show now that the numerical integration trick based on  $BDM$  spaces that we introduced in the previous section can be rather easily extended (in more than one way) to the three-dimensional case.

For this we remark that a vector  $\mathbf{v} \in BDM(K)$  is determined in a unique way by the value of its normal component  $v_n$  of the faces of  $K$  (see<sup>5</sup>). As the normal component is linear on each face, it will be enough to choose, in an arbitrary way, three points on every face, and prescribe the value of  $v_n$  at these points. If we look now at the assembled space  $V_{BDM}$  (as defined in (10)) we see that we need to fix three points in every face  $f_k$  of  $\mathcal{T}_h$  and prescribe the values of  $v_{n^k}$  there. We point out that the normal component must be continuous. Hence, the three points must be the same for both tetrahedrons having  $f_k$  as a face, and the assigned values must be the same.

The first (and, somehow, simpler) way of picking three points per face is to take the three vertices, which is the exact counterpart of what we did in the previous section when we took the two endpoints of the edge. Remember that  $\mathbf{v}$  does not need to be continuous. Hence we do not have to assign the values of  $\mathbf{v}$  at each vertex  $S$  of  $\mathcal{T}_h$ , but rather assign *for each face having  $S$  as a vertex* the value of the normal component of  $\mathbf{v}$  on that face (which will affect the values of  $\mathbf{v}$  only in the two tetrahedra having that face in common). This choice determines a natural choice of a basis  $\mathbf{v}_k^s$  in  $V_{BDM}$ , similar to what we did in (49), where now  $k$  ranges from 1 to  $NF$  (total number of faces in  $\mathcal{T}_h$ ), and  $s$  ranges from 1 to 3 (number of vertices on the  $k$ -th face).

We can now use the integration formula, quite similar to (50):

$$\tilde{a}_{K,h}(\mathbf{u}, \mathbf{v}) = \frac{|K|}{4} \sum_{r=1}^4 \mathbf{u}(V_r) \cdot \mathbf{v}(V_r), \tag{55}$$

where  $V_1, V_2, V_3, V_4$  are the four vertices of  $K$ , and then define (as in (51))

$$\tilde{a}_h(\mathbf{u}, \mathbf{v}) := \sum_K \tilde{a}_{K,h}(\mathbf{u}, \mathbf{v}). \tag{56}$$

It should be reasonably clear that the resulting matrix  $A$ , associated with  $\tilde{a}_h$  (as defined in (56)) will be block diagonal, each block corresponding to a vertex of  $\mathcal{T}_h$  and having a dimension equal to the number of faces sharing that vertex. Each block will be a rather sparse matrix, but its dimension could be easily of the order of 20 – 25. This allows the direct inversion of  $A$  (hence reaching the desired form (6),) but at a nonnegligible cost.

A way to reduce the dimension of the blocks would be to choose, for each face  $f_k$ , the midpoints of its three edges instead of the three vertices. This will produce, in a natural way, a different basis, that we can denote by  $\{\mathbf{w}_k^m\}$  where  $k$  ranges again from 1 to  $NF$ , and  $m$  ranges from 1 to 3 (number of midpoints of the edges of  $f_k$ ). Clearly, the dimension is the same as before ( $= 3NF$ ). In order to make the resulting matrix block diagonal with respect to this last basis, we would need, for each tetrahedron  $K$ , a bilinear form

$$a_{K,h}^*(\mathbf{u}, \mathbf{v}) := \sum_{i=1}^6 \alpha_K^i \mathbf{u}_{n^i}(M_i) \cdot \mathbf{v}_{n^i}(M_i), \tag{57}$$

where  $M_i$  represents the midpoint of the edge  $e_i$  and  $\mathbf{u}_{n^i}$  and  $\mathbf{v}_{n^i}$  represent *the part of  $\mathbf{u}$  (resp.  $\mathbf{v}$ ) which is normal to the edge  $e_i$* , that we can make precise as

$$\mathbf{v}_{n^i} := \mathbf{v} - (\mathbf{v} \cdot \mathbf{t}^i) \mathbf{t}^i \quad (58)$$

where  $\mathbf{t}^i$  is tangential direction to  $e_i$  (the sign, in the context of (58), is immaterial). The problem is now to find the weights  $\alpha_K^i$  in such a way that we have

$$a_{K,h}^*(\mathbf{u}, \mathbf{v}) = a_K(\mathbf{u}, \mathbf{v}) \equiv \int_K \mathbf{u} \cdot \mathbf{v} \, dx \quad (59)$$

whenever  $\mathbf{u}$  and  $\mathbf{v}$  are both constants. Luckily enough, the existence (and the actual values) of the coefficients  $\alpha_K^i$  can be deduced from the *true 3D analogue* of the formula (23) that we used in the two-dimensional case when performing the BMO trick on Raviart-Thomas spaces. The formula, introduced by Haugazeau-Lacoste<sup>11</sup> for  $H(\text{curl}; \Omega)$  spaces, reads

$$a_{K,h}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^6 \beta_K^i (\mathbf{u}(M_i) \cdot \mathbf{t}^i) (\mathbf{v}(M_i) \cdot \mathbf{t}^i). \quad (60)$$

It is proved in<sup>11</sup> that it is possible to find the coefficients  $\beta_K^i$  in such a way that, as in (24),

$$\sum_{i=1}^6 \beta_K^i (\mathbf{u}(M_i) \cdot \mathbf{t}^i) (\mathbf{v}(M_i) \cdot \mathbf{t}^i) = a_K(\mathbf{u}, \mathbf{v}) \equiv \int_K \mathbf{u} \cdot \mathbf{v} \, dx \quad (61)$$

hold for constant  $\mathbf{u}$  and  $\mathbf{v}$ . Taking  $\mathbf{u} = \mathbf{v} = \mathbf{e}_1 = (1, 0, 0)$ , then  $\mathbf{u} = \mathbf{v} = \mathbf{e}_2 = (0, 1, 0)$ , and finally  $\mathbf{u} = \mathbf{v} = \mathbf{e}_3 = (0, 0, 1)$  and summing, we easily get

$$\sum_{j=1}^3 \sum_{i=1}^6 \beta_K^i (\mathbf{e}_j \cdot \mathbf{t}^i)^2 = \sum_{i=1}^6 \beta_K^i \|\mathbf{t}^i\|^2 = \sum_{i=1}^6 \beta_K^i. \quad (62)$$

On the other hand, using (61) for each  $\mathbf{e}_j$  we have

$$\sum_{j=1}^3 \sum_{i=1}^6 \beta_K^i (\mathbf{e}_j \cdot \mathbf{t}^i)^2 = \sum_{j=1}^3 a_K(\mathbf{e}_j, \mathbf{e}_j) = 3|K|, \quad (63)$$

and comparing (62) and (63) we have

$$\sum_{i=1}^6 \beta_K^i = 3|K|. \quad (64)$$

We set now for  $i = 1, \dots, 6$

$$\alpha_K^i := \frac{\beta_K^i}{2}. \quad (65)$$

For every constant  $\mathbf{v}$  we have, using (57), then (58), then the fact that  $\|\mathbf{v}(M_i)\|$  does not depend on  $i$  ( $\mathbf{v}$  is constant!), (65) and (61), and then again (65) and (64) we have

$$\begin{aligned} a_{K,h}^*(\mathbf{v}, \mathbf{v}) &= \sum_{i=1}^6 \alpha_K^i \|\mathbf{v}_{n^i}(M_i)\|^2 = \sum_{i=1}^6 \alpha_K^i (\|\mathbf{v}(M_i)\|^2 - \|\mathbf{v}(M_i) \cdot \mathbf{t}^i\|^2) \\ &= \|\mathbf{v}\|^2 \sum_{i=1}^6 \alpha_K^i - \frac{|K|}{2} \|\mathbf{v}\|^2 = \|\mathbf{v}\|^2 \left( \frac{3|K|}{2} - \frac{|K|}{2} \right) = \|\mathbf{v}\|^2 |K|, \end{aligned} \quad (66)$$

implying that the required property (59) holds for  $\mathbf{u} = \mathbf{v} = \text{constant}$ . Looking now at the bilinear form

$$D(\mathbf{u}, \mathbf{v}) := a_{K,h}^*(\mathbf{u}, \mathbf{v}) - a_K(\mathbf{u}, \mathbf{v}), \tag{67}$$

we see that it is symmetric and equal to zero for  $\mathbf{u} = \mathbf{v} = \text{constant}$ . Hence

$$D(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(D(\mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v}) - D(\mathbf{u}, \mathbf{u}) - D(\mathbf{v}, \mathbf{v})) = 0 \tag{68}$$

for every constant  $\mathbf{u}, \mathbf{v}$  as desired.

It is clear that, with respect to the last basis  $\{\mathbf{w}_k^m\}$  the bilinear form

$$a_h^*(\mathbf{u}, \mathbf{v}) = \sum_K a_{K,h}^*(\mathbf{u}, \mathbf{v}) \tag{69}$$

will be again block-diagonal, each block corresponding now to a midpoint  $M$  of an edge. The dimension of each block will be equal to the number of faces having in common the edge  $e$  containing  $M$ . With an argument similar to that used at the end of Section 3, it can be seen that each row has again only three nonzero coefficients. In a general decomposition into tetrahedra, the typical number of elements having an edge in common is quite lower than the typical number of elements having a vertex in common, so that the new formulation allows an easier explicit inversion of the matrix  $A$  (compared with what we had in the previous *BDM* three-dimensional trick).

It can be easily seen that in the two cases (*BDM* three-dimensional tricks using the vertices or the midpoints of the edges) we can interpret the equation  $U = A^{-1}P$  as a way of recovering the flux from the piecewise constant values of the pressure. More precisely, the flux on each face  $f_k$  is recovered:

- taking, for each vertex  $S$  of  $f_k$ , a suitable average of the jumps of  $p_h$  in the tetrahedra having that vertex in common, in order to obtain “the value of  $u_{n^k}$  on  $f_k$  at  $S$ ”, and finally reconstructing a linear expression of  $u_{n^k}$  on  $f_k$  using its values at the three vertices (for the former vertex-based trick)
- taking, for each midpoint  $M$  of an edge  $e$  of  $f_k$ , a suitable average of the jumps of  $p_h$  in the tetrahedra having  $e$  as an edge, in order to obtain “the value of  $u_{n^k}$  on  $f_k$  at  $M$ ”, and finally reconstructing a linear expression of  $u_{n^k}$  on  $f_k$  using its values at the three midpoints (for the latter midpoint-based trick).

In both cases the coefficients of the “suitable combination” come out of the solution of the corresponding block equation.

It is very likely that Finite Volume schemes based on these formulae already exist in the literature. However we were not able to find them, and we are reasonably convinced that the Finite Volume motivation would be different from the present one. The analysis of these methods will be presented in.<sup>7</sup>

## REFERENCES

[1] P. Alotto and I. Perugia, Tree-cotree implicit condensation in Magnetostatics, *IEEE Trans. on Magnetics*, **36**, 1523-1526 (2000).

- [2] D.N. Arnold and F. Brezzi, Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates, *RAIRO Modél. Math. Anal. Numér.*, **19**, 7–32 (1985).
- [3] J. Baranger, J.F. Maitre, and F. Oudin, Connection between finite volume and mixed finite element methods, *M<sup>2</sup>AN*, **30** (4), 445–465 (1996).
- [4] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer-Verlag, to appear.
- [5] F. Brezzi, J. Douglas, jr., and L.D. Marini, Two families of mixed finite elements for second order elliptic problems, *Numer.Math.*, **47**, 217–235 (1985).
- [6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, (1991).
- [7] F. Brezzi, M. Fortin, and L.D. Marini, Error analysis of piecewise constant approximations of Darcy's law, in preparation.
- [8] F. Brezzi, T.J.R. Hughes, L.D. Marini, and A. Masud, Mixed Discontinuous Galerkin methods for Darcy flow, in preparation.
- [9] R. Eymard, T. Gallouët, and R. Herbin, Finite Volumes Methods, in *Handbook of Numerical Analysis*, (P.G. Ciarlet and J.L. Lions, eds.), Elsevier (2000).
- [10] B.X. Fraeijs de Veubeke, Displacement and equilibrium models in the finite element method, in *Stress Analysis*, (O.C. Zienkiewicz and G. Hollister, eds.), New York (1965).
- [11] Y. Haugazeau, P. Lacoste, Condensation de la matrice masse pour les éléments finis mixtes de  $H(\text{rot})$ , *C. R. Acad. Sci. Paris*, **316**, série I, 509–512 (1993).
- [12] T.J.R. Hughes and A. Masud, A stabilized Mixed Finite Element Method for Darcy Flow, in preparation.
- [13] 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 and E. Magenes, eds.), Lecture Notes in Math., Springer-Verlag, New York, **606**, 292–315 (1977).