

# Finite Element Approximation of Steady Flows of Colloidal Solutions

Andrea Bonito\*, Vivette Girault†, Diane Guignard‡,  
Kumbakonam R. Rajagopal§, and Endre Süli¶

August 9, 2021

## Abstract

We consider the mathematical analysis and numerical approximation of a system of nonlinear partial differential equations that arises in models that have relevance to steady isochoric flows of colloidal suspensions. The symmetric velocity gradient is assumed to be a monotone nonlinear function of the deviatoric part of the Cauchy stress tensor. We prove the existence of a weak solution to the problem, and under the additional assumption that the nonlinearity involved in the constitutive relation is Lipschitz continuous we also prove uniqueness of the weak solution. We then construct mixed finite element approximations of the system using both conforming and nonconforming finite element spaces. For both of these we prove the convergence of the method to the unique weak solution of the problem, and in the case of the conforming method we provide a bound on the error between the analytical solution and its finite element approximation in terms of the best approximation error from the finite element spaces. We propose first a Lions–Mercier type iterative method and next a classical fixed-point algorithm to solve the finite-dimensional problems resulting from the finite element discretisation of the system of nonlinear partial differential equations under consideration and present numerical experiments that illustrate the practical performance of the proposed numerical method.

*Keywords:* Non-Newtonian fluids, implicit constitutive theory, existence of weak solutions, mixed finite element approximation, convergence analysis

*2020 Mathematics Subject Classification.* Primary 35Q35, 65N12, 76D03, 76M10; Secondary 76A05

## 1 Introduction

The classical incompressible Navier–Stokes constitutive equation and its usual generalisations, the constitutive relations for the incompressible Stokesian fluid, are explicit expressions for the Cauchy stress in terms of the symmetric part of the velocity gradient. The Stokesian fluid is defined by the constitutive expression

$$\mathbf{T} = -p\mathbf{I} + \mathbf{f}(\mathbf{D}), \quad (1.1)$$

where  $\mathbf{T}$  is the Cauchy stress,  $-p\mathbf{I}$  is the indeterminate part of the stress due to the constraint of incompressibility and  $\mathbf{D}$  is the symmetric part of the velocity gradient,  $\mathbf{D} = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^t)$ . The incompressible

---

\*Texas A&M University, Department of Mathematics, 3368 TAMU, College Station, TX 77843-3368, USA. email: bonito@math.tamu.edu. Supported by the NSF Grant DMS-1817691;

†Sorbonne-Université, CNRS, Université de Paris, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France. email: girault@ann.jussieu.fr;

‡University of Ottawa, Department of Mathematics and Statistics, 150 Louis-Pasteur Pvt, Ottawa, ON, Canada K1N 6N5. email: dguignar@uottawa.ca;

§Texas A&M University, Department of Mechanical Engineering, College Station, TX 77843-3123, USA. email: krajagopal@tamu.edu;

¶Mathematical Institute, University of Oxford, Andrew Wiles Building, Woodstock Road, Oxford OX2 6GG, United Kingdom. email: sul@maths.ox.ac.uk.

Navier–Stokes fluid is a special sub-class of (1.1) that is linear in the symmetric part of the velocity gradient and is defined through:

$$\mathbf{T} = -p\mathbf{I} + 2\mu\mathbf{D}, \quad (1.2)$$

where  $\mu$  is the viscosity of the fluid. Power-law fluids are another popular sub-class of (1.1), the power-law fluid being defined through the constitutive equation

$$\mathbf{T} = -p\mathbf{I} + 2\mu_0(1 + \alpha \text{tr}(\mathbf{D}^2))^m \mathbf{D}, \quad (1.3)$$

where  $\mu_0$  and  $\alpha$  are positive constants and  $m$  is a constant; if  $m$  is zero we recover the Navier–Stokes fluid model, if it is negative we have a shear-thinning fluid model and if it is positive we have a shear-thickening fluid model. There are however many fluids that cannot be described by constitutive equations of the form (1.1) but require “relations”, in the true mathematical sense of the term, between the Cauchy stress and the symmetric part of the velocity gradient. Implicit constitutive relations that involve higher time derivatives of the stress and the symmetric part of the velocity gradient have been proposed to describe the response of non-Newtonian fluids that exhibit viscoelastic response<sup>1</sup> (see Burgers (1939) [13], Oldroyd (1950) [30]); that is fluids that exhibit phenomena like stress relaxation. However, purely implicit algebraic relationship between the stress and the symmetric part of the velocity gradient were not considered to describe non-Newtonian fluids until recently. Such models are critical if one is interested in describing the response of fluids which do not exhibit viscoelasticity but whose material properties depend on the mean value of the stress and the shear rate, a characteristic exhibited by many fluids and colloids, as borne out by numerous experiments. Consider for example an incompressible fluid whose viscosity depends on the mechanical pressure<sup>2</sup> (mean value of the stress) and is shear-thinning, whose constitutive relation takes the form

$$\mathbf{T} = -p\mathbf{I} + 2\mu(p, \text{tr}(\mathbf{D}^2))\mathbf{D}. \quad (1.4)$$

Since  $\text{tr}(\mathbf{D}) = \text{div}(\mathbf{u}) = 0$ ,

$$\text{tr}(\mathbf{T}) = -3p, \quad \text{i.e.,} \quad p = -\frac{1}{3}\text{tr}(\mathbf{T}), \quad (1.5)$$

the above equation takes the form

$$\mathbf{T} = \frac{1}{3}(\text{tr}(\mathbf{T}))\mathbf{I} + 2\mu\left(\frac{1}{3}\text{tr}(\mathbf{T}), \text{tr}(\mathbf{D}^2)\right)\mathbf{D}. \quad (1.6)$$

(The factor  $1/3$  is related to the number of space dimensions  $d = 3$ ; in two dimensions it would be replaced by  $1/2$ .) The above expression is of the form

$$\mathbf{f}(\mathbf{T}, \mathbf{D}) = \mathbf{0}, \quad (1.7)$$

which is an implicit relationship between the stress and the symmetric part of the velocity gradient. Rajagopal (2003) [34], (2006) [35] introduced the implicit relationship of the above form (and also the much more general implicit relationship between the history of the stress and the history of the deformation gradient) to describe materials whose properties depend upon the pressure and the shear rate. In fact, the properties of all fluids depend upon the pressure: it is just a matter of how large the variation of the pressure is in order for one to take the variation of the properties into account. The book by Bridgman (1931) [9] entitled “Physics of High Pressures” provides copious references to the experimental literature before 1931 on the variation of the viscosity of fluids with pressure, and one can find recent references to the experimental literature on the dependence of viscosity on pressure in Málek and Rajagopal (2006) [27]. Stokes (1845) [40] recognised that the viscosity of fluids varies with pressure, but in the case of sufficiently slow flows in channels and pipes he assumed that the viscosity could be considered a constant. Suffice to say, constitutive relations of the class (1.7) are necessary to describe the response of fluids whose viscosity depends on the pressure. Also as mentioned earlier, the implicit constitutive relation (1.7) is useful to describe the behaviour of colloids.

<sup>1</sup>While the Maxwell fluid (see Maxwell (1866) [28]) is defined through a constitutive relation involving the derivative of the stress, it is not an implicit model in that the symmetric part of the velocity gradient can be explicitly defined in terms of the stress and the time derivative of the stress.

<sup>2</sup>The terminology “pressure” is often misused, especially in nonlinear fluids; for a detailed discussion of the same see Rajagopal (2015) [37].

Recently, Perláková and Průša (2015) [32] (see also LeRoux and Rajagopal (2013) [39]) used an implicit model belonging to a sub-class of (1.7) to describe the response of colloidal solutions as presented in the experimental work of Boltzenhagen *et al.* (1997) [4], Hu *et al.* (1998) [21], Lopez-Diaz *et al.* (2010) [26] among others. Notice that while one always expresses the incompressible Navier–Stokes fluid by the representation (1.2), it is perfectly reasonable to describe it as

$$\mathbf{D} = \varphi \mathbf{I} + \frac{1}{2\mu} \mathbf{T}, \quad \text{where } \varphi = \frac{p}{2\mu}. \quad (1.8)$$

In fact, it is the representation (1.8) that is in keeping with causality as the stress is the cause and the velocity and hence its gradient is the effect, and this fact cannot be overemphasised. Such a representation would imply that the Stokes assumption that is often appealed to is incorrect (see Rajagopal (2013) [36] for a detailed discussion of the same). Málek *et al.* (2010) [33] generalised (1.8) to stress power-law fluids, namely constitutive relations of the form:

$$\begin{aligned} \mathbf{T} &= -p\mathbf{I} + \mathbf{T}^{\mathbf{d}}, \\ \mathbf{D} &= \gamma[1 + \beta \text{tr}((\mathbf{T}^{\mathbf{d}})^2)]^n \mathbf{T}^{\mathbf{d}}, \end{aligned} \quad (1.9)$$

where  $\mathbf{T}^{\mathbf{d}}$  is the deviatoric part of the Cauchy stress,  $\gamma$  and  $\beta$  are positive constants, and  $n$  is a constant that can be positive, negative or zero. The constitutive relation (1.9) is capable of describing phenomena that the classical power-law models are incapable of describing. For instance, the constitutive models (1.9) can describe limiting strain rate as well as fluids which allow the possibility of the strain rate initially increasing with stress and later decreasing with stress; both such responses cannot be described by the classical power-law fluid model (1.3) (see the discussion in Málek *et al.* (2010) [33] with regard to the difference in the response characteristics of the stress power-law fluid and the classical power-law fluid). We are interested in a further generalization of the constitutive relation of the form (1.9) that is appropriate for describing the response of colloidal solutions. This constitutive relation takes the form:

$$\mathbf{D} = \left\{ \gamma[1 + \beta \text{tr}((\mathbf{T}^{\mathbf{d}})^2)]^n + \alpha \right\} \mathbf{T}^{\mathbf{d}}, \quad (1.10)$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are positive constants,  $n$  is a real number, and  $\mathbf{T}^{\mathbf{d}}$  is the deviatoric part of the Cauchy stress. The shear stress in a fluid undergoing simple shear flow, that is described by the constitutive relation given above, increases from zero to a maximum, then decreases to a local minimum, and then increases monotonically as the shear stress increases from zero. As discussed by Le Roux and Rajagopal [39], and Perláková and Průša [32], many colloids exhibit such behavior. The constitutive relation that we introduce first in (2.10) and next in (3.1) includes (1.10) as a special sub-class. It can be posed within a Hilbert space setting owing to the presence of the coefficient  $\alpha$  in (1.10), but nevertheless, it is a challenging problem as it involves two nonlinearities: the monotone part in the constitutive relation and the inertial (convective) term. The problem without the inertial term, see Subsection 2.2 below, has already been analysed in [5], while the analysis of the steady-state incompressible Navier–Stokes equations is well-established, see for instance [41, 20]. With both nonlinearities present in the model, proving the existence of a weak solution, for instance, to the best of our knowledge cannot be done by simply coupling the techniques used for these two problems, namely the Browder–Minty theorem and the Galerkin method combined with Brouwer’s fixed point theorem and a weak compactness argument. More refined arguments are needed; they are crucial to the proofs of Lemmas 4 and 5 below.

This work is organised as follows. The notation and the functional-analytic setting are recalled in the next subsection. In Section 2, both linear and fully nonlinear versions of the formulation are briefly analysed for the Stokes system, i.e., without the inertial (convective) term. The theoretical analysis of the complete nonlinear system is carried out in Section 3. The main results of this section are Theorem 1 for the existence of a solution and Proposition 2 for the uniqueness of a solution under additional assumptions on the input data. In Section 4, conforming finite element approximations of these models are proposed and error estimates are derived. The cases of both simplicial and hexahedral elements are discussed. The analysis of the latter is less satisfactory as it requires subdivisions consisting of parallelepipeds and suffers from a higher computational cost. This motivates the introduction of nonconforming approximations in Section 5. In Section 6, two decoupling algorithms are presented and compared: a Lions–Mercier algorithm adapted

to a system with a monotone part and an elliptic part, and a classical fixed-point algorithm alternating between the approximation of a Navier-Stokes system and the nonlinear constitutive relation for the stress. Numerical experiments are performed with conforming finite elements on a square mesh in two dimensions. The theoretically established convergence of the scheme is confirmed and convergence of both decoupled algorithms is observed.

## 1.1 Notation and preliminaries

Let  $\Omega \subset \mathbb{R}^d$ ,  $d \in \{2, 3\}$ , be a bounded, open, simply connected Lipschitz domain. We consider the function spaces

$$Q := L_0^2(\Omega), \quad V := H_0^1(\Omega)^d \quad \text{and} \quad M := \{\mathbf{S} \in L^2(\Omega)_{\text{sym}}^{d \times d} : \text{tr}(\mathbf{S}) = 0\}, \quad (1.11)$$

for the pressure, the velocity, and the deviatoric stress tensor, respectively. As usual,

$$L_0^2(\Omega) = \left\{ q \in L^2(\Omega) : \int_{\Omega} q = 0 \right\},$$

the zero mean value constraint being introduced to fix the undetermined additive constant in the mechanical pressure. Here the subscript sym indicates that the  $d \times d$  tensors under consideration are assumed to be symmetric. Henceforth, the symmetric gradient of the velocity field  $\mathbf{v}$  (or, briefly, symmetric velocity gradient) will be denoted by

$$\mathbf{D}(\mathbf{v}) := \frac{1}{2}(\nabla \mathbf{v} + (\nabla \mathbf{v})^t) \quad (1.12)$$

and the deviatoric part of a  $d \times d$  tensor  $\mathbf{S}$  is defined by

$$\mathbf{S}^d := \mathbf{S} - \frac{1}{d} \text{tr}(\mathbf{S}) \mathbf{I} \quad (1.13)$$

with  $\mathbf{I}$  the  $d \times d$  identity tensor; thus the trace of  $\mathbf{S}^d$  is zero. We denote by  $\mathcal{V}$  the subspace of  $V$  consisting of all divergence-free functions contained in  $V$ ; that is,

$$\mathcal{V} := \{\mathbf{v} \in V : \text{div}(\mathbf{v}) = 0\}. \quad (1.14)$$

For vector-valued functions  $\mathbf{v} : \Omega \rightarrow \mathbb{R}^d$ , we write

$$\|\mathbf{v}\|_{L^2(\Omega)} := \|\mathbf{v}\|_{L^2(\Omega)} \quad \text{and} \quad \|\mathbf{v}\|_{L^\infty(\Omega)} := \|\mathbf{v}\|_{L^\infty(\Omega)}$$

with  $|\cdot|$  signifying the Euclidean norm on  $\mathbb{R}^d$ , while for tensor-valued functions  $\mathbf{S} : \Omega \rightarrow \mathbb{R}^{d \times d}$ , we define

$$\|\mathbf{S}\|_{L^2(\Omega)} := \|\mathbf{S}\|_{L^2(\Omega)},$$

where now

$$|\mathbf{S}| := \sqrt{\mathbf{S} : \mathbf{S}}$$

is the Frobenius norm of  $\mathbf{S}$ . Clearly,  $M$  is a Hilbert space with this norm. We recall the Poincaré and Korn inequalities, which are, respectively, the following: there exist positive constants  $C_P$  and  $C_K$  such that

$$\|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega) \quad (1.15)$$

and

$$\|\nabla \mathbf{v}\|_{L^2(\Omega)} \leq C_K \|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)} \quad \forall \mathbf{v} \in V. \quad (1.16)$$

We endow  $V$  (and  $\mathcal{V}$ ) with the norm

$$\|\cdot\|_V := \|\mathbf{D}(\cdot)\|_{L^2(\Omega)}. \quad (1.17)$$

Both  $V$  and  $\mathcal{V}$  are Hilbert spaces with this norm, because  $\|\cdot\|_V$  is equivalent to both the  $H^1(\Omega)^{d \times d}$  norm and the  $H^1(\Omega)^{d \times d}$  semi-norm, thanks to (1.15), (1.16) and the trivial relation  $\|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)} \leq \|\nabla \mathbf{v}\|_{L^2(\Omega)}$ .

## 2 Stokes system with linear and nonlinear constitutive relations

In this section we study two preliminary model problems without the inertial term; the first one simply reduces to the Stokes system, while the second model problem involves a monotone nonlinearity treated by the Browder–Minty approach.

### 2.1 The Stokes system

Let us consider the problem

$$\begin{cases} -\operatorname{div}(\mathbf{T}) &= \mathbf{f} & \text{in } \Omega, \\ \mathbf{D}(\mathbf{u}) &= \alpha \mathbf{T}^d & \text{in } \Omega, \\ \operatorname{div}(\mathbf{u}) &= 0 & \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} & \text{on } \partial\Omega, \end{cases} \quad (2.1)$$

where  $\mathbf{f} : \Omega \rightarrow \mathbb{R}^d$  is a prescribed external force,  $\mathbf{D}(\mathbf{u})$  is defined by (1.12), the unknown tensor  $\mathbf{T}$  is symmetric, and  $\alpha$  is a given positive constant, the reciprocal of the viscosity coefficient. Here, we assume that  $\mathbf{f} \in L^2(\Omega)^d$  for simplicity, but a similar analysis holds for the general case  $\mathbf{f} \in V' = H^{-1}(\Omega)^d$ ; see for instance Remark 1 in Section 3. By decomposing the Cauchy stress  $\mathbf{T}$  as  $\mathbf{T} = \mathbf{T}^d + \frac{1}{d}\operatorname{tr}(\mathbf{T})\mathbf{I}$  and inserting this in the first equation of (2.1) we arrive at the following equivalent problem:

$$\begin{cases} -\operatorname{div}(\mathbf{T}^d) - \frac{1}{d}\nabla\operatorname{tr}(\mathbf{T}) &= \mathbf{f} & \text{in } \Omega, \\ \mathbf{D}(\mathbf{u}) &= \alpha \mathbf{T}^d & \text{in } \Omega, \\ \operatorname{div}(\mathbf{u}) &= 0 & \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} & \text{on } \partial\Omega, \end{cases} \quad (2.2)$$

which we recognise to be the Stokes system where the mechanical pressure (mean normal stress) is  $p := -\frac{1}{d}\operatorname{tr}(\mathbf{T})$ . Recalling the spaces  $M, V, Q$  defined in (1.11) and using the relation

$$\mathbf{D}(\mathbf{v}) : \mathbf{S} = \nabla \mathbf{v} : \mathbf{S},$$

which holds<sup>3</sup> for any symmetric tensor  $\mathbf{S}$ , the weak formulation of problem (2.2) can be written as follows: find a triple  $(\mathbf{T}^d, \mathbf{u}, p) \in M \times V \times Q$  such that

$$\begin{aligned} \int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{v}) - \int_{\Omega} p \operatorname{div}(\mathbf{v}) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} & \forall \mathbf{v} \in V, \\ \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{S} - \int_{\Omega} \mathbf{S} : \mathbf{D}(\mathbf{u}) &= 0 & \forall \mathbf{S} \in M, \\ - \int_{\Omega} q \operatorname{div}(\mathbf{u}) &= 0 & \forall q \in Q. \end{aligned} \quad (2.3)$$

For any  $\mathbf{S} \in M$ ,  $\mathbf{v} \in V$ , and  $q \in Q$ , we set

$$\begin{aligned} b_1(\mathbf{S}, \mathbf{v}) &:= \int_{\Omega} \mathbf{S} : \mathbf{D}(\mathbf{v}), \\ b_2(\mathbf{v}, q) &:= - \int_{\Omega} q \operatorname{div}(\mathbf{v}). \end{aligned}$$

As is usual for the Stokes problem, the unknown pressure can be eliminated from (2.3) by restricting the test functions  $\mathbf{v}$  to  $\mathcal{V}$ . In addition, the variable  $\mathbf{u}$  can also be eliminated by treating the first line of (2.3) as a constraint, thus leading to an equivalent (reduced) problem for which the two variables  $p$  and  $\mathbf{u}$  are eliminated. The equivalence is based on the following (inf-sup) conditions

$$\inf_{\mathbf{v} \in \mathcal{V}} \sup_{\mathbf{S} \in M} \frac{b_1(\mathbf{S}, \mathbf{v})}{\|\mathbf{S}\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)}} \geq 1 \quad (2.4)$$

---

<sup>3</sup>For any  $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{d \times d}$ , with  $\mathbf{S}$  symmetric, we have that  $\mathbf{S} : \mathbf{R} = \left( \frac{\mathbf{S} + \mathbf{S}^t}{2} \right) : \mathbf{R} = \frac{1}{2} \mathbf{S} : \mathbf{R} + \frac{1}{2} \mathbf{S}^t : \mathbf{R} = \frac{1}{2} \mathbf{S} : \mathbf{R} + \frac{1}{2} \mathbf{S} : \mathbf{R}^t = \mathbf{S} : \left( \frac{\mathbf{R} + \mathbf{R}^t}{2} \right)$ .

and

$$\exists \beta > 0 : \inf_{q \in Q} \sup_{\mathbf{v} \in V} \frac{b_2(\mathbf{v}, q)}{\|q\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)}} \geq \inf_{q \in Q} \sup_{\mathbf{v} \in V} \frac{b_2(\mathbf{v}, q)}{\|q\|_{L^2(\Omega)} \|\nabla \mathbf{v}\|_{L^2(\Omega)}} \geq \beta, \quad (2.5)$$

where we have used that  $\|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)} \leq \|\nabla \mathbf{v}\|_{L^2(\Omega)}$ . It is well-known that the spaces  $V$  and  $Q$  defined in (1.11) satisfy the inf-sup condition (2.5), see for instance [20], while the relation (2.4) can be easily shown by observing that, for a given  $\mathbf{v} \in \mathcal{V}$ , we have  $\mathbf{R} := \mathbf{D}(\mathbf{v}) \in M$  since  $\text{tr}(\mathbf{R}) = \text{tr}(\mathbf{D}(\mathbf{v})) = \text{div}(\mathbf{v}) = 0$  and  $\mathbf{D}(\mathbf{v})$  is symmetric. Therefore,  $b_1(\mathbf{R}, \mathbf{v}) = \|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)}^2$  and thus

$$\sup_{\mathbf{S} \in M} \frac{b_1(\mathbf{S}, \mathbf{v})}{\|\mathbf{S}\|_{L^2(\Omega)}} \geq \frac{b_1(\mathbf{R}, \mathbf{v})}{\|\mathbf{R}\|_{L^2(\Omega)}} = \|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)}.$$

We can then eliminate the incompressibility constraint by seeking  $\mathbf{u} \in \mathcal{V}$ , yielding the (partially reduced) problem: find  $(\mathbf{T}^d, \mathbf{u}) \in M \times \mathcal{V}$  such that

$$\begin{aligned} \int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{v}) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \quad \forall \mathbf{v} \in \mathcal{V}, \\ \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{S} - \int_{\Omega} \mathbf{S} : \mathbf{D}(\mathbf{u}) &= 0 \quad \forall \mathbf{S} \in M. \end{aligned} \quad (2.6)$$

Clearly, each solution of (2.3) satisfies (2.6). Conversely, it follows from the inf-sup condition (2.5) that for any solution  $(\mathbf{T}^d, \mathbf{u})$  of (2.6) there exists a unique  $p \in Q$  such that  $(\mathbf{T}^d, \mathbf{u}, p)$  is the solution of (2.3); see [20]. Hence these two problems are equivalent. Furthermore, we can eliminate the unknown  $\mathbf{u}$  by proceeding as follows; see [5]. First, we introduce the decomposition  $M = \mathcal{M} \oplus \mathcal{M}^\perp$  with

$$\mathcal{M} := \{\mathbf{S} \in M : b_1(\mathbf{S}, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathcal{V}\}, \quad (2.7)$$

the kernel of  $b_1$ , and

$$\mathcal{M}^\perp := \{\mathbf{S} \in M : \int_{\Omega} \mathbf{S} : \mathbf{R} = 0 \quad \forall \mathbf{R} \in \mathcal{M}\}$$

its orthogonal complement in  $M$ , and we write  $\mathbf{T}^d = \mathbf{T}_0^d + \mathbf{T}_f^d$  with  $\mathbf{T}_0^d \in \mathcal{M}$  and  $\mathbf{T}_f^d \in \mathcal{M}^\perp$ . The condition (2.4) ensures the existence and uniqueness of  $\mathbf{T}_f^d \in \mathcal{M}^\perp$  satisfying

$$b_1(\mathbf{T}_f^d, \mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \quad \forall \mathbf{v} \in \mathcal{V} \quad \text{and} \quad \|\mathbf{T}_f^d\|_{L^2(\Omega)} \leq C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} \quad (2.8)$$

with  $C_P$  and  $C_K$  the constants in Poincaré's and Korn's inequalities (1.15) and (1.16), respectively. We finally get the (fully reduced) problem: find  $\mathbf{T}_0^d \in \mathcal{M}$  such that

$$\alpha \int_{\Omega} \mathbf{T}_0^d : \mathbf{S} = -\alpha \int_{\Omega} \mathbf{T}_f^d : \mathbf{S} \quad \forall \mathbf{S} \in \mathcal{M}. \quad (2.9)$$

The well-posedness of problem (2.9) follows from the Lax–Milgram lemma, while its equivalence to the original problem (2.3) is guaranteed by (2.4) and (2.5).

Of course, in this simple model with a linear constitutive relation,  $\mathbf{T}_0^d = \mathbf{0}$  since the right-hand side of (2.9) vanishes and  $\alpha(\cdot, \cdot)_\Omega$  is an inner product on  $\mathcal{M}$ . However, the framework developed here will be used in the sequel in a more general setting.

## 2.2 Stokes model with a nonlinear constitutive relation

Next, we consider the following Stokes-like system with a nonlinear relation between the stress and the symmetric velocity gradient:

$$\begin{cases} -\text{div}(\mathbf{T}^d) - \frac{1}{d} \nabla \text{tr}(\mathbf{T}) &= \mathbf{f} & \text{in } \Omega, \\ \mathbf{D}(\mathbf{u}) &= \alpha \mathbf{T}^d + \gamma \mu(|\mathbf{T}^d|) \mathbf{T}^d & \text{in } \Omega, \\ \text{div}(\mathbf{u}) &= 0 & \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} & \text{on } \partial\Omega \end{cases} \quad (2.10)$$

with  $\gamma$  a given positive constant, and where  $\mu \in \mathcal{C}^1((0, +\infty)) \cap \mathcal{C}^0([0, +\infty))$  is a given function satisfying

$$\frac{d}{da}(\mu(a)a) > 0 \quad \forall a \in \mathbb{R}_{>0} \quad (2.11)$$

and

$$\mu(a) > 0 \quad \text{and} \quad \mu(a)a \leq C_1 \quad \forall a \in \mathbb{R}_{\geq 0} \quad (2.12)$$

for some positive constant  $C_1$ . Since  $\mu$  is continuous on any subinterval of  $\mathbb{R}_{\geq 0}$ , the second part of (2.12) implies that  $\mu$  is bounded above and we denote its maximum by  $\mu_{\max}$ ,

$$0 < \mu(a) \leq \mu_{\max} \quad \forall a \in \mathbb{R}_{\geq 0}. \quad (2.13)$$

Moreover, proceeding as in the proof of [12, Lemma 4.1], we deduce from (2.11) and (2.12) that for any  $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{d \times d}$ , the following monotonicity property hold:

$$(\mu(|\mathbf{R}|)\mathbf{R} - \mu(|\mathbf{S}|)\mathbf{S}) : (\mathbf{R} - \mathbf{S}) \geq 0, \quad (2.14)$$

with equality if and only if  $\mathbf{R} = \mathbf{S}$ .

Introducing again  $p := -\frac{1}{d}\text{tr}(\mathbf{T})$ , the weak formulation of problem (2.10) reads as follows: find a triple  $(\mathbf{T}^d, \mathbf{u}, p) \in M \times V \times Q$  such that

$$\begin{aligned} \int_{\Omega} \mathbf{T}^d : D(\mathbf{v}) - \int_{\Omega} p \operatorname{div}(\mathbf{v}) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} & \forall \mathbf{v} \in V, \\ \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{S} - \int_{\Omega} \mathbf{S} : D(\mathbf{u}) &= 0 & \forall \mathbf{S} \in M, \\ - \int_{\Omega} q \operatorname{div}(\mathbf{u}) &= 0 & \forall q \in Q. \end{aligned} \quad (2.15)$$

Proceeding exactly as in Section 2.1, we first eliminate the pressure, and we thus deduce that problem (2.15) is equivalent to the following problem: find  $(\mathbf{T}^d, \mathbf{u}) \in M \times \mathcal{V}$  such that

$$\begin{aligned} \int_{\Omega} \mathbf{T}^d : D(\mathbf{v}) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} & \forall \mathbf{v} \in \mathcal{V}, \\ \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{S} - \int_{\Omega} \mathbf{S} : D(\mathbf{u}) &= 0 & \forall \mathbf{S} \in M, \end{aligned} \quad (2.16)$$

which is further equivalent to the following problem: find  $\mathbf{T}_0^d \in \mathcal{M}$  such that

$$\alpha \int_{\Omega} (\mathbf{T}_0^d + \mathbf{T}_f^d) : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{T}_0^d + \mathbf{T}_f^d|) (\mathbf{T}_0^d + \mathbf{T}_f^d) : \mathbf{S} = 0 \quad \forall \mathbf{S} \in \mathcal{M}, \quad (2.17)$$

with  $\mathbf{T}_f^d \in \mathcal{M}^\perp$  the solution of (2.8). The Browder–Minty theorem, see for instance [29], guarantees the existence of a solution to problem (2.17). Indeed, let  $\mathcal{A} : M \rightarrow M'$  be defined for  $\mathbf{R}, \mathbf{S} \in M$  by

$$\langle \mathcal{A}(\mathbf{R}), \mathbf{S} \rangle_M := \alpha \int_{\Omega} \mathbf{R} : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{R}|) \mathbf{R} : \mathbf{S}, \quad (2.18)$$

where  $\langle \cdot, \cdot \rangle_M$  denotes the duality pairing between  $M$  and its dual space,  $M'$ . It then easily follows that the mapping  $\mathbf{T}_0^d \mapsto \mathcal{A}(\mathbf{T}_0^d + \mathbf{T}_f^d)$  is bounded, monotone, coercive and hemi-continuous. By the Browder–Minty theorem these imply surjectivity of  $\mathcal{A}$  and thereby existence of a solution, while its uniqueness follows from the strict monotonicity of  $\mathcal{A}$ .

### 3 Navier–Stokes with nonlinear constitutive relation

Now, we focus on our problem of interest, where a convective term is added to the first equation of (2.10), i.e., we consider the problem

$$\begin{cases} (\mathbf{u} \cdot \nabla) \mathbf{u} - \operatorname{div}(\mathbf{T}) &= \mathbf{f} & \text{in } \Omega, \\ \mathbf{D}(\mathbf{u}) &= \alpha \mathbf{T}^d + \gamma \mu(|\mathbf{T}^d|) \mathbf{T}^d & \text{in } \Omega, \\ \operatorname{div}(\mathbf{u}) &= 0 & \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} & \text{on } \partial\Omega. \end{cases} \quad (3.1)$$

We prove a priori estimates, construct a solution, and give sufficient conditions for global uniqueness.

#### 3.1 Reformulation

By introducing the pressure  $p := -\frac{1}{d} \operatorname{tr}(\mathbf{T})$ , problem (3.1) can be rewritten as follows:

$$\begin{cases} (\mathbf{u} \cdot \nabla) \mathbf{u} - \operatorname{div}(\mathbf{T}^d) + \nabla p &= \mathbf{f} & \text{in } \Omega, \\ \mathbf{D}(\mathbf{u}) &= \alpha \mathbf{T}^d + \gamma \mu(|\mathbf{T}^d|) \mathbf{T}^d & \text{in } \Omega, \\ \operatorname{div}(\mathbf{u}) &= 0 & \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} & \text{on } \partial\Omega. \end{cases} \quad (3.2)$$

In order to bring forth an elliptic term on the left-hand side of the first equation of (3.2), we rewrite the second equation in (3.2) as

$$\mathbf{T}^d = \frac{1}{\alpha} \mathbf{D}(\mathbf{u}) - \frac{\gamma}{\alpha} \mu(|\mathbf{T}^d|) \mathbf{T}^d, \quad (3.3)$$

and thus by substituting this relation into the first equation of (3.2) we get

$$\begin{cases} (\mathbf{u} \cdot \nabla) \mathbf{u} - \frac{1}{\alpha} \operatorname{div}(\mathbf{D}(\mathbf{u})) + \nabla p &= \mathbf{f} - \frac{\gamma}{\alpha} \operatorname{div}(\mu(|\mathbf{T}^d|) \mathbf{T}^d) & \text{in } \Omega, \\ \alpha \mathbf{T}^d + \gamma \mu(|\mathbf{T}^d|) \mathbf{T}^d &= \mathbf{D}(\mathbf{u}) & \text{in } \Omega, \\ \operatorname{div}(\mathbf{u}) &= 0 & \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0} & \text{on } \partial\Omega. \end{cases} \quad (3.4)$$

The weak formulation of (3.4) reads: find  $(\mathbf{T}^d, \mathbf{u}, p) \in M \times V \times Q$  such that

$$\begin{aligned} \int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{u}] \cdot \mathbf{v} + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v}) - \int_{\Omega} p \operatorname{div}(\mathbf{v}) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{D}(\mathbf{v}), \\ \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{S} &= \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{S}, \\ \int_{\Omega} q \operatorname{div}(\mathbf{u}) &= 0 \end{aligned} \quad (3.5)$$

for all  $(\mathbf{S}, \mathbf{v}, q) \in M \times V \times Q$ .

As previously, we eliminate the pressure by restricting the test functions to  $\mathcal{V}$ , and we thus obtain the following equivalent reduced problem: find  $(\mathbf{T}^d, \mathbf{u}) \in M \times \mathcal{V}$  such that

$$\int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{u}] \cdot \mathbf{v} + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{D}(\mathbf{v}), \quad (3.6)$$

$$\alpha \int_{\Omega} \mathbf{T}^d : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{S} = \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{S} \quad (3.7)$$

for all  $(\mathbf{S}, \mathbf{v}) \in M \times \mathcal{V}$ .

Interestingly, (3.6), (3.7) can be further reduced by observing that, given  $\mathbf{u}$ , (3.7) uniquely determines  $\mathbf{T}^d$  thanks to the Browder–Minty theorem; see the end of Section 2.2. Thus, we define the mapping  $\mathcal{G} : \mathcal{V} \rightarrow M$  by  $\mathbf{u} \mapsto \mathbf{T}^d$  with  $\mathbf{T}^d \in M$  being the unique solution of

$$\langle \mathcal{A}(\mathbf{T}^d), \mathbf{S} \rangle_M = \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{S} \quad \forall \mathbf{S} \in M, \quad (3.8)$$



where we recall that  $\mathcal{A}$  is defined in (2.18). With this mapping, (3.6), (3.7) is equivalent to the following problem: find  $\mathbf{u} \in \mathcal{V}$  such that

$$\int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{u}] \cdot \mathbf{v} + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathcal{G}(\mathbf{u})|) \mathcal{G}(\mathbf{u}) : \mathbf{D}(\mathbf{v}). \quad (3.9)$$

Before embarking on the proof of existence of a solution to problem (3.6), (3.7) we establish a series of *a priori* estimates under the assumption that a solution exists.

### 3.2 *A priori* estimates

Assuming that problem (3.6), (3.7) has a solution, the following *a priori* estimates hold for any solution  $(\mathbf{T}^{\mathbf{d}}, \mathbf{u}) \in M \times \mathcal{V}$ .

**Lemma 1.** (*First a priori estimates*) Let  $|\Omega|$  denote the measure of  $\Omega$ . Then,

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \leq \alpha C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} + \gamma C_1 |\Omega|^{\frac{1}{2}} \quad (3.10)$$

and

$$\|\mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \leq C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} + \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}} \quad (3.11)$$

with  $C_P$  and  $C_K$  signifying the constants in Poincaré's and Korn's inequality, respectively, and  $C_1$  the constant in (2.12).

*Proof.* Taking  $\mathbf{S} = \mathbf{T}^{\mathbf{d}}$  in (3.7) yields

$$\alpha \|\mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)}^2 + \gamma \int_{\Omega} \mu(|\mathbf{T}^{\mathbf{d}}|) |\mathbf{T}^{\mathbf{d}}|^2 = \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{T}^{\mathbf{d}} \leq \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \|\mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)}.$$

Using then the positivity of  $\mu$ , see (2.12), we get

$$\|\mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \leq \frac{1}{\alpha} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}. \quad (3.12)$$

To obtain a bound for  $\mathbf{u}$ , we recall the well-known relation

$$\int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{v}] \cdot \mathbf{v} = 0 \quad \forall \mathbf{u} \in \mathcal{V}, \quad \forall \mathbf{v} \in V, \quad (3.13)$$

which is easily obtained by integration by parts, as follows:

$$\int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{v}] \cdot \mathbf{v} = \frac{1}{2} \int_{\Omega} \mathbf{u} \cdot \nabla (|\mathbf{v}|^2) = -\frac{1}{2} \int_{\Omega} \operatorname{div}(\mathbf{u}) |\mathbf{v}|^2 = 0.$$

Therefore, taking  $\mathbf{v} = \mathbf{u}$  in (3.6) and using (2.12) we obtain

$$\begin{aligned} \frac{1}{\alpha} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}^2 &= \int_{\Omega} \mathbf{f} \cdot \mathbf{u} + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}^{\mathbf{d}}|) \mathbf{T}^{\mathbf{d}} : \mathbf{D}(\mathbf{u}) \\ &\leq \left( C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} + \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}} \right) \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}, \end{aligned}$$

from which we directly deduce (3.10); (3.11) follows by applying (3.10) to (3.12).  $\square$

**Lemma 2.** (*Second a priori estimates*) Recall that  $\mu_{\max} := \sup_{s \in [0, \infty)} \mu(s)$ . We also have

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \leq (\alpha + \gamma \mu_{\max}) C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} \quad (3.14)$$

and

$$\|\mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \leq \frac{1}{\alpha} (\alpha + \gamma \mu_{\max}) C_P C_K \|\mathbf{f}\|_{L^2(\Omega)}. \quad (3.15)$$

The advantage of the estimates (3.14) and (3.15) is that if  $\mathbf{f} = \mathbf{0}$ , then we can directly deduce that  $\mathbf{u} = \mathbf{0}$  and  $\mathbf{T}^d = \mathbf{0}$  (and consequently  $p = 0$ ).

*Proof.* The ingredients of the proof are similar to those used in the proof of Lemma 1 and only the derivation of the bound for  $\mathbf{D}(\mathbf{u})$  is different. First notice that combining (3.6) and (3.7) we have

$$\int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{u}] \cdot \mathbf{v} + \int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \quad \forall \mathbf{v} \in \mathcal{V}. \quad (3.16)$$

Taking  $\mathbf{v} = \mathbf{u}$  in (3.16) we then find that

$$\int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{u}) \leq C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}. \quad (3.17)$$

Notice that  $\mathbf{T}^d : \mathbf{D}(\mathbf{u}) \geq 0$  a.e. in  $\Omega$ . Indeed, from (3.7) we have that

$$(\alpha + \gamma \mu(|\mathbf{T}^d|)) \mathbf{T}^d = \mathbf{D}(\mathbf{u}) \quad \text{in } M' \quad (3.18)$$

and thus

$$\underbrace{(\alpha + \gamma \mu(|\mathbf{T}^d|))}_{>0} \mathbf{T}^d : \mathbf{D}(\mathbf{u}) = |\mathbf{D}(\mathbf{u})|^2 \geq 0 \quad \text{a.e. in } \Omega.$$

Therefore, taking  $\mathbf{S} = \mathbf{D}(\mathbf{u})$  in (3.7) and using the upper bound  $\mu_{\max}$  for  $\mu$  and the bound (3.17) we have

$$\begin{aligned} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}^2 &= \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{u}) + \gamma \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{D}(\mathbf{u}) \\ &\leq (\alpha + \gamma \mu_{\max}) \int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{u}) \\ &\leq (\alpha + \gamma \mu_{\max}) C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}, \end{aligned}$$

which yields (3.14). Finally, the bound (3.15) for  $\mathbf{T}^d$  is obtained by substituting (3.14) in (3.12).  $\square$

**Remark 1.** Similar *a priori* estimates can be derived in the case when  $\mathbf{f} \in V'$  (with  $V' = H^{-1}(\Omega)^d$ ). More precisely, all occurrences of  $C_P C_K \|\mathbf{f}\|_{L^2(\Omega)}$  can be replaced by  $\|\mathbf{f}\|_{V'}$ , where

$$\|\mathbf{f}\|_{V'} := \sup_{\mathbf{v} \in \mathcal{V}} \frac{\langle \mathbf{f}, \mathbf{v} \rangle_V}{\|\mathbf{v}\|_V} = \sup_{\mathbf{v} \in \mathcal{V}} \frac{\langle \mathbf{f}, \mathbf{v} \rangle_V}{\|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)}}, \quad (3.19)$$

and  $\langle \cdot, \cdot \rangle_V$  denotes the duality pairing between  $V'$  and  $V$ . The same observation holds for all that follows.

**Remark 2.** By a direct argument we can also prove that

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \leq \frac{1}{\alpha} (\alpha + \gamma \mu_{\max})^2 C_P C_K \|\mathbf{f}\|_{L^2(\Omega)}. \quad (3.20)$$

This leads to the same *a priori* bound (3.15) for  $\mathbf{T}^d$ ,

$$\|\mathbf{T}^d\|_{L^2(\Omega)} \leq \frac{1}{\alpha} (\alpha + \gamma \mu_{\max}) C_P C_K \|\mathbf{f}\|_{L^2(\Omega)}.$$

Indeed, the choice  $\mathbf{S} = \mathbf{D}(\mathbf{u})$  in (3.7) gives directly (without invoking (3.18))

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}^2 = \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{u}) + \gamma \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{D}(\mathbf{u}) \leq (\alpha + \gamma \mu_{\max}) \|\mathbf{T}^d\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}.$$

Hence

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \leq (\alpha + \gamma \mu_{\max}) \|\mathbf{T}^d\|_{L^2(\Omega)}$$

and

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}^2 \leq (\alpha + \gamma \mu_{\max})^2 \|\mathbf{T}^d\|_{L^2(\Omega)}^2.$$

Then (3.20) follows by substituting the bound

$$\alpha \|\mathbf{T}^d\|_{L^2(\Omega)}^2 \leq \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{T}^d = \int_{\Omega} \mathbf{f} \cdot \mathbf{u} \leq C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)}$$

into the preceding inequality. This also yields (3.15).

### 3.3 Construction of a solution

In this subsection we prove the existence of a solution in a bounded Lipschitz domain without any restrictions on the data, other than those stated at the beginning of Section 2.2. The first part of the construction is fairly standard: a suitable sequence of (finite-dimensional) Galerkin approximations to the infinite-dimensional problem is constructed, followed by Brouwer's fixed point theorem to prove that each finite-dimensional problem in the sequence has a solution; uniform *a priori* estimates, similar to those derived in Lemma 1, are established for the Galerkin solutions, which are then used for passing to the (weak) limit, via a weak compactness argument. However, because of the combined effect of the nonlinearities, identifying the limit as a solution to the infinite-dimensional problem requires a more refined argument.

For the sake of clarity, the argument is split into several steps.

**Step 1** (Finite-dimensional approximation). Formulation (3.9) lends itself readily to a Galerkin discretisation. Since the only unknown is  $\mathbf{u}$  in  $\mathcal{V}$ , a separable Hilbert space, we introduce a countably infinite basis  $\{\mathbf{w}_1, \mathbf{w}_2, \dots\}$  of orthonormal functions of  $\mathcal{V}$  with respect to the inner product

$$(\mathbf{u}, \mathbf{v}) := \int_{\Omega} D(\mathbf{u}) : D(\mathbf{v}), \quad (3.21)$$

whose span is dense in  $\mathcal{V}$ . Next, we truncate this basis, i.e., for each  $m \geq 1$  we define

$$\mathcal{V}_m := \text{span}\{\mathbf{w}_1, \dots, \mathbf{w}_m\},$$

and for  $\mathbf{u}_m \in \mathcal{V}_m$  we denote by  $\hat{\mathbf{u}}_m \in \mathbb{R}^m$  its representation with respect to this basis. Finally, we fix  $m \geq 1$  and consider the following finite-dimensional problem: find  $\mathbf{u}_m \in \mathcal{V}_m$  such that, for all  $1 \leq j \leq m$ ,

$$\int_{\Omega} [(\mathbf{u}_m \cdot \nabla) \mathbf{u}_m] \cdot \mathbf{w}_j + \frac{1}{\alpha} \int_{\Omega} D(\mathbf{u}_m) : D(\mathbf{w}_j) = \int_{\Omega} \mathbf{f} \cdot \mathbf{w}_j + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}_m^{\mathbf{d}}|) \mathbf{T}_m^{\mathbf{d}} : D(\mathbf{w}_j) \quad (3.22)$$

with  $\mathbf{T}_m^{\mathbf{d}} := \mathcal{G}(\mathbf{u}_m)$ . In other words,  $\mathbf{T}_m^{\mathbf{d}} \in M$  solves

$$\alpha \int_{\Omega} \mathbf{T}_m^{\mathbf{d}} : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{T}_m^{\mathbf{d}}|) \mathbf{T}_m^{\mathbf{d}} : \mathbf{S} = \int_{\Omega} D(\mathbf{u}_m) : \mathbf{S} \quad \forall \mathbf{S} \in M. \quad (3.23)$$

Problem (3.22), which can be seen as the *projection* of (3.9) onto  $\mathcal{V}_m$ , is equivalent to the following: find  $\hat{\mathbf{u}}_m \in \mathbb{R}^m$  such that

$$\mathbf{F}(\hat{\mathbf{u}}_m) = \mathbf{0},$$

where  $\mathbf{F} = (F_1, \dots, F_m)^t : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is the continuous function defined, for  $j = 1, \dots, m$ , by

$$F_j(\hat{\mathbf{u}}_m) := \int_{\Omega} [(\mathbf{u}_m \cdot \nabla) \mathbf{u}_m] \cdot \mathbf{w}_j + \frac{1}{\alpha} \int_{\Omega} D(\mathbf{u}_m) : D(\mathbf{w}_j) - \int_{\Omega} \mathbf{f} \cdot \mathbf{w}_j - \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}_m^{\mathbf{d}}|) \mathbf{T}_m^{\mathbf{d}} : D(\mathbf{w}_j).$$

**Step 2** (Existence of a discrete solution). Problem (3.22) is a system of  $m$  nonlinear equations in  $m$  unknowns. The existence of a solution to this problem can be established by the following variant of Brouwer's fixed point theorem (see e.g. [17, 20]).

**Lemma 3.** *Let  $\mathbf{F} : \mathbb{R}^m \rightarrow \mathbb{R}^m$  be a continuous function that satisfies*

$$\mathbf{F}(\mathbf{x}) \cdot \mathbf{x} \geq 0 \quad \text{if} \quad |\mathbf{x}| = r$$

*for some  $r > 0$ . Then, there exists a point  $\mathbf{x} \in B_m(\mathbf{0}, r) := \{\mathbf{x} \in \mathbb{R}^m : |\mathbf{x}| \leq r\}$  such that*

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}.$$

**Proposition 1.** *Problem (3.22) has at least one solution  $\mathbf{u}_m \in \mathcal{V}_m$  that satisfies the uniform bound*

$$\|D(\mathbf{u}_m)\|_{L^2(\Omega)} \leq \alpha C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} + \gamma C_1 |\Omega|^{\frac{1}{2}}. \quad (3.24)$$

*Moreover,  $\mathbf{T}_m^{\mathbf{d}} = \mathcal{G}(\mathbf{u}_m)$  satisfies the uniform bound*

$$\|\mathbf{T}_m^{\mathbf{d}}\|_{L^2(\Omega)} \leq \frac{1}{\alpha} \|D(\mathbf{u}_m)\|_{L^2(\Omega)} \leq C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} + \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}}. \quad (3.25)$$

*Proof.* We infer from Lemma 3 that  $\mathbf{F}$  has a zero in the ball  $B_m(\mathbf{0}, r)$  with

$$r := \alpha C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} + \gamma C_1 |\Omega|^{\frac{1}{2}}.$$

Indeed, using the antisymmetry property (3.13), which holds because  $\mathbf{u}_m \in \mathcal{V}_m \subset \mathcal{V}$ , we get

$$\begin{aligned} \mathbf{F}(\hat{\mathbf{u}}_m) \cdot \hat{\mathbf{u}}_m &= \frac{1}{\alpha} \int_{\Omega} |\mathbf{D}(\mathbf{u}_m)|^2 - \int_{\Omega} \mathbf{f} \cdot \mathbf{u}_m - \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}_m^{\mathbf{d}}|) \mathbf{T}_m^{\mathbf{d}} : \mathbf{D}(\mathbf{u}_m) \\ &\geq \left( \frac{1}{\alpha} \|\mathbf{D}(\mathbf{u}_m)\|_{L^2(\Omega)} - C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} - \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}} \right) \|\mathbf{D}(\mathbf{u}_m)\|_{L^2(\Omega)}, \end{aligned}$$

where we have used Poincaré's and Korn's inequalities (1.15) and (1.16), respectively, to bound the second term and the relation (2.12) for the third one. As  $\|\mathbf{D}(\mathbf{u}_m)\|_{L^2(\Omega)} = |\hat{\mathbf{u}}_m|$ , we deduce from the last inequality that if  $|\hat{\mathbf{u}}_m| = r$  with  $r$  as defined above, then

$$\mathbf{F}(\hat{\mathbf{u}}_m) \cdot \hat{\mathbf{u}}_m \geq \left( \frac{1}{\alpha} |\hat{\mathbf{u}}_m| - C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} - \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}} \right) |\hat{\mathbf{u}}_m| = 0.$$

Thanks to Lemma 3, there exists a point  $\hat{\mathbf{u}}_m \in B_m(\mathbf{0}, r)$  such that  $\mathbf{F}(\hat{\mathbf{u}}_m) = \mathbf{0}$ , i.e., problem (3.22) has a solution  $\mathbf{u}_m \in \mathcal{V}_m$  that satisfies the uniform bound (3.24). Finally, it is easily shown that  $\mathbf{T}_m^{\mathbf{d}} := \mathcal{G}(\mathbf{u}_m)$  satisfies the bound (3.25).  $\square$

**Step 3** (Passage to the limit  $m \rightarrow \infty$  and identification of the limit). We consider the sequences  $(\mathbf{u}_m)_{m \geq 1}$  and  $(\mathbf{T}_m^{\mathbf{d}})_{m \geq 1}$  with  $\mathbf{u}_m \in \mathcal{V}_m$  and  $\mathbf{T}_m^{\mathbf{d}} = \mathcal{G}(\mathbf{u}_m) \in M$ . Thanks to the uniform estimates (3.24) and (3.25) there exist two subsequences (not relabelled) such that

$$\begin{aligned} \lim_{m \rightarrow \infty} \mathbf{u}_m &= \bar{\mathbf{u}} \quad \text{weakly in } H_0^1(\Omega)^d \text{ (and thus also in } \mathcal{V}), \\ \lim_{m \rightarrow \infty} \mathbf{u}_m &= \bar{\mathbf{u}} \quad \text{strongly in } L^q(\Omega)^d \text{ with } 1 \leq q < \infty \text{ if } d = 2, \text{ and } 1 \leq q < 6 \text{ if } d = 3, \\ \lim_{m \rightarrow \infty} \mathbf{T}_m^{\mathbf{d}} &= \bar{\mathbf{T}}^{\mathbf{d}} \quad \text{weakly in } L^2(\Omega)^{d \times d} \text{ (and thus also in } M), \end{aligned}$$

for some  $\bar{\mathbf{u}} \in \mathcal{V}$  and  $\bar{\mathbf{T}}^{\mathbf{d}} \in M$ . Our objective is to show that the pair  $(\bar{\mathbf{T}}^{\mathbf{d}}, \bar{\mathbf{u}}) \in M \times \mathcal{V}$  is a solution to the problem under consideration by passing to the limit in (3.22), (3.23).

Passing to the limit in (3.22), (3.23) is however not straightforward because of the lack of strong convergence of  $\mathbf{T}_m^{\mathbf{d}}$  in  $M$ . Identifying the pair  $(\bar{\mathbf{T}}^{\mathbf{d}}, \bar{\mathbf{u}}) \in M \times \mathcal{V}$  as a solution will be achieved by means of the following two lemmas, the first of which (Lemma 4) relies on the equations and the strong convergence of the sequence  $(\mathbf{u}_m)_{m \geq 1}$  in  $L^q(\Omega)^d$  shown above, and the second lemma (Lemma 5) follows from the monotonicity property (2.14).

The proof, included below, that the pair  $(\bar{\mathbf{T}}^{\mathbf{d}}, \bar{\mathbf{u}})$  satisfies (3.7) is inspired by the arguments in [11], where a more general constitutive relation than (3.3) was considered. Specifically, the conclusion of Lemma 5 follows from [11, Lemma 2.4.1], the hypothesis (2.12) of [11, Lemma 2.4.1] being fulfilled thanks to Lemma 4; however we provide a proof here that is directly tailored to our problem.

**Lemma 4.** *The following limit holds:*

$$\lim_{m \rightarrow \infty} \int_{\Omega} \mathbf{T}_m^{\mathbf{d}} : \mathbf{D}(\mathbf{u}_m) = \int_{\Omega} \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{D}(\bar{\mathbf{u}}). \quad (3.26)$$

*Proof.* By testing equation (3.23) with  $\mathbf{S} = \mathbf{D}(\mathbf{w}_j)$  and substituting into (3.22) we deduce that

$$\int_{\Omega} [(\mathbf{u}_m \cdot \nabla) \mathbf{u}_m] \cdot \mathbf{w}_j + \int_{\Omega} \mathbf{T}_m^{\mathbf{d}} : \mathbf{D}(\mathbf{w}_j) = \int_{\Omega} \mathbf{f} \cdot \mathbf{w}_j \quad \forall 1 \leq j \leq m. \quad (3.27)$$

Multiplying (3.27) by  $(\hat{\mathbf{u}}_m)_j$ , summing over  $j$ , and applying (3.13), we derive

$$\int_{\Omega} \mathbf{T}_m^{\mathbf{d}} : \mathbf{D}(\mathbf{u}_m) = \int_{\Omega} \mathbf{f} \cdot \mathbf{u}_m. \quad (3.28)$$

Thus we obtain on the one hand

$$\lim_{m \rightarrow \infty} \int_{\Omega} \mathbf{T}_m^{\mathbf{d}} : \mathbf{D}(\mathbf{u}_m) = \lim_{m \rightarrow \infty} \int_{\Omega} \mathbf{f} \cdot \mathbf{u}_m = \int_{\Omega} \mathbf{f} \cdot \bar{\mathbf{u}}. \quad (3.29)$$

On the other hand, letting  $m$  tend to infinity in (3.27) for fixed  $j$  and considering the strong convergence of  $\mathbf{u}_m$ , we infer that

$$\int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}}] \cdot \mathbf{w}_j + \int_{\Omega} \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{D}(\mathbf{w}_j) = \int_{\Omega} \mathbf{f} \cdot \mathbf{w}_j \quad \forall j \geq 1,$$

and the density of  $\bigcup_{m \geq 1} \mathcal{V}_m$  in  $\mathcal{V}$  therefore implies that

$$\int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}}] \cdot \mathbf{v} + \int_{\Omega} \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{D}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \quad \forall \mathbf{v} \in \mathcal{V}. \quad (3.30)$$

In view of (3.13), the choice  $\mathbf{v} = \bar{\mathbf{u}}$  in (3.30) yields

$$\int_{\Omega} \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{D}(\bar{\mathbf{u}}) = \int_{\Omega} \mathbf{f} \cdot \bar{\mathbf{u}}, \quad (3.31)$$

and (3.26) then follows from (3.29) and (3.31).  $\square$

**Lemma 5.** *We have that*

$$\bar{\mathbf{T}}^{\mathbf{d}} = \mathcal{G}(\bar{\mathbf{u}}). \quad (3.32)$$

*Proof.* Let  $\tilde{\mathbf{T}}^{\mathbf{d}} := \mathcal{G}(\bar{\mathbf{u}})$ ; since  $\mathbf{T}_m^{\mathbf{d}} := \mathcal{G}(\mathbf{u}_m)$ , we have by definition

$$\alpha \int_{\Omega} (\mathbf{T}_m^{\mathbf{d}} - \tilde{\mathbf{T}}^{\mathbf{d}}) : \mathbf{S} + \gamma \int_{\Omega} (\mu(|\mathbf{T}_m^{\mathbf{d}}|) \mathbf{T}_m^{\mathbf{d}} - \mu(|\tilde{\mathbf{T}}^{\mathbf{d}}|) \tilde{\mathbf{T}}^{\mathbf{d}}) : \mathbf{S} = \int_{\Omega} (\mathbf{D}(\mathbf{u}_m - \bar{\mathbf{u}})) : \mathbf{S}$$

for all  $\mathbf{S} \in M$ . Taking then  $\mathbf{S} = \mathbf{T}_m^{\mathbf{d}} - \tilde{\mathbf{T}}^{\mathbf{d}}$  and using the monotonicity property (2.14) we get

$$\begin{aligned} \alpha \int_{\Omega} |\mathbf{T}_m^{\mathbf{d}} - \tilde{\mathbf{T}}^{\mathbf{d}}|^2 &\leq \int_{\Omega} (\mathbf{D}(\mathbf{u}_m - \bar{\mathbf{u}})) : (\mathbf{T}_m^{\mathbf{d}} - \tilde{\mathbf{T}}^{\mathbf{d}}) \\ &= \int_{\Omega} \left[ \mathbf{D}(\mathbf{u}_m) : \mathbf{T}_m^{\mathbf{d}} - \mathbf{D}(\mathbf{u}_m) : \tilde{\mathbf{T}}^{\mathbf{d}} - \mathbf{D}(\bar{\mathbf{u}}) : (\mathbf{T}_m^{\mathbf{d}} - \tilde{\mathbf{T}}^{\mathbf{d}}) \right]. \end{aligned}$$

Finally, we take the limit  $m \rightarrow \infty$  of both sides and apply (3.26) to obtain

$$\alpha \lim_{m \rightarrow \infty} \int_{\Omega} |\mathbf{T}_m^{\mathbf{d}} - \tilde{\mathbf{T}}^{\mathbf{d}}|^2 \leq \int_{\Omega} \left[ \mathbf{D}(\bar{\mathbf{u}}) : \bar{\mathbf{T}}^{\mathbf{d}} - \mathbf{D}(\bar{\mathbf{u}}) : \tilde{\mathbf{T}}^{\mathbf{d}} - \mathbf{D}(\bar{\mathbf{u}}) : (\bar{\mathbf{T}}^{\mathbf{d}} - \tilde{\mathbf{T}}^{\mathbf{d}}) \right] = 0,$$

which implies (3.32) as well as the *strong* convergence in  $M$  of  $\mathbf{T}_m^{\mathbf{d}}$  to  $\bar{\mathbf{T}}^{\mathbf{d}}$ .  $\square$

**Theorem 1.** (*Existence of a solution*) *The pair  $(\bar{\mathbf{T}}^{\mathbf{d}}, \bar{\mathbf{u}}) \in M \times \mathcal{V}$  solves (3.6), (3.7).*

*Proof.* It follows from Lemma 5 that on the one hand  $(\bar{\mathbf{T}}^{\mathbf{d}}, \bar{\mathbf{u}})$  solves (3.7) and on the other hand,

$$\lim_{m \rightarrow \infty} \mu(|\mathbf{T}_m^{\mathbf{d}}|) \mathbf{T}_m^{\mathbf{d}} = \mu(|\bar{\mathbf{T}}^{\mathbf{d}}|) \bar{\mathbf{T}}^{\mathbf{d}} \text{ weakly in } M.$$

Indeed, passing to the limit in (3.23) gives, for any  $\mathbf{S} \in M$ ,

$$\begin{aligned} \lim_{m \rightarrow \infty} \gamma \int_{\Omega} \mu(|\mathbf{T}_m^{\mathbf{d}}|) \mathbf{T}_m^{\mathbf{d}} : \mathbf{S} &= \lim_{m \rightarrow \infty} \left( \int_{\Omega} \mathbf{D}(\mathbf{u}_m) : \mathbf{S} - \alpha \int_{\Omega} \mathbf{T}_m^{\mathbf{d}} : \mathbf{S} \right) \\ &= \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \mathbf{S} - \alpha \int_{\Omega} \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{S} \\ &= \gamma \int_{\Omega} \mu(|\bar{\mathbf{T}}^{\mathbf{d}}|) \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{S}. \end{aligned}$$

Therefore, taking the limit as  $m \rightarrow \infty$  in (3.22) we get

$$\int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}}] \cdot \mathbf{w}_j + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \mathbf{D}(\mathbf{w}_j) = \int_{\Omega} \mathbf{f} \cdot \mathbf{w}_j + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\bar{\mathbf{T}}^{\mathbf{d}}|) \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{D}(\mathbf{w}_j)$$

for each  $j = 1, 2, \dots$ , and thus the density of  $\bigcup_{m \geq 1} \mathcal{V}_m$  in  $\mathcal{V}$  implies that

$$\int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}}] \cdot \mathbf{v} + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \mathbf{D}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\bar{\mathbf{T}}^{\mathbf{d}}|) \bar{\mathbf{T}}^{\mathbf{d}} : \mathbf{D}(\mathbf{v}) \quad \forall \mathbf{v} \in \mathcal{V},$$

which is precisely (3.6).  $\square$

### 3.4 Global conditional uniqueness

We now prove global uniqueness of the solution under additional assumptions on the function  $\mu$  and the input data. The notion of uniqueness we establish is global and conditional in the sense that it holds under suitable restrictions on the data, but it is also global because no other solution exists.

Let  $\mathbb{R}_{\text{sym},0}^{d \times d}$  denote the space of symmetric  $d \times d$  matrices with vanishing trace and let  $C_S$  be the smallest positive constant in the following Sobolev embedding:

$$\|\mathbf{v}\|_{L^4(\Omega)} \leq C_S \|\nabla \mathbf{v}\|_{L^2(\Omega)} \quad \forall \mathbf{v} \in V. \quad (3.33)$$

**Proposition 2.** (Uniqueness) Assume that the function  $\mathbf{T}^{\mathbf{d}} \mapsto \mu(|\mathbf{T}^{\mathbf{d}}|) \mathbf{T}^{\mathbf{d}}$  is Lipschitz continuous in  $\mathbb{R}_{\text{sym},0}^{d \times d}$ , i.e., there exists a positive constant  $\Lambda$  such that

$$|\mu(|\mathbf{T}^{\mathbf{d}}|) \mathbf{T}^{\mathbf{d}} - \mu(|\mathbf{S}^{\mathbf{d}}|) \mathbf{S}^{\mathbf{d}}| \leq \Lambda |\mathbf{T}^{\mathbf{d}} - \mathbf{S}^{\mathbf{d}}| \quad \forall \mathbf{T}^{\mathbf{d}}, \mathbf{S}^{\mathbf{d}} \in \mathbb{R}_{\text{sym},0}^{d \times d}. \quad (3.34)$$

If the input data satisfy

$$\frac{\gamma}{\alpha} \Lambda + \alpha^2 C_S^2 C_P C_K^4 \|\mathbf{f}\|_{L^2(\Omega)} + \alpha \gamma C_S^2 C_K^3 C_1 |\Omega|^{\frac{1}{2}} < 1 \quad (3.35)$$

then the solution of problem (3.6), (3.7) is unique.

*Proof.* We use a variational argument. Suppose that  $(\mathbf{T}_1^{\mathbf{d}}, \mathbf{u}_1), (\mathbf{T}_2^{\mathbf{d}}, \mathbf{u}_2) \in M \times \mathcal{V}$  are solutions of (3.6), (3.7). Let us write  $\delta \mathbf{T}^{\mathbf{d}} := \mathbf{T}_1^{\mathbf{d}} - \mathbf{T}_2^{\mathbf{d}}$  and  $\delta \mathbf{u} := \mathbf{u}_1 - \mathbf{u}_2$ . Subtracting the equations solved by  $(\mathbf{T}_2^{\mathbf{d}}, \mathbf{u}_2)$  from those solved by  $(\mathbf{T}_1^{\mathbf{d}}, \mathbf{u}_1)$  we get for all  $(\mathbf{S}, \mathbf{v}) \in M \times \mathcal{V}$  the following pair of equalities:

$$\int_{\Omega} [(\mathbf{u}_1 \cdot \nabla) \mathbf{u}_1 - (\mathbf{u}_2 \cdot \nabla) \mathbf{u}_2] \cdot \mathbf{v} + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\delta \mathbf{u}) : \mathbf{D}(\mathbf{v}) = \frac{\gamma}{\alpha} \int_{\Omega} (\mu(|\mathbf{T}_1^{\mathbf{d}}|) \mathbf{T}_1^{\mathbf{d}} - \mu(|\mathbf{T}_2^{\mathbf{d}}|) \mathbf{T}_2^{\mathbf{d}}) : \mathbf{D}(\mathbf{v}), \quad (3.36)$$

$$\alpha \int_{\Omega} \delta \mathbf{T}^{\mathbf{d}} : \mathbf{S} + \gamma \int_{\Omega} (\mu(|\mathbf{T}_1^{\mathbf{d}}|) \mathbf{T}_1^{\mathbf{d}} - \mu(|\mathbf{T}_2^{\mathbf{d}}|) \mathbf{T}_2^{\mathbf{d}}) : \mathbf{S} = \int_{\Omega} \mathbf{D}(\delta \mathbf{u}) : \mathbf{S}. \quad (3.37)$$

The choice  $\mathbf{S} = \delta \mathbf{T}^{\mathbf{d}}$  in (3.37), thanks to the monotonicity property (2.14), leads to

$$\|\delta \mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \leq \frac{1}{\alpha} \|\mathbf{D}(\delta \mathbf{u})\|_{L^2(\Omega)}. \quad (3.38)$$

Then, by noting that

$$\int_{\Omega} [(\mathbf{u}_1 \cdot \nabla) \mathbf{u}_1 - (\mathbf{u}_2 \cdot \nabla) \mathbf{u}_2] \cdot \mathbf{v} = \int_{\Omega} [(\delta \mathbf{u} \cdot \nabla) \mathbf{u}_1] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}_2 \cdot \nabla) \delta \mathbf{u}] \cdot \mathbf{v},$$

by testing (3.36) with  $\mathbf{v} = \delta \mathbf{u}$ , and recalling (3.13) we obtain

$$\begin{aligned} \frac{1}{\alpha} \|\mathbf{D}(\delta \mathbf{u})\|_{L^2(\Omega)}^2 &= \frac{\gamma}{\alpha} \int_{\Omega} (\mu(|\mathbf{T}_1^{\mathbf{d}}|) \mathbf{T}_1^{\mathbf{d}} - \mu(|\mathbf{T}_2^{\mathbf{d}}|) \mathbf{T}_2^{\mathbf{d}}) : \mathbf{D}(\delta \mathbf{u}) - \int_{\Omega} [(\delta \mathbf{u} \cdot \nabla) \mathbf{u}_1] \cdot \delta \mathbf{u} \\ &\stackrel{(3.34)}{\leq} \frac{\gamma}{\alpha} \Lambda \|\delta \mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \|\mathbf{D}(\delta \mathbf{u})\|_{L^2(\Omega)} + \|\delta \mathbf{u}\|_{L^4(\Omega)}^2 \|\nabla \mathbf{u}_1\|_{L^2(\Omega)} \\ &\stackrel{(3.33), (1.16)}{\leq} \frac{\gamma}{\alpha} \Lambda \|\delta \mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \|\mathbf{D}(\delta \mathbf{u})\|_{L^2(\Omega)} + C_S^2 C_K^3 \|\mathbf{D}(\mathbf{u}_1)\|_{L^2(\Omega)} \|\mathbf{D}(\delta \mathbf{u})\|_{L^2(\Omega)}^2 \\ &\stackrel{(3.38), (3.10)}{\leq} \left[ \frac{\gamma}{\alpha^2} \Lambda + C_S^2 C_K^3 \left( \alpha C_P C_K \|\mathbf{f}\|_{L^2(\Omega)} + \gamma C_1 |\Omega|^{\frac{1}{2}} \right) \right] \|\mathbf{D}(\delta \mathbf{u})\|_{L^2(\Omega)}^2. \end{aligned}$$

The assumption (3.35) on the data guarantees that the factor on the right-hand side of the last inequality is strictly smaller than  $\frac{1}{\alpha}$ , thus implying that  $\|\mathbf{D}(\delta \mathbf{u})\|_{L^2(\Omega)} = 0$ , i.e.,  $\mathbf{u}_1 = \mathbf{u}_2$ . Finally, applying this result to (3.38) yields  $\mathbf{T}_1^{\mathbf{d}} = \mathbf{T}_2^{\mathbf{d}}$ .  $\square$

**Remark 3.** The strategy used in deriving the second *a priori* estimate stated in Lemma 2 leads to uniqueness when (3.35) is replaced by

$$\frac{\gamma}{\alpha} \Lambda + \alpha C_S^2 C_P C_K^4 (\alpha + \gamma \mu_{\max}) \|\mathbf{f}\|_{L^2(\Omega)} < 1. \quad (3.39)$$

In fact both strategies lead to the same condition (3.39); namely, we also get (3.39) by using (3.14) instead of (3.10) to bound  $\|\mathbf{D}(\mathbf{u}_1)\|_{L^2(\Omega)}$  in the proof of Proposition 2.

Note that both (3.35) and (3.39) hold when  $\gamma$  and  $\mathbf{f}$  are sufficiently small.

**Remark 4.** Under the Lipschitz condition (3.34), the proof of (3.20) and (3.15) is valid with  $\mu_{\max}$  replaced by  $\Lambda$ , and the  $L^2(\Omega)^d$  norm of  $\mathbf{f}$  (multiplied by  $C_P C_K$ ) replaced by its norm in  $\mathcal{V}'$ , see Remark 1. More precisely,

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \leq \frac{1}{\alpha} (\alpha + \gamma \Lambda)^2 \|\mathbf{f}\|_{\mathcal{V}'}, \quad (3.40)$$

$$\|\mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \leq \frac{1}{\alpha} (\alpha + \gamma \Lambda) \|\mathbf{f}\|_{\mathcal{V}'}. \quad (3.41)$$

### 3.5 Comparison of the *a priori* estimates

At this stage, it is useful to compare the *a priori* estimates derived in the previous sections. We have

$$\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \leq C_{\mathbf{u}} := \min \left\{ (\alpha + \gamma \mu_{\max}) \|\mathbf{f}\|_{\mathcal{V}'}, \alpha \|\mathbf{f}\|_{\mathcal{V}'} + \gamma C_1 |\Omega|^{\frac{1}{2}}, \frac{1}{\alpha} (\alpha + \gamma \Lambda)^2 \|\mathbf{f}\|_{\mathcal{V}'} \right\}, \quad (3.42)$$

$$\|\mathbf{T}^{\mathbf{d}}\|_{L^2(\Omega)} \leq C_{\mathbf{T}^{\mathbf{d}}} := \min \left\{ \frac{1}{\alpha} (\alpha + \gamma \mu_{\max}) \|\mathbf{f}\|_{\mathcal{V}'}, \|\mathbf{f}\|_{\mathcal{V}'} + \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}}, \frac{1}{\alpha} (\alpha + \gamma \Lambda) \|\mathbf{f}\|_{\mathcal{V}'} \right\}, \quad (3.43)$$

where  $\Lambda$  is replaced by  $\mu_{\max}$  if we do not make the Lipschitz assumption (3.34). For  $p$  we have

$$\|p\|_{L^2(\Omega)} \leq \frac{1}{\beta} \left( C_S^2 C_K^2 C_{\mathbf{u}}^2 + \|\mathbf{f}\|_{(\mathcal{V}^\perp)'} + \min \left\{ \min \left( 1, \frac{\gamma \mu_{\max}}{\alpha}, \frac{\gamma \Lambda}{\alpha} \right) C_{\mathbf{T}^{\mathbf{d}}}, \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}} \right\} \right),$$

where  $\mathcal{V}^\perp$  denotes the orthogonal complement of  $\mathcal{V}$  in  $V$  with respect to the inner product (3.21).

**Remark 5.** We can replace  $C_S^2$  by the product  $C_p C_r$  of the smallest constants  $C_p$  and  $C_r$  from the Sobolev embedding of  $H^1(\Omega)^d$  into  $L^p(\Omega)^d$  and  $L^r(\Omega)^d$ , respectively, with  $p = 6$  and  $r = 3$ . We could also use the best constant  $\hat{C}$  such that

$$\int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{v}] \cdot \mathbf{w} \leq \hat{C} \|\nabla \mathbf{u}\|_{L^2(\Omega)} \|\nabla \mathbf{v}\|_{L^2(\Omega)} \|\nabla \mathbf{w}\|_{L^2(\Omega)},$$

or even

$$\int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{v}] \cdot \mathbf{w} \leq \hat{C} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{w})\|_{L^2(\Omega)}.$$

In the former case,  $\hat{C} \leq C_p C_r$  while in the latter case,  $\hat{C} \leq C_K^3 C_p C_r$ .

## 4 Conforming finite element approximation

In this section, we study conforming finite element approximations of problem (3.2), where conformity refers to the discrete velocity space. To facilitate the implementation, it is useful to relax the zero trace restriction on the discrete tensor space, but this is not quite a nonconformity because the theoretical analysis of the preceding sections holds without this condition. In particular, the inf-sup condition (2.4) is still valid (supremum over a larger space).

We start with the numerical analysis of general conforming approximations, including existence of discrete solutions, convergence, and error estimates, and give specific examples further on.

#### 4.1 General conforming approximation

As stated above, here  $M = L^2(\Omega)_{\text{sym}}^{d \times d}$ . Up to this modification, we propose to discretise the formulation derived from (3.2): find  $(\mathbf{T}^d, \mathbf{u}, p) \in M \times V \times Q$  such that

$$\begin{aligned} \int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{u}] \cdot \mathbf{v} + \int_{\Omega} \mathbf{T}^d : \mathbf{D}(\mathbf{v}) + b_2(\mathbf{v}, p) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v} & \forall \mathbf{v} \in V, \\ \alpha \int_{\Omega} \mathbf{T}^d : \mathbf{S} + \gamma \int_{\Omega} \mu(|\mathbf{T}^d|) \mathbf{T}^d : \mathbf{S} &= \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{S} & \forall \mathbf{S} \in M, \\ b_2(\mathbf{u}, q) &= 0 & \forall q \in Q. \end{aligned} \quad (4.1)$$

Note that, since  $\text{div}(\mathbf{u}) = 0$ , by taking  $\mathbf{S} = \mathbf{I}$  the second line of (4.1) implies that the solution  $\mathbf{T}^d$  of (4.1) satisfies  $\text{tr}(\mathbf{T}^d) = 0$  a.e. in  $\Omega$ , even though this condition was not explicitly imposed on elements of  $M$ .

Let  $h > 0$  be a discretisation parameter that will tend to zero and, for each  $h$ , let  $V_h \subset V$ ,  $Q_h \subset Q$  and  $M_h \subset M$  be three finite-dimensional spaces satisfying the following basic approximation properties, for all  $\mathbf{S} \in M$ ,  $\mathbf{v} \in V$  and  $q \in Q$ :

$$\lim_{h \rightarrow 0} \inf_{\mathbf{S}_h \in M_h} \|\mathbf{S}_h - \mathbf{S}\|_{L^2(\Omega)} = 0, \quad \lim_{h \rightarrow 0} \inf_{\mathbf{v}_h \in V_h} \|\mathbf{D}(\mathbf{v}_h - \mathbf{v})\|_{L^2(\Omega)} = 0, \quad \lim_{h \rightarrow 0} \inf_{q_h \in Q_h} \|q_h - q\|_{L^2(\Omega)} = 0.$$

Moreover, let

$$V_{h,0} := \{\mathbf{v}_h \in V_h : b_2(\mathbf{v}_h, q_h) = 0 \quad \forall q_h \in Q_h\}. \quad (4.2)$$

We assume on the one hand that the pair  $(V_h, Q_h)$  is uniformly stable for the divergence, i.e.,

$$\inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in V_h} \frac{b_2(\mathbf{v}_h, q_h)}{\|\mathbf{D}(\mathbf{v}_h)\|_{L^2(\Omega)} \|q_h\|_{L^2(\Omega)}} \geq \beta^* \quad (4.3)$$

for some constant  $\beta^* > 0$ , independent of  $h$ , and on the other hand that  $M_h$  and  $V_{h,0}$  are compatible in the sense that

$$\mathbf{D}(\mathbf{v}_h) \in M_h \quad \forall \mathbf{v}_h \in V_{h,0}. \quad (4.4)$$

Note that the latter assumption may be prohibitive when considering conforming finite elements on quadrilateral ( $d = 2$ ) or hexahedral ( $d = 3$ ) meshes, see Subsection 4.2; this motivates the study of non-conforming finite elements considered in Section 5. The inf-sup condition (4.3) guarantees that

$$\lim_{h \rightarrow 0} \inf_{\mathbf{v}_h \in V_{h,0}} \|\mathbf{D}(\mathbf{v}_h - \mathbf{u})\|_{L^2(\Omega)} = 0. \quad (4.5)$$

Indeed, (4.3) implies the relation

$$\inf_{\mathbf{v}_{h,0} \in V_{h,0}} \|\mathbf{D}(\mathbf{u} - \mathbf{v}_{h,0})\|_{L^2(\Omega)} \leq \left(1 + \frac{c_b}{\beta^*}\right) \inf_{\mathbf{v}_h \in V_h} \|\mathbf{D}(\mathbf{u} - \mathbf{v}_h)\|_{L^2(\Omega)}, \quad (4.6)$$

which can be shown using a standard argument; see for instance [20]. Here,  $c_b$  denotes the continuity constant of  $b_2(\cdot, \cdot)$  on  $V \times Q$ .

As the divergence of functions of  $V_{h,0}$  is not necessarily zero, the antisymmetry property (3.13) does not hold in the discrete spaces. Since this property is a crucial ingredient in the analysis of our problem, it is standard (see for instance [41, 20]) to introduce the trilinear form  $d : V \times V \times V \rightarrow \mathbb{R}$  defined by

$$d(\mathbf{u}; \mathbf{v}, \mathbf{w}) := \frac{1}{2} \int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{v}] \cdot \mathbf{w} - \frac{1}{2} \int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{w}] \cdot \mathbf{v}. \quad (4.7)$$

The trilinear form  $d$  is obviously antisymmetric and it is consistent thanks to the fact that

$$d(\mathbf{u}; \mathbf{v}, \mathbf{w}) = \int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{v}] \cdot \mathbf{w} \quad \forall \mathbf{u} \in \mathcal{V}, \quad \forall \mathbf{v}, \mathbf{w} \in V.$$

Moreover, a standard computation shows that there exists a constant  $\hat{D} \leq \min(C_S^2, C_3 C_6) C_K^3$  such that

$$d(\mathbf{u}; \mathbf{v}, \mathbf{w}) \leq \hat{D} \|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{v})\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{w})\|_{L^2(\Omega)} \quad \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in V. \quad (4.8)$$



We then consider the following approximation of problem (4.1): find  $(\mathbf{T}_h, \mathbf{u}_h, p_h) \in M_h \times V_h \times Q_h$  such that

$$\begin{aligned} d(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + \int_{\Omega} \mathbf{T}_h : \mathbf{D}(\mathbf{v}_h) + b_2(\mathbf{v}_h, p_h) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h & \forall \mathbf{v}_h \in V_h, \\ \alpha \int_{\Omega} \mathbf{T}_h : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\mathbf{T}_h|) \mathbf{T}_h : \mathbf{S}_h &= \int_{\Omega} \mathbf{D}(\mathbf{u}_h) : \mathbf{S}_h & \forall \mathbf{S}_h \in M_h, \\ b_2(\mathbf{u}_h, q_h) &= 0 & \forall q_h \in Q_h. \end{aligned} \quad (4.9)$$

#### 4.1.1 Existence of a discrete solution

Existence of a solution to problem (4.9) without restrictions on the data is established by Brouwer's fixed point theorem, as in Section 3.3. To begin with, for any function  $\mathbf{v} \in V$ , we define the discrete analogue of the mapping  $\mathcal{G}$ , see (3.8); namely,  $\mathcal{G}_h(\mathbf{v}) \in M_h$  is the unique solution of

$$\alpha \int_{\Omega} \mathcal{G}_h(\mathbf{v}) : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\mathcal{G}_h(\mathbf{v})|) \mathcal{G}_h(\mathbf{v}) : \mathbf{S}_h = \int_{\Omega} \mathbf{D}(\mathbf{v}) : \mathbf{S}_h \quad \forall \mathbf{S}_h \in M_h. \quad (4.10)$$

This finite-dimensional square system has one and only one solution  $\mathcal{G}_h(\mathbf{v})$  thanks to the properties of the left-hand side: the first term is elliptic and the second term is monotone. As in Section 3.3, in view of the inf-sup condition (4.3), problem (4.9) is equivalent to finding  $\mathbf{u}_h \in V_{h,0}$  solution of

$$d(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + \int_{\Omega} \mathbf{T}_h : \mathbf{D}(\mathbf{v}_h) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h \quad \forall \mathbf{v}_h \in V_{h,0}, \quad (4.11)$$

where  $\mathbf{T}_h := \mathcal{G}_h(\mathbf{u}_h)$ . By proceeding as in Proposition 1, it is easy to prove that problem (4.11) has at least one solution  $\mathbf{u}_h \in V_{h,0}$ , and by the above equivalence, each solution  $\mathbf{u}_h$  determines a pair  $(\mathbf{T}_h, p_h) \in M_h \times Q_h$  so that  $(\mathbf{T}_h, \mathbf{u}_h, p_h)$  solves problem (4.9). Moreover, each solution of problem (4.9) satisfies the same estimates as in (3.10) and (3.11). For the sake of simplicity, since the approximation is conforming, we state them in terms of the norm of  $\mathbf{f}$  in  $H^{-1}(\Omega)^d$ ,

$$\|\mathbf{D}(\mathbf{u}_h)\|_{L^2(\Omega)} \leq \alpha \|\mathbf{f}\|_{H^{-1}(\Omega)} + \gamma C_1 |\Omega|^{\frac{1}{2}} \quad (4.12)$$

and

$$\|\mathbf{T}_h\|_{L^2(\Omega)} \leq \|\mathbf{f}\|_{H^{-1}(\Omega)} + \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}}. \quad (4.13)$$

Regarding the other *a priori* bounds, (3.20) and (3.15) are satisfied by  $\mathbf{u}_h$  and  $\mathbf{T}_h$  and, if (3.34) holds, so are (3.41) and (3.40), all up to the above norm for  $\mathbf{f}$ . In contrast, however, we do not have enough information to claim that (3.14) is valid because it relies on the nonnegativity of  $\mathbf{T}_h : \mathbf{D}(\mathbf{u}_h)$  almost everywhere in  $\Omega$ ; the integral average is positive but this does not always guarantee pointwise nonnegativity. Thus we replace the constant  $C_{\mathbf{u}}$  of (3.42) by the constant  $\widetilde{C}_{\mathbf{u}}$  in the following inequality:

$$\|\mathbf{D}(\mathbf{u}_h)\|_{L^2(\Omega)} \leq \widetilde{C}_{\mathbf{u}} := \min \left\{ \frac{1}{\alpha} (\alpha + \gamma \mu_{\max})^2 \|\mathbf{f}\|_{H^{-1}(\Omega)}, \alpha \|\mathbf{f}\|_{H^{-1}(\Omega)} + \gamma C_1 |\Omega|^{\frac{1}{2}}, \frac{1}{\alpha} (\alpha + \gamma \Lambda)^2 \|\mathbf{f}\|_{H^{-1}(\Omega)} \right\}, \quad (4.14)$$

where the last term is included when (3.34) holds. Because  $C_{\mathbf{u}} \leq \widetilde{C}_{\mathbf{u}}$ , we shall use  $\widetilde{C}_{\mathbf{u}}$  to bound both  $\mathbf{u}$  and  $\mathbf{u}_h$  in order to simplify the constants in the computations that will now follow.

Finally, let us establish the convergence of the sequence of discrete solutions in the limit of  $h \rightarrow 0$ . The above uniform *a priori* estimates imply that, up to a subsequence of the discretisation parameter  $h$ ,

$$\begin{aligned} \lim_{h \rightarrow 0} \mathbf{u}_h &= \bar{\mathbf{u}} \quad \text{weakly in } H_0^1(\Omega)^d, \\ \lim_{h \rightarrow 0} \mathbf{u}_h &= \bar{\mathbf{u}} \quad \text{strongly in } L^q(\Omega)^d \text{ with } 1 \leq q < \infty \text{ if } d = 2, \text{ and } 1 \leq q < 6 \text{ if } d = 3, \\ \lim_{h \rightarrow 0} \mathbf{T}_h &= \bar{\mathbf{T}} \quad \text{weakly in } L^2(\Omega)^{d \times d}, \end{aligned}$$

for some  $\bar{\mathbf{u}} \in H_0^1(\Omega)^d$  and  $\bar{\mathbf{T}} \in L^2(\Omega)^{d \times d}$ . Clearly, the symmetry of  $\mathbf{T}_h$  implies that of  $\bar{\mathbf{T}}$  and  $\text{div}(\bar{\mathbf{u}}) = 0$  follows from the fact that  $\mathbf{u}_h$  belongs to  $V_{h,0}$ . Then the approximation properties of the discrete spaces and (4.5) permit to replicate the steps of the proof of Lemma 4 and yield

$$\lim_{h \rightarrow 0} \int_{\Omega} \mathbf{T}_h : \mathbf{D}(\mathbf{u}_h) = \int_{\Omega} \bar{\mathbf{T}} : \mathbf{D}(\bar{\mathbf{u}}). \quad (4.15)$$

To fully identify the limit, in addition to  $\tilde{\mathbf{T}}^{\mathbf{d}} := \mathcal{G}(\bar{\mathbf{u}})$ , which has trace zero since  $\text{div}(\bar{\mathbf{u}}) = 0$ , we introduce the auxiliary tensor  $\tilde{\mathbf{T}}_h := \mathcal{G}_h(\bar{\mathbf{u}})$ . On the one hand

$$\alpha \int_{\Omega} (\tilde{\mathbf{T}}_h - \tilde{\mathbf{T}}^{\mathbf{d}}) : \mathbf{S}_h + \gamma \int_{\Omega} (\mu(|\tilde{\mathbf{T}}_h|)\tilde{\mathbf{T}}_h - \mu(|\tilde{\mathbf{T}}^{\mathbf{d}}|)\tilde{\mathbf{T}}^{\mathbf{d}}) : \mathbf{S}_h = 0 \quad \forall \mathbf{S}_h \in M_h,$$

thus implying that, for all  $\mathbf{S}_h$  in  $M_h$ ,

$$\begin{aligned} & \alpha \|\tilde{\mathbf{T}}_h - \tilde{\mathbf{T}}^{\mathbf{d}}\|_{L^2(\Omega)}^2 + \gamma \int_{\Omega} (\mu(|\tilde{\mathbf{T}}_h|)\tilde{\mathbf{T}}_h - \mu(|\tilde{\mathbf{T}}^{\mathbf{d}}|)\tilde{\mathbf{T}}^{\mathbf{d}}) : (\tilde{\mathbf{T}}_h - \tilde{\mathbf{T}}^{\mathbf{d}}) \\ &= \alpha \int_{\Omega} (\tilde{\mathbf{T}}_h - \tilde{\mathbf{T}}^{\mathbf{d}}) : (\mathbf{S}_h - \tilde{\mathbf{T}}^{\mathbf{d}}) + \gamma \int_{\Omega} (\mu(|\tilde{\mathbf{T}}_h|)\tilde{\mathbf{T}}_h - \mu(|\tilde{\mathbf{T}}^{\mathbf{d}}|)\tilde{\mathbf{T}}^{\mathbf{d}}) : (\mathbf{S}_h - \tilde{\mathbf{T}}^{\mathbf{d}}). \end{aligned}$$

Since both  $\tilde{\mathbf{T}}_h$  and  $\tilde{\mathbf{T}}^{\mathbf{d}}$  are bounded in  $M$  uniformly with respect to  $h$ , and

$$\|\mu(|\tilde{\mathbf{T}}_h|)\tilde{\mathbf{T}}_h - \mu(|\tilde{\mathbf{T}}^{\mathbf{d}}|)\tilde{\mathbf{T}}^{\mathbf{d}}\|_{L^2(\Omega)} \leq 2C_1|\Omega|^{\frac{1}{2}},$$

again a uniform bound, then the approximation properties of  $M_h$  and the monotonicity property (2.14) imply that

$$\lim_{h \rightarrow 0} \|\tilde{\mathbf{T}}_h - \tilde{\mathbf{T}}^{\mathbf{d}}\|_{L^2(\Omega)} = 0. \quad (4.16)$$

On the other hand, the auxiliary tensor  $\tilde{\mathbf{T}}_h$  permits us to argue as in the proof of Lemma 5. Indeed, the monotonicity property (2.14) yields

$$\begin{aligned} \alpha \|\mathbf{T}_h - \tilde{\mathbf{T}}_h\|_{L^2(\Omega)}^2 &\leq \int_{\Omega} \mathbf{D}(\mathbf{u}_h) : (\mathbf{T}_h - \tilde{\mathbf{T}}_h) - \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : (\mathbf{T}_h - \tilde{\mathbf{T}}_h) \\ &= \int_{\Omega} \mathbf{D}(\mathbf{u}_h) : \mathbf{T}_h - \int_{\Omega} \mathbf{D}(\mathbf{u}_h) : \tilde{\mathbf{T}}_h - \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : (\mathbf{T}_h - \tilde{\mathbf{T}}_h). \end{aligned}$$

From (4.15) and (4.16), we easily derive that the above right-hand side tends to zero. Hence

$$\lim_{h \rightarrow 0} \|\mathbf{T}_h - \tilde{\mathbf{T}}_h\|_{L^2(\Omega)} = 0,$$

and then combining this with (4.16) we infer that

$$\lim_{h \rightarrow 0} \|\mathbf{T}_h - \tilde{\mathbf{T}}^{\mathbf{d}}\|_{L^2(\Omega)} = 0. \quad (4.17)$$

Hence uniqueness of the limit implies that  $\bar{\mathbf{T}} = \tilde{\mathbf{T}}^{\mathbf{d}} = \mathcal{G}(\bar{\mathbf{u}})$ . This, and (4.3), permit to identify the limit as in Lemma 5 and Theorem 1 and proves convergence to a weak solution without restrictions on the data. Thus we have proved the following result.

**Theorem 2.** (Convergence for all data) Under the above approximation properties and compatibility of the discrete spaces, up to a subsequence,

$$\begin{aligned} \lim_{h \rightarrow 0} \mathbf{u}_h &= \mathbf{u} && \text{weakly in } H_0^1(\Omega)^d, \\ \lim_{h \rightarrow 0} \mathbf{u}_h &= \mathbf{u} && \text{strongly in } L^q(\Omega)^d \text{ with } 1 \leq q < \infty \text{ if } d = 2, \text{ and } 1 \leq q < 6 \text{ if } d = 3, \\ \lim_{h \rightarrow 0} \mathbf{T}_h &= \mathbf{T}^{\mathbf{d}} && \text{strongly in } L^2(\Omega)^{d \times d}, \\ \lim_{h \rightarrow 0} p_h &= p && \text{weakly in } L^2(\Omega), \end{aligned}$$

where  $(\mathbf{T}^{\mathbf{d}}, \mathbf{u}, p)$  is a solution of (3.6), (3.7).

#### 4.1.2 Error estimate

We now prove an *a priori* error estimate between  $(\mathbf{T}^d, \mathbf{u}, p)$  and  $(\mathbf{T}_h, \mathbf{u}_h, p_h)$ , under the assumption (3.34) that has not been used so far, and the small data condition (4.18) below. Note that this small data condition is in fact the same as the uniqueness condition (3.35), upon replacing  $C_{\mathbf{u}}$  by  $\widetilde{C}_{\mathbf{u}}$ . To simplify the notation and compress some of the long displayed lines of mathematics, we shall write  $\|\cdot\|_V$ ,  $\|\cdot\|_M$  and  $\|\cdot\|_Q$  instead of  $\|D(\cdot)\|_{L^2(\Omega)}$  (as a norm on  $V$ ),  $\|\cdot\|_{L^2(\Omega)}$  (as a norm on  $M$ ) and  $\|\cdot\|_{L^2(\Omega)}$  (as a norm on  $Q$ ), respectively.

**Theorem 3.** *In addition to (3.34), let the input data satisfy*

$$\frac{\gamma}{\alpha}\Lambda + \alpha\hat{D}\widetilde{C}_{\mathbf{u}} \leq \theta < 1, \quad (4.18)$$

where  $0 < \theta < 1$  and  $\hat{D}$  is the constant from (4.8). Then, there exists a constant  $C > 0$  independent of  $h$  such that the difference between the solution  $(\mathbf{T}_h, \mathbf{u}_h, p_h)$  of (4.9) and  $(\mathbf{T}^d, \mathbf{u}, p)$  of (4.1) satisfies

$$\|\mathbf{u} - \mathbf{u}_h\|_V + \|\mathbf{T}^d - \mathbf{T}_h\|_M + \|p - p_h\|_Q \leq C \left[ \inf_{\mathbf{v}_h \in V_h} \|\mathbf{u} - \mathbf{v}_h\|_V + \inf_{\mathbf{S}_h \in M_h} \|\mathbf{T}^d - \mathbf{S}_h\|_M + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right]. \quad (4.19)$$

*Proof.* Since we are using conforming finite element spaces, taking  $(\mathbf{S}, \mathbf{v}, q) = (\mathbf{S}_h, \mathbf{v}_h, q_h)$  in (4.1) and subtracting the equations of (4.9) we easily get

$$\begin{aligned} d(\mathbf{u}; \mathbf{u}, \mathbf{v}_h) - d(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + \int_{\Omega} (\mathbf{T}^d - \mathbf{T}_h) : \mathbf{D}(\mathbf{v}_h) + b_2(\mathbf{v}_h, p - p_h) &= 0 \quad \forall \mathbf{v}_h \in V_h, \\ \alpha \int_{\Omega} (\mathbf{T}^d - \mathbf{T}_h) : \mathbf{S}_h + \gamma \int_{\Omega} (\mu(|\mathbf{T}^d|)\mathbf{T}^d - \mu(|\mathbf{T}_h|)\mathbf{T}_h) : \mathbf{S}_h &= \int_{\Omega} \mathbf{D}(\mathbf{u} - \mathbf{u}_h) : \mathbf{S}_h \quad \forall \mathbf{S}_h \in M_h, \\ b_2(\mathbf{u} - \mathbf{u}_h, q_h) &= 0 \quad \forall q_h \in Q_h. \end{aligned} \quad (4.20)$$

The rest of the proof is divided into three steps.

**Step 1** (Error bound for the pressure). By the triangle inequality we have, for any  $q_h \in Q_h$ ,

$$\|p - p_h\|_Q \leq \|p - q_h\|_Q + \|q_h - p_h\|_Q,$$

and it therefore suffices to derive a bound on  $\|q_h - p_h\|_Q$ . From the (discrete) inf-sup condition we have

$$\beta^* \|p_h - q_h\|_Q \leq \sup_{\mathbf{v}_h \in V_h} \frac{b_2(\mathbf{v}_h, p_h - q_h)}{\|\mathbf{v}_h\|_V}.$$

Again, using the first equation of (4.20) we have

$$\begin{aligned} b_2(\mathbf{v}_h, p_h - q_h) &= b_2(\mathbf{v}_h, p_h - p) + b_2(\mathbf{v}_h, p - q_h) \\ &= d(\mathbf{u}; \mathbf{u}, \mathbf{v}_h) - d(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + \int_{\Omega} (\mathbf{T}^d - \mathbf{T}_h) : \mathbf{D}(\mathbf{v}_h) + b_2(\mathbf{v}_h, p - q_h) \\ &\leq \left[ (\hat{D}\|\mathbf{u}\|_V + \hat{D}\|\mathbf{u}_h\|_V) \|\mathbf{u} - \mathbf{u}_h\|_V + \|\mathbf{T}^d - \mathbf{T}_h\|_M + c_b \|p - q_h\|_Q \right] \|\mathbf{v}_h\|_V \\ &\leq \left[ 2\hat{D}\widetilde{C}_{\mathbf{u}} \|\mathbf{u} - \mathbf{u}_h\|_V + \|\mathbf{T}^d - \mathbf{T}_h\|_M + c_b \|p - q_h\|_Q \right] \|\mathbf{v}_h\|_V, \end{aligned}$$

where we can take  $c_b = C_K$  using the relation  $\|\operatorname{div}(\mathbf{v})\|_{L^2(\Omega)}^2 + \|\operatorname{rot}(\mathbf{v})\|_{L^2(\Omega)}^2 = \|\nabla \mathbf{v}\|_{L^2(\Omega)}^2$  that holds because we have homogeneous Dirichlet boundary conditions (otherwise take  $c_b = \sqrt{d}C_K$ ). Thus, we obtain

$$\|p - p_h\|_Q \leq \frac{2\hat{D}\widetilde{C}_{\mathbf{u}}}{\beta^*} \|\mathbf{u} - \mathbf{u}_h\|_V + \frac{1}{\beta^*} \|\mathbf{T}^d - \mathbf{T}_h\|_M + \left(1 + \frac{c_b}{\beta^*}\right) \|p - q_h\|_Q \quad (4.21)$$

for any  $q_h \in Q_h$ .

**Step 2** (Error bound for the stress tensor). Again, we start with the triangle inequality; for any  $\mathbf{S}_h \in M_h$  we have that

$$\|\mathbf{T}^d - \mathbf{T}_h\|_M \leq \|\mathbf{T}^d - \mathbf{S}_h\|_M + \|\mathbf{S}_h - \mathbf{T}_h\|_M,$$

and we then bound  $\|\mathbf{T}_h - \mathbf{S}_h\|_M$ . Thanks to the monotonicity property (2.14) and the second equation of (4.20), we have

$$\begin{aligned}
\alpha \|\mathbf{T}_h - \mathbf{S}_h\|_M^2 &\leq \alpha \int_{\Omega} |\mathbf{T}_h - \mathbf{S}_h|^2 + \gamma \int_{\Omega} (\mu(|\mathbf{T}_h|)\mathbf{T}_h - \mu(|\mathbf{S}_h|)\mathbf{S}_h) : (\mathbf{T}_h - \mathbf{S}_h) \\
&= \alpha \int_{\Omega} (\mathbf{T}_h - \mathbf{T}^d + \mathbf{T}^d - \mathbf{S}_h) : (\mathbf{T}_h - \mathbf{S}_h) + \gamma \int_{\Omega} (\mu(|\mathbf{T}_h|)\mathbf{T}_h - \mu(|\mathbf{S}_h|)\mathbf{S}_h) : (\mathbf{T}_h - \mathbf{S}_h) \\
&= \int_{\Omega} \mathbf{D}(\mathbf{u}_h - \mathbf{u}) : (\mathbf{T}_h - \mathbf{S}_h) + \gamma \int_{\Omega} (\mu(|\mathbf{T}^d|)\mathbf{T}^d - \mu(|\mathbf{S}_h|)\mathbf{S}_h) : (\mathbf{T}_h - \mathbf{S}_h) + \alpha \int_{\Omega} (\mathbf{T}^d - \mathbf{S}_h) : (\mathbf{T}_h - \mathbf{S}_h) \\
&\leq \left[ \|\mathbf{u} - \mathbf{u}_h\|_V + \alpha \|\mathbf{T}^d - \mathbf{S}_h\|_M + \gamma \Lambda \|\mathbf{T}^d - \mathbf{S}_h\|_M \right] \|\mathbf{T}_h - \mathbf{S}_h\|_M,
\end{aligned}$$

and thus

$$\|\mathbf{T}^d - \mathbf{T}_h\|_M \leq \frac{1}{\alpha} \|\mathbf{u} - \mathbf{u}_h\|_V + \left(2 + \frac{\gamma \Lambda}{\alpha}\right) \|\mathbf{T}^d - \mathbf{S}_h\|_M \quad (4.22)$$

for any  $\mathbf{S}_h \in M_h$ .

**Step 3** (Error bound for the velocity). Recalling the definition of  $V_{h,0}$  in (4.2), let  $\mathbf{v}_{h,0} \in V_{h,0}$  and let  $\mathbf{v}_h := \mathbf{v}_{h,0} - \mathbf{u}_h \in V_{h,0}$ . We will first show the relation (4.19) by taking the infimum over  $V_{h,0}$  instead of  $V_h$ . As before, we use the triangle inequality to get

$$\|\mathbf{u} - \mathbf{u}_h\|_V \leq \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V.$$

Thanks to the assumption (4.4), we can take  $\mathbf{S}_h = \mathbf{D}(\mathbf{v}_h)$  in the second equation of (4.20) yielding

$$\int_{\Omega} \mathbf{D}(\mathbf{u} - \mathbf{u}_h) : \mathbf{D}(\mathbf{v}_h) = \alpha \int_{\Omega} (\mathbf{T}^d - \mathbf{T}_h) : \mathbf{D}(\mathbf{v}_h) + \gamma \int_{\Omega} (\mu(|\mathbf{T}^d|)\mathbf{T}^d - \mu(|\mathbf{T}_h|)\mathbf{T}_h) : \mathbf{D}(\mathbf{v}_h).$$

Using the first equation of (4.20), we can easily derive the equality

$$\begin{aligned}
\|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V^2 &= \int_{\Omega} \mathbf{D}(\mathbf{v}_{h,0} - \mathbf{u}_h) : \mathbf{D}(\mathbf{v}_h) = \int_{\Omega} \mathbf{D}(\mathbf{v}_{h,0} - \mathbf{u}) : \mathbf{D}(\mathbf{v}_h) - \alpha(d(\mathbf{u}; \mathbf{u}, \mathbf{v}_h) - d(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h)) \\
&\quad - \alpha b_2(\mathbf{v}_h, p - q_h) + \gamma \int_{\Omega} (\mu(|\mathbf{T}^d|)\mathbf{T}^d - \mu(|\mathbf{T}_h|)\mathbf{T}_h) : \mathbf{D}(\mathbf{v}_h),
\end{aligned}$$

thanks to the fact that  $b_2(\mathbf{v}_h, q_h - p_h) = 0$ . To bound the convective term, we use

$$\begin{aligned}
d(\mathbf{u}; \mathbf{u}, \mathbf{v}_h) - d(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) &= d(\mathbf{u} - \mathbf{u}_h; \mathbf{u}, \mathbf{v}_h) + d(\mathbf{u}_h; \mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) \\
&= d(\mathbf{u} - \mathbf{v}_{h,0}; \mathbf{u}, \mathbf{v}_h) + d(\mathbf{v}_{h,0} - \mathbf{u}_h; \mathbf{u}, \mathbf{v}_h) \\
&\quad + d(\mathbf{u}_h; \mathbf{u} - \mathbf{v}_{h,0}, \mathbf{v}_h) + \underbrace{d(\mathbf{u}_h; \mathbf{v}_{h,0} - \mathbf{u}_h, \mathbf{v}_h)}_{=0} \\
&\leq \left[ \hat{D}\|\mathbf{u}\|_V \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + \hat{D}\|\mathbf{u}\|_V \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V + \hat{D}\|\mathbf{u}_h\|_V \|\mathbf{u} - \mathbf{v}_{h,0}\|_V \right] \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V \\
&\leq \left[ 2\hat{D}\widetilde{C}_{\mathbf{u}} \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + \hat{D}\widetilde{C}_{\mathbf{u}} \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V \right] \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V,
\end{aligned}$$

from which we get

$$\begin{aligned}
\|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V &\leq \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + 2\alpha\hat{D}\widetilde{C}_{\mathbf{u}} \|\mathbf{u} - \mathbf{v}_{h,0}\|_V \\
&\quad + \alpha c_b \|p - q_h\|_Q + \gamma \Lambda \|\mathbf{T}^d - \mathbf{T}_h\|_M + \alpha\hat{D}\widetilde{C}_{\mathbf{u}} \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V.
\end{aligned}$$

Now using (4.22) we arrive at

$$\begin{aligned}
\|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V &\leq (1 + 2\alpha\hat{D}\widetilde{C}_{\mathbf{u}}) \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + \alpha c_b \|p - q_h\|_Q + \frac{\gamma \Lambda}{\alpha} \|\mathbf{u} - \mathbf{u}_h\|_V \\
&\quad + \gamma \Lambda \left(2 + \frac{\gamma \Lambda}{\alpha}\right) \|\mathbf{T}^d - \mathbf{S}_h\|_M + \alpha\hat{D}\widetilde{C}_{\mathbf{u}} \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V \\
&\leq \left(1 + 2\alpha\hat{D}\widetilde{C}_{\mathbf{u}} + \frac{\gamma \Lambda}{\alpha}\right) \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + \gamma \Lambda \left(2 + \frac{\gamma \Lambda}{\alpha}\right) \|\mathbf{T}^d - \mathbf{S}_h\|_M \\
&\quad + \alpha c_b \|p - q_h\|_Q + \left(\frac{\gamma \Lambda}{\alpha} + \alpha\hat{D}\widetilde{C}_{\mathbf{u}}\right) \|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V.
\end{aligned}$$

Therefore, using the assumption (4.18) on the input data, we obtain

$$\|\mathbf{v}_{h,0} - \mathbf{u}_h\|_V \leq \frac{1}{1-\theta} \left[ \left( 1 + 2\alpha\hat{D}\widetilde{C}_{\mathbf{u}} + \frac{\gamma\Lambda}{\alpha} \right) \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + \gamma\Lambda \left( 2 + \frac{\gamma\Lambda}{\alpha} \right) \|\mathbf{T}^{\mathbf{d}} - \mathbf{S}_h\|_M + \alpha c_b \|p - q_h\|_Q \right]$$

and thus

$$\|\mathbf{u} - \mathbf{u}_h\|_V \leq \left( 1 + \frac{1 + 2\alpha\hat{D}\widetilde{C}_{\mathbf{u}} + \frac{\gamma\Lambda}{\alpha}}{1-\theta} \right) \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + \frac{\gamma\Lambda \left( 2 + \frac{\gamma\Lambda}{\alpha} \right)}{1-\theta} \|\mathbf{T}^{\mathbf{d}} - \mathbf{S}_h\|_M + \frac{\alpha c_b}{1-\theta} \|p - q_h\|_Q \quad (4.23)$$

for any  $\mathbf{v}_{h,0} \in V_{h,0}$ . Finally, combining (4.21), (4.22) and (4.23) we obtain

$$\|\mathbf{u} - \mathbf{u}_h\|_V + \|\mathbf{T}^{\mathbf{d}} - \mathbf{T}_h\|_M + \|p - p_h\|_Q \leq c_1 \inf_{\mathbf{v}_{h,0} \in V_{h,0}} \|\mathbf{u} - \mathbf{v}_{h,0}\|_V + c_2 \inf_{\mathbf{S}_h \in M_h} \|\mathbf{T}^{\mathbf{d}} - \mathbf{S}_h\|_M + c_3 \inf_{q_h \in Q_h} \|p - q_h\|_Q$$

with

$$\begin{aligned} c_1 &:= \left[ 1 + \frac{2\hat{D}\widetilde{C}_{\mathbf{u}}}{\beta^*} + \frac{1}{\alpha} \left( 1 + \frac{1}{\beta^*} \right) \right] \left( 1 + \frac{1 + 2\alpha\hat{D}\widetilde{C}_{\mathbf{u}} + \frac{\gamma\Lambda}{\alpha}}{1-\theta} \right), \\ c_2 &:= \left( 1 + \frac{1}{\beta^*} \right) \left( 2 + \frac{\gamma\Lambda}{\alpha} \right) + \left[ 1 + \frac{2\hat{D}\widetilde{C}_{\mathbf{u}}}{\beta^*} + \frac{1}{\alpha} \left( 1 + \frac{1}{\beta^*} \right) \right] \frac{\gamma\Lambda \left( 2 + \frac{\gamma\Lambda}{\alpha} \right)}{1-\theta}, \\ c_3 &:= \left( 1 + \frac{c_b}{\beta^*} \right) + \left[ 1 + \frac{2\hat{D}\widetilde{C}_{\mathbf{u}}}{\beta^*} + \frac{1}{\alpha} \left( 1 + \frac{1}{\beta^*} \right) \right] \left( \frac{\alpha c_b}{1-\theta} \right). \end{aligned}$$

We can then conclude the proof using (4.6).  $\square$

## 4.2 Examples of conforming approximation

From now on, we assume that the boundary of the Lipschitz domain  $\Omega \subset \mathbb{R}^d$  is a polygonal line (when  $d = 2$ ) or a polyhedral surface (when  $d = 3$ ), so that it can be exactly meshed. For each  $h$ , let  $\mathcal{T}_h$  be a conforming mesh on  $\overline{\Omega}$  consisting of elements  $E$ , triangles or quadrilaterals in two dimensions, tetrahedra or hexahedra (all planar-faced) in three dimensions, conforming in the sense that the mesh has no hanging nodes. As usual, the diameter of  $E$  is denoted by  $h_E$ ,

$$h = \sup_{E \in \mathcal{T}_h} h_E,$$

and  $\varrho_E$  is the diameter of the largest ball inscribed in  $E$ .

### 4.2.1 The simplicial case

In the case of simplices, the family of meshes  $\mathcal{T}_h$  is assumed to be *regular* in the sense of Ciarlet [14]: i.e., it is assumed that there exists a constant  $\sigma > 0$ , independent of  $h$ , such that

$$\frac{h_E}{\varrho_E} \leq \sigma \quad \forall E \in \mathcal{T}_h. \quad (4.24)$$

This condition guarantees that there is an invertible affine mapping  $\mathcal{F}_E$  that maps the unit reference simplex onto  $E$ .

For any integer  $k \geq 0$ , let  $\mathbb{P}_k$  denote the space of polynomials in  $d$  variables of degree at most  $k$ . In each element  $E$ , the functions will be approximated in the spaces  $\mathbb{P}_k$ . The specific choice of finite element spaces is dictated by two considerations. First, conditions (4.3) and (4.4) must be satisfied. Next, since the number of unknowns in (4.9) is large, the degree  $k$  of the finite element functions should be small. It is well-known that the lowest degree of conforming approximation of  $(\mathbf{u}, p)$  satisfying (4.3), without modification of the bilinear forms, is the Taylor-Hood  $\mathbb{P}_2^d - \mathbb{P}_1$  element, see [20, 3], provided each element has at least one interior vertex.

In view of (4.4), this implies that  $\mathbf{T}^d$  is approximated by  $\mathbb{P}_1^{d \times d}$ . Thus the corresponding finite element spaces are

$$\begin{aligned} V_h &:= \{\mathbf{v}_h \in H_0^1(\Omega)^d : \mathbf{v}_h|_E \in \mathbb{P}_2^d \quad \forall E \in \mathcal{T}_h\}, \\ Q_h &:= \{q_h \in H^1(\Omega) \cap L_0^2(\Omega) : q_h|_E \in \mathbb{P}_1 \quad \forall E \in \mathcal{T}_h\}, \\ M_h &:= \{\mathbf{S}_h \in L^2(\Omega)_{\text{sym}}^{d \times d} : \mathbf{S}_h|_E \in (\mathbb{P}_1)_{\text{sym}}^{d \times d} \quad \forall E \in \mathcal{T}_h\}. \end{aligned}$$

It is easy to check that with these spaces on a simplicial mesh, under condition (4.24), problem (4.9) has at least one solution. Furthermore, if the data satisfy (4.18), then Theorem 3 yields

$$\|\mathbf{D}(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} + \|\mathbf{T}^d - \mathbf{T}_h\|_{L^2(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq C h^2, \quad (4.25)$$

provided that the solution is sufficiently smooth, namely  $\mathbf{u} \in H^3(\Omega)^d \cap H_0^1(\Omega)^d$ ,  $\mathbf{T}^d \in H^2(\Omega)^{d \times d}$ , and  $p \in H^2(\Omega) \cap L_0^2(\Omega)$ . Therefore the scheme has order two for an optimal number of degrees of freedom, i.e., this order of convergence cannot be achieved with fewer degrees of freedom.

#### 4.2.2 The quadrilateral/hexahedral case

The notion of regularity is more complex for quadrilateral and much more complex for hexahedral elements. In the case of quadrilaterals [20], the family of meshes is regular if the elements are convex and, moreover, the subtriangles associated to each vertex (there is one per vertex) all satisfy (4.24). In the case of hexahedra with plane faces, convexity and the validity of (4.24) for the subtetrahedra associated to each vertex are necessary but not sufficient. This difficulty has been investigated by many authors, see for instance [42, 23]; the most relevant publication concerning hexahedra with plane faces is however [22], where the minimum of the Jacobian in the reference cube  $\hat{E}$  is bounded below by the minimum of the coefficients of its Bézier expansion and this minimum is determined by an efficient algorithm. The details of this are beyond the scope of this work, and we shall simply assume here that the minimum of these Bézier coefficients is strictly positive and that furthermore, denoting by  $\mathcal{J}_E$  the Jacobian determinant of  $\mathcal{F}_E$ ,

$$\mathcal{J}_E(\hat{\mathbf{x}}) \geq \hat{c} \varrho_E^3 \quad \forall \hat{\mathbf{x}} \in \hat{E} \quad (4.26)$$

with a constant  $\hat{c}$  independent of  $E$  and  $h$ . If these conditions hold, there is an invertible bi-affine mapping  $\mathcal{F}_E$  in two dimensions or tri-affine in three dimensions that maps the unit reference square or cube onto  $E$ .

We let  $\mathbb{Q}_k$  be the space of polynomials in  $d$  variables of degree at most  $k$  in *each* variable. In contrast to the case of simplicial meshes, the space  $\mathbb{Q}_k$  is not invariant under the composition with  $\mathcal{F}_E$ , which makes the compatibility condition (4.4) between  $\mathbf{D}(V_h)$  and  $M_h$  problematic. To circumvent this issue, we restrict ourselves to affine maps  $\mathcal{F}_E$ , thereby allowing subdivisions consisting of parallelograms/parallelepipeds. In addition, the situation is less satisfactory when a quadrilateral or hexahedral mesh is used, because although the Taylor-Hood  $\mathbb{Q}_2$ – $\mathbb{Q}_1$  element satisfies (4.3), the second condition (4.4) does not hold if  $\mathbf{T}^d$  is approximated by  $\mathbb{Q}_1^{d \times d}$  since the components of the gradient of  $\mathbb{Q}_2$  functions belong to a space, intermediate between  $\mathbb{Q}_2$  and  $\mathbb{Q}_1$ , that is strictly larger than both  $\mathbb{Q}_1$  and  $\mathbb{P}_2$ . Therefore, in order to satisfy (4.4), the simplest option is to discretise each component of  $\mathbf{T}^d$  by  $\mathbb{Q}_2$ . The corresponding finite element spaces are

$$\begin{aligned} V_h &:= \{\mathbf{v}_h \in H_0^1(\Omega)^d : \mathbf{v}_h|_E \in \mathbb{Q}_2^d \quad \forall E \in \mathcal{T}_h\}, \\ Q_h &:= \{q_h \in H^1(\Omega) \cap L_0^2(\Omega) : q_h|_E \in \mathbb{Q}_1 \quad \forall E \in \mathcal{T}_h\}, \\ M_h &:= \{\mathbf{S}_h \in L^2(\Omega)_{\text{sym}}^{d \times d} : \mathbf{S}_h|_E \in (\mathbb{Q}_2)_{\text{sym}}^{d \times d} \quad \forall E \in \mathcal{T}_h\}. \end{aligned}$$

With these spaces and under the above regularity conditions, problem (4.9) has at least one solution and the error estimate (4.25) holds if the data satisfy (4.18). However, this triple of spaces is no longer optimal, because the degree two approximation of  $\mathbf{T}^d$  now requires far too many degrees of freedom with no gain in accuracy. For instance, when  $d = 3$ , its approximation by  $(\mathbb{Q}_2)_{\text{sym}}^{3 \times 3}$  requires  $27 \times 6 = 162$  unknowns inside each element instead of  $8 \times 6 = 48$  unknowns for  $(\mathbb{Q}_1)_{\text{sym}}^{3 \times 3}$ .

The nonconforming finite element approximations discussed in Section 5 do not require an affine mapping  $\mathcal{F}_E$  and, by considering  $\mathbb{P}$ -type approximations on the physical element  $E$ , do not suffer from the computational cost overhead mentioned above.

## 5 Nonconforming finite element approximation

The nonconforming approximations developed here will not only allow the use of elements of degree one for  $\mathbf{u}$ , but will also lead to locally mass-conserving schemes. Because of the discontinuity of the finite element functions, the proofs are in some cases more complex; this is true in particular for the proof of the inf-sup condition for the discrete divergence.

### 5.1 The quadrilateral/planar-faced hexahedral case

Here we consider quadrilateral/hexahedral grids  $\mathcal{T}_h$  with planar faces, satisfying the regularity assumptions stated in Section 4.2. There is a wide choice of possible approximations with nonconforming finite elements. Here we propose globally discontinuous velocities in  $\mathbb{P}_k^d$ ,  $k \geq 1$ , in each cell associated with globally discontinuous pressures and stresses both of degree at most  $k-1$ . Thus we consider  $V_h \subset L^2(\Omega)^d$ ,  $Q_h \subset L_0^2(\Omega)$  and  $M_h \subset L^2(\Omega)^{d \times d}_{\text{sym}}$  defined by

$$V_h := \{\mathbf{v}_h \in L^2(\Omega)^d : \mathbf{v}_h|_E \in \mathbb{P}_k^d \quad \forall E \in \mathcal{T}_h\}, \quad (5.1)$$

$$Q_h := \{q_h \in L_0^2(\Omega) : q_h|_E \in \mathbb{P}_{k-1} \quad \forall E \in \mathcal{T}_h\}, \quad (5.2)$$

$$M_h := \{\mathbf{S}_h \in L^2(\Omega)^{d \times d}_{\text{sym}} : \mathbf{S}_h|_E \in (\mathbb{P}_{k-1})^{d \times d}_{\text{sym}} \quad \forall E \in \mathcal{T}_h\}. \quad (5.3)$$

As usual, the full nonconformity of  $V_h$  is compensated by adding to the forms consistent jumps and averages on edges when  $d = 2$  or faces when  $d = 3$ ; see for instance [38]. Let  $\Gamma_h = \Gamma_h^i \cup \Gamma_h^b$  denote the set of all edges when  $d = 2$  or all faces when  $d = 3$  with  $\Gamma_h^i$  and  $\Gamma_h^b$  signifying the set of all interior and the set of all boundary edges ( $d = 2$ ) or faces ( $d = 3$ ), respectively. A unit normal vector  $\mathbf{n}_e$  is attributed to each  $e \in \Gamma_h$ ; its direction can be freely chosen. Here, the following rule is applied: if  $e \in \Gamma_h^b$ , then  $\mathbf{n}_e = \mathbf{n}_\Omega$ , the exterior unit normal to  $\Omega$ ; if  $e \in \Gamma_h^i$ , then  $\mathbf{n}_e$  points from  $E_i$  to  $E_j$ , where  $E_i$  and  $E_j$  are the two elements of  $\mathcal{T}_h$  adjacent to  $e$  and the number  $i$  of  $E_i$  is smaller than that of  $E_j$ . The jumps and averages of any function  $f$  on  $e$  (smooth enough to have a trace) are defined by

$$[f(x)]_e := f(x)|_{E_i} - f(x)|_{E_j}, \quad \text{when } \mathbf{n}_e \text{ points from } E_i \text{ to } E_j,$$

$$\{f(x)\}_e := \frac{1}{2}(f(x)|_{E_i} + f(x)|_{E_j}).$$

When  $e \in \Gamma_h^b$ , the jump and average are defined to coincide with the trace on  $e$ .

The terms involving jumps and averages that are added to each form are not unique; here we make the following fairly standard choice:

$$\int_\Omega \mathbf{S} : \mathbf{D}(\mathbf{v}) \simeq b_{1h}(\mathbf{S}_h, \mathbf{v}_h) := \sum_{E \in \mathcal{T}_h} \int_E \mathbf{S}_h : \mathbf{D}(\mathbf{v}_h) - \sum_{e \in \Gamma_h} \int_e \{\mathbf{S}_h\}_e \mathbf{n}_e \cdot [\mathbf{v}_h]_e. \quad (5.4)$$

The trilinear form  $d$  is approximated by a centred discretisation, as follows:

$$\begin{aligned} d_h(\mathbf{u}_h; \mathbf{v}_h, \mathbf{w}_h) &:= \sum_{E \in \mathcal{T}_h} \int_E [(\mathbf{u}_h \cdot \nabla) \mathbf{v}_h] \cdot \mathbf{w}_h + \frac{1}{2} \sum_{E \in \mathcal{T}_h} \int_E \text{div}(\mathbf{u}_h) (\mathbf{v}_h \cdot \mathbf{w}_h) \\ &\quad - \frac{1}{2} \sum_{e \in \Gamma_h} \int_e [\mathbf{u}_h]_e \cdot \mathbf{n}_e \{\mathbf{v}_h \cdot \mathbf{w}_h\}_e - \sum_{e \in \Gamma_h^i} \int_e \{\mathbf{u}_h\}_e \cdot \mathbf{n}_e [\mathbf{v}_h]_e \cdot \{\mathbf{w}_h\}_e. \end{aligned} \quad (5.5)$$

The divergence form  $b_2$  is approximated by

$$b_{2h}(\mathbf{v}_h, q_h) := - \sum_{E \in \mathcal{T}_h} \int_E q_h \text{div}(\mathbf{v}_h) + \sum_{e \in \Gamma_h} \int_e [\mathbf{v}_h]_e \cdot \mathbf{n}_e \{q_h\}_e. \quad (5.6)$$

Clearly, the jump terms in (5.4) and (5.6) vanish when  $\mathbf{v}_h$  belongs to  $H_0^1(\Omega)^d$ . Likewise, the jump and divergence terms in (5.5) vanish when  $\mathbf{u}_h$  and  $\mathbf{v}_h$  belong to  $H_0^1(\Omega)^d$  and  $\text{div}(\mathbf{u}_h) = 0$ . Moreover, (5.5) is constructed so that  $d_h$  is antisymmetric,

$$d_h(\mathbf{u}_h; \mathbf{v}_h, \mathbf{w}_h) = -d_h(\mathbf{u}_h; \mathbf{w}_h, \mathbf{v}_h) \quad \forall \mathbf{u}_h, \mathbf{v}_h, \mathbf{w}_h \in V_h. \quad (5.7)$$

Finally, the following positive definite form acts as a penalty to compensate the nonconformity of  $\mathbf{u}_h$ :

$$J_h(\mathbf{u}_h, \mathbf{v}_h) := \sum_{e \in \Gamma_h} \frac{\sigma_e}{h_e} \int_e [\mathbf{u}_h]_e \cdot [\mathbf{v}_h]_e, \quad (5.8)$$

where  $h_e$  is the average of the diameter of the two elements adjacent to  $e$ , if  $e \in \Gamma_h^i$ , or the diameter of the element adjacent to  $e$  otherwise. The parameters  $\sigma_e > 0$  will be chosen below to guarantee stability of the scheme, see (5.28) and (5.24). This form is also used to define the norm on  $V_h$  by

$$\|\mathbf{v}_h\|_{V_h} := \left( \|\mathbf{D}(\mathbf{v}_h)\|_h^2 + J_h(\mathbf{v}_h, \mathbf{v}_h) \right)^{\frac{1}{2}}, \quad (5.9)$$

where

$$\|\mathbf{D}(\mathbf{v}_h)\|_h := \left( \sum_{E \in \mathcal{T}_h} \|\mathbf{D}(\mathbf{v}_h)\|_{L^2(E)}^2 \right)^{\frac{1}{2}} \quad (5.10)$$

denotes the associated semi-norm. Also, in view of (5.6), we define the space of discretely divergence-free functions,

$$V_{h,0} := \{\mathbf{v}_h \in V_h : b_{2h}(\mathbf{v}_h, q_h) = 0 \quad \forall q_h \in Q_h\}. \quad (5.11)$$

The discrete scheme reads: find  $(\mathbf{T}_h, \mathbf{u}_h, p_h) \in M_h \times V_h \times Q_h$  solution of

$$\begin{aligned} d_h(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + b_{1h}(\mathbf{T}_h, \mathbf{v}_h) + b_{2h}(\mathbf{v}_h, p_h) + J_h(\mathbf{u}_h, \mathbf{v}_h) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h & \forall \mathbf{v}_h \in V_h, \\ \alpha \int_{\Omega} \mathbf{T}_h : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\mathbf{T}_h|) \mathbf{T}_h : \mathbf{S}_h &= b_{1h}(\mathbf{S}_h, \mathbf{u}_h) & \forall \mathbf{S}_h \in M_h, \\ b_{2h}(\mathbf{u}_h, q_h) &= 0 & \forall q_h \in Q_h. \end{aligned} \quad (5.12)$$

As expected,  $b_{2h}(\mathbf{v}_h, 1) = 0$ , and therefore the system (5.12) is unchanged when the zero mean value constraint is lifted from the functions of  $Q_h$ .

### 5.1.1 Properties of the norm and forms

All constants below depend on the regularity of the mesh but are independent of  $h$ . In particular, we shall use  $C$  to denote such generic constant independent of  $h$ . In addition, we shall use the following “edge to interior” inequality. There exists a constant  $\hat{C}$ , depending only on the dimension  $d$  and the degree of the polynomials, such that for all  $\mathbf{v}_h \in V_h$ , all  $e \in \Gamma_h$  and any element  $E$ , adjacent to  $e$ ,

$$\|\mathbf{v}_h\|_{L^2(e)} \leq \hat{C} \left( \frac{|e|}{|E|} \right)^{\frac{1}{2}} \|\mathbf{v}_h\|_{L^2(E)}. \quad (5.13)$$

It is easy to check that (5.9) defines a norm on  $V_h$ . Next, the results in [6, 7] yield the following consequences of a discrete Korn inequality:

$$\|\mathbf{v}_h\|_{L^2(\Omega)} \leq C \|\mathbf{v}_h\|_{V_h} \quad \forall \mathbf{v}_h \in V_h, \quad (5.14)$$

and

$$\|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)} \leq C \|\mathbf{v}_h\|_{V_h} \quad \forall \mathbf{v}_h \in V_h, \quad (5.15)$$

where  $\nabla_h \mathbf{v}_h$  is the broken gradient (i.e., the local gradient in each element). Moreover, by following the work in [19, 24, 10], this can be generalised for all finite  $p \geq 1$  when  $d = 2$  and all  $p \in [1, 6]$  when  $d = 3$ , to

$$\|\mathbf{v}_h\|_{L^p(\Omega)} \leq C(p) \|\mathbf{v}_h\|_{V_h} \quad \forall \mathbf{v}_h \in V_h. \quad (5.16)$$

With this norm, the following compactness result holds for any sequence  $\mathbf{v}_h$  in  $V_h$ , see [10, 19, 2]: if there exists a constant  $C$  independent of  $h$  such that

$$\|\mathbf{v}_h\|_{V_h} \leq C,$$



then there exists a function  $\bar{\mathbf{v}} \in H_0^1(\Omega)^d$  such that for all finite  $p \geq 1$  when  $d = 2$  and all  $p \in [1, 6)$  when  $d = 3$ ,

$$\lim_{h \rightarrow 0} \|\mathbf{v}_h - \bar{\mathbf{v}}\|_{L^p(\Omega)} = 0. \quad (5.17)$$

Regarding the forms, a straightforward finite-dimensional argument shows that, for all  $\mathbf{u}_h, \mathbf{v}_h, \mathbf{w}_h \in V_h$ ,

$$\left| \sum_{e \in \Gamma_h} \int_e [\mathbf{u}_h]_e \cdot \mathbf{n}_e \{ \mathbf{v}_h \cdot \mathbf{w}_h \}_e \right| \leq C (J_h(\mathbf{u}_h, \mathbf{u}_h))^{\frac{1}{2}} \|\mathbf{v}_h\|_{L^4(\Omega)} \|\mathbf{w}_h\|_{L^4(\Omega)}, \quad (5.18)$$

$$\left| \sum_{e \in \Gamma_h^i} \int_e \{ \mathbf{u}_h \}_e \cdot \mathbf{n}_e [\mathbf{v}_h]_e \cdot \{ \mathbf{w}_h \}_e \right| \leq C \|\mathbf{u}_h\|_{L^4(\Omega)} (J_h(\mathbf{v}_h, \mathbf{v}_h))^{\frac{1}{2}} \|\mathbf{w}_h\|_{L^4(\Omega)}. \quad (5.19)$$

Hence we have, for all  $\mathbf{u}_h, \mathbf{v}_h, \mathbf{w}_h \in V_h$ ,

$$\begin{aligned} |d_h(\mathbf{u}_h; \mathbf{v}_h, \mathbf{w}_h)| &\leq C \|\mathbf{u}_h\|_{L^4(\Omega)} (J_h(\mathbf{v}_h, \mathbf{v}_h))^{\frac{1}{2}} \|\mathbf{w}_h\|_{L^4(\Omega)} \\ &+ \left[ \|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)} \|\mathbf{u}_h\|_{L^4(\Omega)} + \frac{1}{2} \left( \left( \sum_{E \in \mathcal{T}_h} \|\operatorname{div}(\mathbf{u}_h)\|_{L^2(E)}^2 \right)^{\frac{1}{2}} + C (J_h(\mathbf{u}_h, \mathbf{u}_h))^{\frac{1}{2}} \right) \|\mathbf{v}_h\|_{L^4(\Omega)} \right] \|\mathbf{w}_h\|_{L^4(\Omega)}. \end{aligned} \quad (5.20)$$

Similarly,

$$\left| b_{2h}(\mathbf{v}_h, q_h) \right| \leq \left( \left( \sum_{E \in \mathcal{T}_h} \|\operatorname{div}(\mathbf{v}_h)\|_{L^2(E)}^2 \right)^{\frac{1}{2}} + C (J_h(\mathbf{v}_h, \mathbf{v}_h))^{\frac{1}{2}} \right) \|q_h\|_{L^2(\Omega)} \quad \forall \mathbf{v}_h \in V_h, q_h \in Q_h, \quad (5.21)$$

$$\left| b_{1h}(\mathbf{S}_h, \mathbf{v}_h) \right| \leq \left( \|\mathbf{D}(\mathbf{v}_h)\|_h + C (J_h(\mathbf{v}_h, \mathbf{v}_h))^{\frac{1}{2}} \right) \|\mathbf{S}_h\|_{L^2(\Omega)} \quad \forall \mathbf{v}_h \in V_h, \mathbf{S}_h \in M_h. \quad (5.22)$$

Finally, the inequality below is used in choosing  $\sigma_e$ . Its proof is fairly straightforward, but it is included here for the reader's convenience.

**Proposition 3.** *For any  $\mathbf{u}_h \in V_h$ , any choice of  $\sigma_e > 0$  and any real number  $\delta > 0$ , we have*

$$\left| \sum_{e \in \Gamma_h} \int_e \{ \mathbf{D}(\mathbf{u}_h) \}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e \right| \leq \frac{1}{2} \left( \frac{1}{\delta} J_h(\mathbf{u}_h, \mathbf{u}_h) + \delta \frac{C_h}{\min_{e \in \Gamma_h} \sigma_e} \|\mathbf{D}(\mathbf{u}_h)\|_h^2 \right), \quad (5.23)$$

where

$$C_h := 2d \hat{C}^2 \max \left( \max_{e \in \Gamma_h^i} \left( h_e \max_{j=1,2} \frac{|e|}{|E_j|} \right), \max_{e \in \Gamma_h^b} \left( h_e \frac{|e|}{|E|} \right) \right), \quad (5.24)$$

$E_1$  and  $E_2$  are the elements that share the face  $e \in \Gamma_h^i$ ,  $E$  is the element that has face  $e \in \Gamma_h^b$ , and  $\hat{C}$  is the constant appearing in inequality (5.13) solely depending on  $d$  and the polynomial degree.

*Proof.* For a face  $e \in \Gamma_h^i$ , which is shared by elements  $E_1$  and  $E_2$ , we have

$$\begin{aligned} \left| \int_e \{ \mathbf{D}(\mathbf{u}_h) \}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e \right| &\leq \left( \frac{\sigma_e}{h_e} \right)^{\frac{1}{2}} \|\mathbf{u}_h\|_{L^2(e)} \left( \frac{h_e}{\sigma_e} \right)^{\frac{1}{2}} \frac{\hat{C}}{2} \sum_{j=1}^2 \left( \frac{|e|}{|E_j|} \right)^{\frac{1}{2}} \|\mathbf{D}(\mathbf{u}_h)\|_{L^2(E_j)} \\ &\leq \frac{1}{2} \left( \frac{1}{\delta} \frac{\sigma_e}{h_e} \|\mathbf{u}_h\|_{L^2(e)}^2 + \delta \frac{h_e}{2\sigma_e} \hat{C}^2 \max_{j=1,2} \frac{|e|}{|E_j|} (\|\mathbf{D}(\mathbf{u}_h)\|_{L^2(E_1)}^2 + \|\mathbf{D}(\mathbf{u}_h)\|_{L^2(E_2)}^2) \right). \end{aligned}$$

Similarly, for  $e \in \Gamma_h^b$ , which is the face of an element  $E$  adjacent to  $\partial\Omega$ , we have

$$\left| \int_e \{ \mathbf{D}(\mathbf{u}_h) \}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e \right| \leq \frac{1}{2} \left( \frac{1}{\delta} \frac{\sigma_e}{h_e} \|\mathbf{u}_h\|_{L^2(e)}^2 + \delta \frac{h_e}{\sigma_e} \hat{C}^2 \frac{|e|}{|E|} \|\mathbf{D}(\mathbf{u}_h)\|_{L^2(E)}^2 \right).$$

By using the last two inequalities in

$$\left| \sum_{e \in \Gamma_h} \int_e \{ \mathbf{D}(\mathbf{u}_h) \}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e \right| \leq \sum_{e \in \Gamma_h} \left| \int_e \{ \mathbf{D}(\mathbf{u}_h) \}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e \right|$$

and splitting the sum on the right-hand side into sums over the disjoint sets  $\Gamma_h^b$  and  $\Gamma_h^i$ , we have that

$$\begin{aligned} \left| \sum_{e \in \Gamma_h} \int_e \{D(\mathbf{u}_h)\}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e \right| &\leq \frac{1}{2} \left( \frac{1}{\delta} \sum_{e \in \Gamma_h^b} \frac{\sigma_e}{h_e} \|\mathbf{u}_h\|_{L^2(e)}^2 + \delta \sum_{e \in \Gamma_h^b} \frac{h_e}{\sigma_e} \hat{C}^2 \frac{|e|}{|E|} \|D(\mathbf{u}_h)\|_{L^2(E)}^2 \right) \\ &+ \frac{1}{2} \left( \frac{1}{\delta} \sum_{e \in \Gamma_h^i} \frac{\sigma_e}{h_e} \|\mathbf{u}_h\|_{L^2(e)}^2 + \delta \sum_{e \in \Gamma_h^i} \frac{h_e}{2\sigma_e} \hat{C}^2 \max_{j=1,2} \frac{|e|}{|E_j|} (\|D(\mathbf{u}_h)\|_{L^2(E_1)}^2 + \|D(\mathbf{u}_h)\|_{L^2(E_2)}^2) \right) \end{aligned}$$

with the notational convention that when summing over  $e \in \Gamma_h^b$  the element  $E$  under the summation sign is the element adjacent to  $\partial\Omega$  with face  $e$ , and when summing over  $e \in \Gamma_h^i$  the elements  $E_1$  and  $E_2$  under the summation sign are the ones that share the face  $e$ . Hence,

$$\begin{aligned} \left| \sum_{e \in \Gamma_h} \int_e \{D(\mathbf{u}_h)\}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e \right| &\leq \frac{1}{2} \left( \frac{1}{\delta} J_h(\mathbf{u}_h, \mathbf{u}_h) + \frac{\delta}{\min_{e \in \Gamma_h^b} \sigma_e} \hat{C}^2 \max_{e \in \Gamma_h^b} \left( h_e \frac{|e|}{|E|} \right) \sum_{e \in \Gamma_h^b} \|D(\mathbf{u}_h)\|_{L^2(E)}^2 \right) \\ &+ \frac{1}{2} \left( \frac{\delta}{\min_{e \in \Gamma_h^i} \sigma_e} \hat{C}^2 \max_{e \in \Gamma_h^i} \left( h_e \max_{j=1,2} \frac{|e|}{|E_j|} \right) \sum_{e \in \Gamma_h^i} \frac{1}{2} (\|D(\mathbf{u}_h)\|_{L^2(E_1)}^2 + \|D(\mathbf{u}_h)\|_{L^2(E_2)}^2) \right). \end{aligned}$$

The asserted result (5.23) follows from the last inequality by noting that, for each  $E \in \mathcal{T}_h$ , the factor  $\|D(\mathbf{u}_h)\|_{L^2(E)}^2$  appears at most  $2d$  times.  $\square$

Concerning the expression appearing in (5.24) we note that, thanks to the regularity assumption on the family of meshes, we have that  $h_e \frac{|e|}{|E|} \leq C$  and so

$$C_h \leq C. \quad (5.25)$$

### 5.1.2 First *a priori* estimates

By testing the first equation of (5.12) with  $\mathbf{v}_h = \mathbf{u}_h$ , applying the third equation and the antisymmetry (5.7) of  $d_h$ , we obtain

$$b_{1h}(\mathbf{T}_h, \mathbf{u}_h) + J_h(\mathbf{u}_h, \mathbf{u}_h) = \int_{\Omega} \mathbf{f} \cdot \mathbf{u}_h.$$

Next, by testing the second equation of (5.12) with  $\mathbf{S}_h = \mathbf{T}_h$  and substituting the above equality, we deduce that

$$\alpha \|\mathbf{T}_h\|_{L^2(\Omega)}^2 + J_h(\mathbf{u}_h, \mathbf{u}_h) \leq \int_{\Omega} \mathbf{f} \cdot \mathbf{u}_h.$$

Thus, in view of (5.14), we have our first bound:

$$\alpha \|\mathbf{T}_h\|_{L^2(\Omega)}^2 + J_h(\mathbf{u}_h, \mathbf{u}_h) \leq C \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{u}_h\|_{V_h}. \quad (5.26)$$

A further bound is arrived at by testing the second equation of (5.12) with  $\mathbf{S}_h = D(\mathbf{u}_h)$ ; hence,

$$\alpha \int_{\Omega} \mathbf{T}_h : D(\mathbf{u}_h) + \gamma \int_{\Omega} \mu(|\mathbf{T}_h|) \mathbf{T}_h : D(\mathbf{u}_h) = \|D(\mathbf{u}_h)\|_h^2 - \sum_{e \in \Gamma_h} \int_e \{D(\mathbf{u}_h)\}_e \mathbf{n}_e \cdot [\mathbf{u}_h]_e.$$

Then Proposition 3 gives, for any  $\delta > 0$ ,

$$\|D(\mathbf{u}_h)\|_h^2 \leq \alpha \|\mathbf{T}_h\|_{L^2(\Omega)} \|D(\mathbf{u}_h)\|_h + \gamma C_1 |\Omega|^{\frac{1}{2}} \|D(\mathbf{u}_h)\|_h + \frac{1}{2} \left( \frac{1}{\delta} J_h(\mathbf{u}_h, \mathbf{u}_h) + \delta \frac{C_h}{\min_{e \in \Gamma_h} \sigma_e} \|D(\mathbf{u}_h)\|_h^2 \right). \quad (5.27)$$

We choose  $\delta = 1$  and, upon recalling (5.25), assume that  $\sigma_e$  is chosen so that

$$\min_{e \in \Gamma_h} \sigma_e \geq C_h. \quad (5.28)$$

Next, by adding  $J_h(\mathbf{u}_h, \mathbf{u}_h)$  to both sides of (5.27), applying (5.26) to bound this term, and using the norm of  $V_h$ , we infer that

$$\|\mathbf{u}_h\|_{V_h}^2 \leq \alpha \|\mathbf{T}_h\|_{L^2(\Omega)} \|\mathbf{D}(\mathbf{u}_h)\|_h + \gamma C_1 |\Omega|^{\frac{1}{2}} \|\mathbf{D}(\mathbf{u}_h)\|_h + C \|\mathbf{f}\|_{L^2(\Omega)} \|\mathbf{u}_h\|_{V_h} + \frac{1}{2} \|\mathbf{u}_h\|_{V_h}^2$$

and thus

$$\frac{1}{2} \|\mathbf{u}_h\|_{V_h} \leq \alpha \|\mathbf{T}_h\|_{L^2(\Omega)} + \gamma C_1 |\Omega|^{\frac{1}{2}} + C \|\mathbf{f}\|_{L^2(\Omega)}. \quad (5.29)$$

To close the estimates, we return to (5.26) and get

$$\alpha \|\mathbf{T}_h\|_{L^2(\Omega)}^2 + J_h(\mathbf{u}_h, \mathbf{u}_h) \leq \frac{1}{2} \left( \delta_2 \|\mathbf{u}_h\|_{V_h}^2 + \frac{C^2}{\delta_2} \|\mathbf{f}\|_{L^2(\Omega)}^2 \right)$$

for any  $\delta_2 > 0$ . Thus

$$\alpha \|\mathbf{T}_h\|_{L^2(\Omega)} \leq \frac{\sqrt{\alpha}}{\sqrt{2\delta_2}} C \|\mathbf{f}\|_{L^2(\Omega)} + \frac{\sqrt{\alpha\delta_2}}{\sqrt{2}} \|\mathbf{u}_h\|_{V_h},$$

and the choice  $\delta_2 = \frac{1}{8\alpha}$  yields

$$\alpha \|\mathbf{T}_h\|_{L^2(\Omega)} \leq 2\alpha C \|\mathbf{f}\|_{L^2(\Omega)} + \frac{1}{4} \|\mathbf{u}_h\|_{V_h}.$$

Thus we have shown the following uniform and unconditional bounds:

$$\|\mathbf{u}_h\|_{V_h} \leq 4C(1 + 2\alpha) \|\mathbf{f}\|_{L^2(\Omega)} + 4\gamma C_1 |\Omega|^{\frac{1}{2}}, \quad \|\mathbf{T}_h\|_{L^2(\Omega)} \leq \frac{C}{\alpha} (4\alpha + 1) \|\mathbf{f}\|_{L^2(\Omega)} + \frac{\gamma}{\alpha} C_1 |\Omega|^{\frac{1}{2}}. \quad (5.30)$$

An *a priori* estimate for the pressure requires an inf-sup condition. This is the subject of the next subsection.

### 5.1.3 An inf-sup condition

In the nonconforming case considered here, the analogue of (4.3) reads

$$\inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in V_h} \frac{b_{2h}(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{V_h} \|q_h\|_{L^2(\Omega)}} \geq \beta^* \quad (5.31)$$

with a constant  $\beta^* > 0$  independent of  $h$ . To check this condition, recall Fortin's lemma; see for instance [20].

**Lemma 6.** *The discrete condition (5.31) holds uniformly with respect to  $h$  if, and only if, there exists an approximation operator  $\Pi_h \in \mathcal{L}(H_0^1(\Omega)^d; V_h)$  such that, for all  $\mathbf{v} \in H_0^1(\Omega)^d$ ,*

$$b_{2h}(\Pi_h(\mathbf{v}) - \mathbf{v}, q_h) = 0 \quad \forall q_h \in Q_h, \quad (5.32)$$

and

$$\|\Pi_h(\mathbf{v})\|_{V_h} \leq C \|\mathbf{v}\|_{H^1(\Omega)} \quad (5.33)$$

with a constant  $C$  independent of  $h$ .

Originally, Fortin's lemma was stated for discrete functions in subspaces of  $H_0^1(\Omega)^d$ , but the extension to spaces of discontinuous functions is straightforward, as long as the form  $b_{2h}(\cdot, \cdot)$  is consistent with the divergence, which is the case here.

As the proof of (5.32), (5.33) is fairly technical, we restrict the discussion to the first order case, i.e.,  $k = 1$ , in hexahedra. The quadrilateral case is much simpler.

### 5.1.4 The inf-sup condition in planar-faced hexahedra for $k = 1$

The construction of a suitable operator  $\Pi_h$  is usually done by correcting a good approximation operator  $R_h$ . For instance, we can use the  $L^2$  projection onto the space of polynomials of degree one defined locally in each element, so that  $R_h(\mathbf{v})$  belongs to  $V_h$  and satisfies optimal approximation properties; see for instance [8]. Then  $R_h(\mathbf{v})$  is corrected by constructing  $\mathbf{c}_h \in V_h$  such that

$$b_{2h}(\mathbf{c}_h, q_h) = b_{2h}(R_h(\mathbf{v}) - \mathbf{v}, q_h) \quad \forall q_h \in Q_h. \quad (5.34)$$

By expanding  $b_{2h}$  and denoting by  $q_E$  the value of  $q_h$  in  $E$ , (5.34) reads

$$- \sum_{E \in \mathcal{T}_h} q_E \int_E \operatorname{div}(\mathbf{c}_h) + \sum_{e \in \Gamma_h} \int_e [\mathbf{c}_h]_e \cdot \mathbf{n}_e \{q_h\}_e = - \sum_{E \in \mathcal{T}_h} q_E \int_E \operatorname{div}(R_h(\mathbf{v}) - \mathbf{v}) + \sum_{e \in \Gamma_h} \int_e [R_h(\mathbf{v}) - \mathbf{v}]_e \cdot \mathbf{n}_e \{q_h\}_e.$$

Green's formula in each element yields

$$- \sum_{E \in \mathcal{T}_h} q_E \int_{\partial E} \mathbf{c}_h \cdot \mathbf{n}_E + \sum_{e \in \Gamma_h} \int_e [\mathbf{c}_h]_e \cdot \mathbf{n}_e \{q_h\}_e = - \sum_{E \in \mathcal{T}_h} q_E \int_{\partial E} (R_h(\mathbf{v}) - \mathbf{v}) \cdot \mathbf{n}_E + \sum_{e \in \Gamma_h} \int_e [R_h(\mathbf{v}) - \mathbf{v}]_e \cdot \mathbf{n}_e \{q_h\}_e \quad (5.35)$$

with  $\mathbf{n}_E$  the unit exterior normal to  $E$ . Consider now an interior face  $e$  shared by  $E_1$  and  $E_2$ , so that  $\mathbf{n}_e$  is interior to  $E_2$ ; the contribution of  $e$  to the left-hand side of (5.35) is

$$-q_{E_1} \int_e \mathbf{c}_h|_{E_1} \cdot \mathbf{n}_e + q_{E_2} \int_e \mathbf{c}_h|_{E_2} \cdot \mathbf{n}_e + \int_e \frac{1}{2} (q_{E_1} + q_{E_2}) (\mathbf{c}_h|_{E_1} - \mathbf{c}_h|_{E_2}) \cdot \mathbf{n}_e = - \int_e [q_h]_e \{ \mathbf{c}_h \}_e \cdot \mathbf{n}_e$$

with a similar contribution to the right-hand side. Notice also that the contribution of a boundary face  $e \in \Gamma_h^b$  is equal to zero on both sides of (5.35). Therefore a sufficient condition for (5.35) is that

$$\int_e \mathbf{c}_h|_E \cdot \mathbf{n}_e = \int_e (R_h(\mathbf{v}) - \mathbf{v})|_E \cdot \mathbf{n}_e. \quad (5.36)$$

We will thus construct  $\mathbf{c}_h \in V_h$  by imposing (5.36) for each element  $E \in \mathcal{T}_h$  and each face  $e \in \partial E$ . To simplify the notation, we will write from now on  $\mathbf{c}_h$  and  $(R_h(\mathbf{v}) - \mathbf{v})$  instead of  $\mathbf{c}_h|_E$  and  $(R_h(\mathbf{v}) - \mathbf{v})|_E$ , respectively.

Let  $E$  be an arbitrary hexahedral element of  $\mathcal{T}_h$  with faces  $e_i$ , centre of face  $\mathbf{b}_i$ , and exterior unit normal  $\mathbf{n}_i$ ,  $1 \leq i \leq 6$ . To be specific, let  $\mathbf{a}_i$ ,  $i = 1, 2, 3, 4$ , be the vertices of  $e_1$ ,  $\mathbf{a}_i$ ,  $i = 1, 3, 5, 6$ , the vertices of  $e_2$ ,  $\mathbf{a}_i$ ,  $i = 1, 2, 5, 7$ , the vertices of  $e_3$ ,  $\mathbf{a}_i$ ,  $i = 5, 6, 7, 8$ , the vertices of  $e_4$ ,  $\mathbf{a}_i$ ,  $i = 2, 4, 7, 8$ , the vertices of  $e_5$ , and  $\mathbf{a}_i$ ,  $i = 3, 4, 6, 8$ , the vertices of  $e_6$ . The ordering of the nodes is illustrated in Figure 1. Note that for  $i = 1, 2, 3$ ,  $e_{i+3}$  is the face opposite to  $e_i$ , opposite in the sense that its intersection with  $e_i$  is empty.

Without loss of generality, we assume that the vertex  $\mathbf{a}_1$  is located at the origin and that the face  $e_1$  lies on the  $x_3 = 0$  plane. Indeed, this situation can be obtained via a rigid motion (translation plus rotation), which preserves all normal vectors. Therefore, the normal to the face  $e_1$  is parallel to the  $x_3$  axis. Now, the idea is to transform  $E$  onto a “reference” element  $\hat{E}$  by an affine mapping  $\mathcal{F}_E$  so that the subtetrahedron  $S_1$  of  $E$  based on  $e_1$  and containing the origin  $\mathbf{a}_1$  is mapped onto the unit tetrahedron  $\hat{S}_1$ . More precisely, as  $e_2$  and  $e_3$  are both adjacent to  $e_1$ ,  $S_1$  is the subtetrahedron with vertices  $\mathbf{a}_1$ ,  $\mathbf{a}_2$ ,  $\mathbf{a}_3$ , and  $\mathbf{a}_5$ , and  $\hat{S}_1$  has vertices  $\hat{\mathbf{a}}_1 = (0, 0, 0)$ ,  $\hat{\mathbf{a}}_2 = (0, 1, 0)$ ,  $\hat{\mathbf{a}}_3 = (1, 0, 0)$ ,  $\hat{\mathbf{a}}_5 = (0, 0, 1)$ , see Figure 1 for an illustration and some notation. This transformation and notation will be used till the end of this subsection. It stems from the regularity of the family of triangulations that there exists a constant  $M$ , independent of  $E$  and  $h$ , such that

$$\operatorname{diameter}(\hat{E}) \leq M. \quad (5.37)$$

The affine mapping  $\mathcal{F}_E$  has the expression

$$\mathbf{x} = \mathcal{F}_E(\hat{\mathbf{x}}) = \mathbf{B}\hat{\mathbf{x}},$$

where the constant term is zero since  $\mathbf{a}_1$  is the origin, and the matrix  $\mathbf{B}$  is nonsingular; its columns are respectively  $\mathbf{a}_3 = (a_3^1, a_3^2, 0)^t$ ,  $\mathbf{a}_2 = (a_2^1, a_2^2, 0)^t$  and  $\mathbf{a}_5 = (a_5^1, a_5^2, a_5^3)^t$ . The image of the remaining vertices of

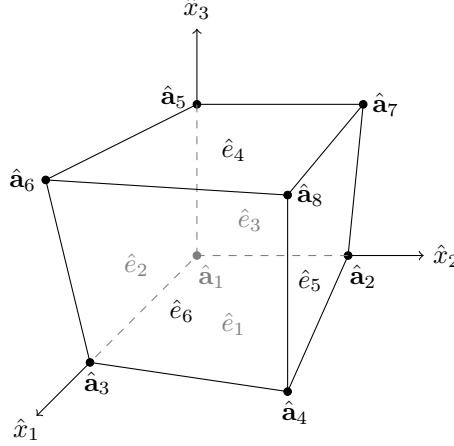


Figure 1: Some notation for the “reference” element  $\hat{E}$ .

$E$  are  $\hat{\mathbf{a}}_i = \mathcal{F}_E^{-1}(\mathbf{a}_i)$ ,  $i = 4, 6, 7, 8$ . As  $\mathcal{F}_E$  is an affine transformation, it transforms faces onto faces, edges onto edges, and vertices onto vertices. Thus, since  $\mathbf{a}_4$  is in the plane  $x_3 = 0$ , then  $\hat{\mathbf{a}}_4$  is in the plane  $\hat{x}_3 = 0$ . Likewise,  $\hat{\mathbf{a}}_6$  is in the plane  $\hat{x}_2 = 0$ ,  $\hat{\mathbf{a}}_7$  in the plane  $\hat{x}_1 = 0$ , and  $\hat{\mathbf{a}}_8$  in the plane determined by  $\hat{\mathbf{a}}_4$ ,  $\hat{\mathbf{a}}_2$ ,  $\hat{\mathbf{a}}_7$ , as well as the plane determined by  $\hat{\mathbf{a}}_7$ ,  $\hat{\mathbf{a}}_5$ ,  $\hat{\mathbf{a}}_6$ , and the plane determined by  $\hat{\mathbf{a}}_6$ ,  $\hat{\mathbf{a}}_3$ ,  $\hat{\mathbf{a}}_4$ , hence in the intersection of these three planes. Therefore  $\hat{E}$  is located in the first octant of  $\mathbb{R}^3$ . Let  $\hat{\mathbf{n}}_i$  denote the unit exterior normal vector to  $\hat{e}_i$ . It is related to  $\mathbf{n}_i$  by the general formula

$$\hat{\mathbf{n}}_i = \frac{\mathbf{B}^t \mathbf{n}_i}{|\mathbf{B}^t \mathbf{n}_i|}. \quad (5.38)$$

The advantage of having  $e_1$  on the plane  $x_3 = 0$  is that  $\hat{\mathbf{n}}_1 = \mathbf{n}_1 = (0, 0, -1)^t$ . We also have  $\hat{\mathbf{n}}_2 = (0, -1, 0)^t$ , and  $\hat{\mathbf{n}}_3 = (-1, 0, 0)^t$ . Thus

$$|\hat{n}_1^3| = |\hat{n}_2^2| = |\hat{n}_3^1| = 1, \quad (5.39)$$

and the regularity of the family  $\mathcal{T}_h$  implies that there exists a constant  $\nu_0$ , independent of  $h$  and  $E$ , such that

$$|\hat{n}_4^3|, |\hat{n}_5^2|, |\hat{n}_6^1| \geq \nu_0. \quad (5.40)$$

With this transformation, and after cancelling  $|\det \mathbf{B}|$  on both sides, (5.36) reads locally

$$\int_{\hat{e}} \hat{\mathbf{c}}_h \cdot (\mathbf{B}^t)^{-1} \hat{\mathbf{n}}_{\hat{e}} = \int_{\hat{e}} (\widehat{R}_h(\hat{\mathbf{v}}) - \hat{\mathbf{v}}) \cdot (\mathbf{B}^t)^{-1} \hat{\mathbf{n}}_{\hat{e}},$$

where the hat denotes composition with  $\mathcal{F}_E$ . Thus, by performing the change of variable

$$\hat{\mathbf{d}}_h = \mathbf{B}^{-1} \hat{\mathbf{c}}_h$$

and defining the face moment

$$m_{\hat{e}}(f) := \frac{1}{|\hat{e}|} \int_{\hat{e}} f,$$

(5.36) is equivalent to

$$m_{\hat{e}_i}(\hat{\mathbf{d}}_h) \cdot \hat{\mathbf{n}}_i = \hat{g}_i := \frac{1}{|\hat{e}_i|} \int_{\hat{e}_i} \mathbf{B}^{-1} (\widehat{R}_h(\hat{\mathbf{v}}) - \hat{\mathbf{v}}) \cdot \hat{\mathbf{n}}_i, \quad 1 \leq i \leq 6. \quad (5.41)$$

This is a linear system of six equations in twelve unknowns, the coefficients of  $\hat{\mathbf{d}}_h$ . Therefore, we can freely choose six coefficients and we have the following existence lemma.

**Lemma 7.** *There exists exactly one polynomial vector  $\hat{\mathbf{d}}_h = (\hat{d}_1, \hat{d}_2, \hat{d}_3)^t$  that satisfies (5.41) and the following six conditions:*

$$m_{\hat{e}_1}(\hat{d}_1) = m_{\hat{e}_5}(\hat{d}_1) = m_{\hat{e}_1}(\hat{d}_2) = m_{\hat{e}_6}(\hat{d}_2) = m_{\hat{e}_5}(\hat{d}_3) = m_{\hat{e}_6}(\hat{d}_3) = 0. \quad (5.42)$$

*Proof.* Once the six conditions (5.42) are prescribed, we are left with a square linear system of six equations in six unknowns. Therefore it suffices to prove that the only solution of the corresponding homogeneous system is the zero solution. To begin with, we consider the lines  $i = 5$  and  $i = 6$  in (5.41). In view of (5.39) and (5.40), the strategy for the choice (5.42) is to set to zero the coefficients of  $\hat{n}_6^2$  and  $\hat{n}_6^3$  and those of  $\hat{n}_5^1$  and  $\hat{n}_5^3$ , i.e., prescribe  $m_{\hat{e}_5}(\hat{d}_1) = m_{\hat{e}_6}(\hat{d}_2) = m_{\hat{e}_5}(\hat{d}_3) = m_{\hat{e}_6}(\hat{d}_3) = 0$ . With this assumption, the lines  $i = 6$  and  $i = 5$  reduce respectively to

$$m_{\hat{e}_6}(\hat{d}_1) = 0, \quad m_{\hat{e}_5}(\hat{d}_2) = 0. \quad (5.43)$$

Next, we consider the line  $i = 1$ . As  $\hat{n}_1^3 = -1$  is the only nonzero component, it reduces to

$$m_{\hat{e}_1}(\hat{d}_3) = 0. \quad (5.44)$$

Similarly, when  $i = 2$  and  $i = 3$  we have, respectively

$$m_{\hat{e}_2}(\hat{d}_2) = 0, \quad m_{\hat{e}_3}(\hat{d}_1) = 0. \quad (5.45)$$

Collecting these results and the two extra assumptions  $m_{\hat{e}_1}(\hat{d}_1) = m_{\hat{e}_1}(\hat{d}_2) = 0$  in (5.42), we find that

$$m_{\hat{e}_1}(\hat{d}_1) = m_{\hat{e}_5}(\hat{d}_1) = m_{\hat{e}_6}(\hat{d}_1) = m_{\hat{e}_3}(\hat{d}_1) = 0$$

$$m_{\hat{e}_1}(\hat{d}_2) = m_{\hat{e}_5}(\hat{d}_2) = m_{\hat{e}_6}(\hat{d}_2) = m_{\hat{e}_2}(\hat{d}_2) = 0.$$

The three faces  $\hat{e}_1, \hat{e}_5, \hat{e}_6$  share the vertex  $\hat{\mathbf{a}}_4$ , and the regularity of the hexahedron implies that the three vectors along the segments  $[\hat{\mathbf{a}}_4, \hat{\mathbf{a}}_3]$ ,  $[\hat{\mathbf{a}}_4, \hat{\mathbf{a}}_8]$ , and  $[\hat{\mathbf{a}}_4, \hat{\mathbf{a}}_2]$  is a set of three linearly independent vectors of  $\mathbb{R}^3$ . Then the regularity of the hexahedron implies that a polynomial of degree one is uniquely determined by its moments on the four faces  $\hat{e}_1, \hat{e}_5, \hat{e}_6, \hat{e}_i$  for any  $i$  in the set  $\{2, 3, 4\}$ . Hence, as  $\hat{d}_1$  (respectively,  $\hat{d}_2$ ) is a polynomial of degree one, the first set (respectively, second set) of equalities and the regularity of the hexahedron imply that  $\hat{d}_1 = 0$ , respectively,  $\hat{d}_2 = 0$ . When  $i = 4$ , this leads to  $m_{\hat{e}_4}(\hat{d}_3) = 0$ . Consequently,

$$m_{\hat{e}_1}(\hat{d}_3) = m_{\hat{e}_5}(\hat{d}_3) = m_{\hat{e}_6}(\hat{d}_3) = m_{\hat{e}_4}(\hat{d}_3) = 0,$$

and  $\hat{d}_3 = 0$ . Thus  $\hat{\mathbf{d}}_h = \mathbf{0}$  and the system has a unique solution.  $\square$

Let  $\mathbf{M}_{\hat{E}}$  be the  $6 \times 6$  matrix of the system (5.41) under the restriction (5.42). It stems from Lemma 7 that  $\mathbf{M}_{\hat{E}}$  is nonsingular. Furthermore, the regularity of the hexahedron implies that  $\mathbf{M}_{\hat{E}}$  is a continuous function of  $\hat{E}$ , thus continuous in a compact set of  $\mathbb{R}^3$ . Hence the norm of its inverse is bounded by a constant  $\hat{C}$ , independent of  $\hat{E}$ ,

$$|\mathbf{M}_{\hat{E}}^{-1}| \leq \hat{C}. \quad (5.46)$$

The stability of the correction follows now easily.

**Lemma 8.** *There exists a constant  $\hat{C}$ , independent of  $h$  and  $E$ , such that for all  $E$  in  $\mathcal{T}_h$  and all  $e$  in  $\Gamma_h$ ,*

$$\|\mathbf{c}_h\|_{L^2(E)} \leq \hat{C} h_E |\mathbf{v}|_{H^1(E)}, \quad |\mathbf{c}_h|_{H^1(E)} \leq \hat{C} |\mathbf{v}|_{H^1(E)}, \quad \left(\frac{\sigma_e}{h_e}\right)^{\frac{1}{2}} \|[\mathbf{c}_h]_e\|_{L^2(e)} \leq \hat{C} (|\mathbf{v}|_{H^1(E_1)} + |\mathbf{v}|_{H^1(E_2)}), \quad (5.47)$$

where  $E_1$  and  $E_2$  are the two elements sharing  $e$ , when  $e$  is an interior face, and the sum is reduced to one term, namely the element  $E$  adjacent to  $e$ , when  $e$  is a boundary face.

*Proof.* The notation  $\hat{C}$  below refers to different constants that are all independent of  $h$  and  $E$ . Recalling (5.41), (5.37) and the transformation from  $\hat{S}_1$  onto  $S_1$ , we observe that, for any  $i$ ,

$$|\hat{g}_i| \leq \frac{\hat{C}}{\varrho_{S_1}} \|\widehat{R}_h(\hat{\mathbf{v}}) - \hat{\mathbf{v}}\|_{L^1(\hat{e}_i)} \leq \frac{\hat{C}}{\varrho_{S_1}} \|\widehat{R}_h(\hat{\mathbf{v}}) - \hat{\mathbf{v}}\|_{L^2(\hat{e}_i)}.$$

By a trace inequality in  $\hat{E}$  and the approximation property of  $\widehat{R}_h$  in  $\hat{E}$ , we have

$$\sum_{i=1}^6 \|\widehat{R}_h(\hat{\mathbf{v}}) - \hat{\mathbf{v}}\|_{L^2(\hat{e}_i)} \leq \hat{C} \|\widehat{R}_h(\hat{\mathbf{v}}) - \hat{\mathbf{v}}\|_{H^1(\hat{E})} \leq \hat{C} |\hat{\mathbf{v}}|_{H^1(\hat{E})}.$$

Then, by reverting to  $E$ ,

$$\sum_{i=1}^6 \|\widehat{R}_h(\hat{\mathbf{v}}) - \hat{\mathbf{v}}\|_{L^2(\hat{e}_i)} \leq \hat{C} \frac{h_{S_1}}{|E|^{\frac{1}{2}}} |\mathbf{v}|_{H^1(E)}.$$

In view of (5.46) and the regularity of the family  $\mathcal{T}_h$ , the above relations lead to the following bound on  $\hat{\mathbf{d}}_h$ :

$$\|\hat{\mathbf{d}}_h\|_{L^\infty(\hat{E})} \leq \frac{\hat{C}}{|E|^{\frac{1}{2}}} \frac{h_{S_1}}{\varrho_{S_1}} |\mathbf{v}|_{H^1(E)} \leq \frac{\hat{C}}{|E|^{\frac{1}{2}}} |\mathbf{v}|_{H^1(E)};$$

with  $\hat{\mathbf{c}}_h = \mathbf{B}\hat{\mathbf{d}}_h$ , this yields

$$\|\mathbf{c}_h\|_{L^\infty(E)} = \|\hat{\mathbf{c}}_h\|_{L^\infty(\hat{E})} \leq \hat{C} \frac{h_{S_1}}{|E|^{\frac{1}{2}}} |\mathbf{v}|_{H^1(E)}. \quad (5.48)$$

Since  $h_{S_1} < h_E$ , we immediately deduce from (5.48) the first two inequalities in (5.47). Finally, the third inequality follows from (5.48) and

$$\left(\frac{\sigma_e}{h_e}\right)^{\frac{1}{2}} \|\mathbf{c}_h\|_{L^2(e)} \leq \left(\frac{\sigma_e}{h_e}\right)^{\frac{1}{2}} |e|^{\frac{1}{2}} \|\mathbf{c}_h\|_{L^\infty(E)}.$$

That completes the proof of the lemma.  $\square$

As a consequence of Lemma 8 we have the following bounds:

$$\|\mathbf{c}_h\|_{L^2(\Omega)} \leq \hat{C} h |\mathbf{v}|_{H^1(\Omega)}, \quad \|\mathbf{c}_h\|_{V_h} \leq \hat{C} |\mathbf{v}|_{H^1(\Omega)}. \quad (5.49)$$

Finally, since the construction of Lemma 7 yields a unique correction, it is easy to check that the mapping  $\mathbf{v} \mapsto \mathbf{c}_h$  defines a linear operator from  $V_h$  into itself, i.e.,  $\mathbf{c}_h = \mathbf{c}_h(\mathbf{v})$ .

On the other hand, we infer from standard approximation properties of  $R_h$  and the regularity of the mesh, that

$$\begin{aligned} \|\mathbf{v} - R_h(\mathbf{v})\|_{L^2(E)} &\leq \hat{C} h_E |\mathbf{v}|_{H^1(E)}, \quad |R_h(\mathbf{v})|_{H^1(E)} \leq \hat{C} |\mathbf{v}|_{H^1(E)}, \\ \left(\frac{\sigma_e}{h_e}\right)^{\frac{1}{2}} \|[R_h(\mathbf{v})]_e\|_{L^2(e)} &\leq \hat{C} (|\mathbf{v}|_{H^1(E_1)} + |\mathbf{v}|_{H^1(E_2)}) \end{aligned} \quad (5.50)$$

and

$$\|\mathbf{v} - R_h(\mathbf{v})\|_{L^2(\Omega)} \leq \hat{C} h |\mathbf{v}|_{H^1(\Omega)}, \quad \|R_h(\mathbf{v})\|_{V_h} \leq \hat{C} |\mathbf{v}|_{H^1(\Omega)}. \quad (5.51)$$

Thus  $\Pi_h(\mathbf{v}) = R_h(\mathbf{v}) - \mathbf{c}_h(\mathbf{v})$  satisfies the conditions (5.32) and (5.33) of Lemma 6. This proves the inf-sup condition as stated in the next theorem.

**Theorem 4.** *Let the family of hexahedra  $\mathcal{T}_h$  be regular in the sense defined above. Then the form  $b_{2h}$  defined in (5.6) with the pair spaces  $V_h$  and  $Q_h$  for  $k = 1$ , see (5.1) and (5.2), satisfies the inf-sup condition (5.31) with a constant  $\beta^* > 0$  independent of  $h$ .*

### 5.1.5 A bound on the pressure

As usual, the inf-sup condition (5.31) yields a bound on the pressure. Indeed, it follows from the first equation of (5.12) together with (5.22), (5.20), (5.15) and (5.16) that

$$|b_{2h}(\mathbf{v}_h, p_h)| \leq C (\|\mathbf{T}_h\|_{L^2(\Omega)} + \|\mathbf{u}_h\|_{V_h} + \|\mathbf{u}_h\|_{V_h}^2 + \|\mathbf{f}\|_{L^2(\Omega)}) \|\mathbf{v}_h\|_{V_h}.$$

Then (5.30) implies, with a constant  $C$  independent of  $h$  (but depending on  $\alpha$ ), that

$$|b_{2h}(\mathbf{v}_h, p_h)| \leq C \|\mathbf{v}_h\|_{V_h} \quad \forall \mathbf{v}_h \in V_h.$$

With the inf-sup condition (5.31), this implies that

$$\|p_h\|_{L^2(\Omega)} \leq C, \quad (5.52)$$

for another constant  $C$  independent of  $h$ .

### 5.1.6 Existence and convergence

The proof of existence of a solution of (5.12) is the same as in the conforming case. Recall that the case of interest is  $k = 1$ , which is assumed for the remainder of this subsection, but all of what follows can be straightforwardly extended to a general polynomial degree  $k \geq 1$  as long as the inf-sup condition (5.31) holds. First, the problem is reduced to one equation by testing the first equation of (5.12) with  $\mathbf{v}_h \in V_{h,0}$  and by observing that the second equation determines for each  $\mathbf{u}_h \in V_h$  a unique  $\mathbf{T}_h$  in  $M_h$ . This is expressed by writing  $\mathbf{T}_h = \mathcal{G}_{h,DG}(\mathbf{u}_h)$ . Then, (5.12) is equivalent to finding a  $\mathbf{u}_h \in V_{h,0}$  such that

$$d_h(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + b_{1h}(\mathcal{G}_{h,DG}(\mathbf{u}_h), \mathbf{v}_h) + J_h(\mathbf{u}_h, \mathbf{v}_h) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h \quad \forall \mathbf{v}_h \in V_{h,0}. \quad (5.53)$$

By means of the *a priori* estimates (5.30), existence of a solution is deduced by Brouwer's fixed point theorem.

Regarding convergence, the *a priori* estimates (5.30) and (5.52) together with (5.17) imply that there exist functions  $\bar{\mathbf{T}} \in L^2(\Omega)_{\text{sym}}^{d \times d}$ ,  $\bar{\mathbf{u}} \in H_0^1(\Omega)^d$ , and  $\bar{p} \in L_0^2(\Omega)$  such that, up to subsequences,

$$\lim_{h \rightarrow 0} \|\mathbf{u}_h - \bar{\mathbf{u}}\|_{L^q(\Omega)} = 0 \quad \text{with } 1 \leq q < \infty \text{ if } d = 2, \text{ and } 1 \leq q < 6 \text{ if } d = 3,$$

$$\lim_{h \rightarrow 0} \mathbf{T}_h = \bar{\mathbf{T}} \quad \text{weakly in } L^2(\Omega)^{d \times d},$$

and

$$\lim_{h \rightarrow 0} p_h = \bar{p} \quad \text{weakly in } L^2(\Omega).$$

However, in order to pass to the limit in the equations of the scheme, following [16], we need to introduce discrete differential operators related to distributional differential operators. These are  $G_h^{\text{sym}}(\mathbf{v}_h) \in M_h$  and  $G_h^{\text{div}}(\mathbf{v}_h) \in \Theta_h$ , defined for all  $\mathbf{v}_h \in V_h$  by, respectively,

$$\int_{\Omega} G_h^{\text{sym}}(\mathbf{v}_h) : \mathbf{R}_h = b_{1h}(\mathbf{R}_h, \mathbf{v}_h) = \sum_{E \in \mathcal{T}_h} \int_E \mathbf{D}(\mathbf{v}_h) : \mathbf{R}_h - \sum_{e \in \Gamma_h} \int_e [\mathbf{v}_h]_e \cdot \{\mathbf{R}_h\}_e \mathbf{n}_e \quad \forall \mathbf{R}_h \in M_h, \quad (5.54)$$

$$\int_{\Omega} G_h^{\text{div}}(\mathbf{v}_h) r_h = b_{2h}(\mathbf{v}_h, r_h) = - \sum_{E \in \mathcal{T}_h} \int_E r_h \operatorname{div}(\mathbf{v}_h) + \sum_{e \in \Gamma_h} \int_e [\mathbf{v}_h]_e \cdot \mathbf{n}_e \{r_h\}_e \quad \forall r_h \in \Theta_h, \quad (5.55)$$

where

$$\Theta_h = \{\theta_h \in L^2(\Omega) : \theta_h|_E \in \mathbb{P}_1 \quad \forall E \in \mathcal{T}_h\}.$$

The polynomial degree one in this space is convenient for proving the convergence of the nonlinear term; see (5.62). The straightforward scaling argument used in proving Proposition 3 shows that

$$\|G_h^{\text{sym}}(\mathbf{v}_h)\|_{L^2(\Omega)} \leq C \|\mathbf{v}_h\|_{V_h} \quad \forall \mathbf{v}_h \in V_h, \quad (5.56)$$

and

$$\|G_h^{\text{div}}(\mathbf{v}_h)\|_{L^2(\Omega)} \leq \left( \sum_{E \in \mathcal{T}_h} \|\operatorname{div}(\mathbf{v}_h)\|_{L^2(E)}^2 \right)^{\frac{1}{2}} + C J_h(\mathbf{v}_h, \mathbf{v}_h)^{\frac{1}{2}},$$

and thus by (5.15)

$$\|G_h^{\text{div}}(\mathbf{v}_h)\|_{L^2(\Omega)} \leq C \|\mathbf{v}_h\|_{V_h} \quad \forall \mathbf{v}_h \in V_h \quad (5.57)$$

with different constants  $C$  independent of  $h$ . At the same time, this gives existence of these two operators. The next proposition relates  $G_h^{\text{sym}}(\mathbf{u}_h)$  and  $\mathbf{D}(\bar{\mathbf{u}})$ . The proof is an easy extension of that written in [16], but we include it below for the reader's convenience.



**Proposition 4.** *Up to a subsequence, we have*

$$\lim_{h \rightarrow 0} G_h^{\text{sym}}(\mathbf{u}_h) = \mathbf{D}(\bar{\mathbf{u}}) \quad \text{weakly in } L^2(\Omega)^{d \times d}. \quad (5.58)$$

*Proof.* On the one hand, the bounds (5.56) and (5.30) imply that there exists a function  $\bar{\mathbf{w}} \in L^2(\Omega)^{d \times d}_{\text{sym}}$  such that, up to a subsequence,

$$\lim_{h \rightarrow 0} G_h^{\text{sym}}(\mathbf{u}_h) = \bar{\mathbf{w}} \quad \text{weakly in } L^2(\Omega)^{d \times d}. \quad (5.59)$$

On the other hand, take any tensor  $\mathbf{F}$  in  $H^1(\Omega)^{d \times d}_{\text{sym}}$  and let  $P_h^0(\mathbf{F})$  be its orthogonal  $L^2(\Omega)^{d \times d}$  projection on constants in each  $E$ . We have

$$\left| \int_{\Omega} G_h^{\text{sym}}(\mathbf{u}_h) : (\mathbf{F} - P_h^0(\mathbf{F})) \right| \leq C \|\mathbf{u}_h\|_{V_h} \|\mathbf{F} - P_h^0(\mathbf{F})\|_{L^2(\Omega)},$$

that tends to zero with  $h$ . Therefore, the definition (5.54) of  $G_h^{\text{sym}}(\mathbf{u}_h)$  implies that

$$\lim_{h \rightarrow 0} \int_{\Omega} G_h^{\text{sym}}(\mathbf{u}_h) : \mathbf{F} = \lim_{h \rightarrow 0} b_{1h}(P_h^0(\mathbf{F}), \mathbf{u}_h) = \lim_{h \rightarrow 0} (b_{1h}(P_h^0(\mathbf{F}) - \mathbf{F}, \mathbf{u}_h) + b_{1h}(\mathbf{F}, \mathbf{u}_h))$$

and a straightforward argument yields that the first term tends to zero. Hence

$$\lim_{h \rightarrow 0} \int_{\Omega} G_h^{\text{sym}}(\mathbf{u}_h) : \mathbf{F} = \lim_{h \rightarrow 0} b_{1h}(\mathbf{F}, \mathbf{u}_h) \quad \forall \mathbf{F} \in H^1(\Omega)^{d \times d}_{\text{sym}}.$$

Now, an application of Green's formula in each  $E$  gives

$$b_{1h}(\mathbf{F}, \mathbf{u}_h) = - \sum_{E \in \mathcal{T}_h} \int_E \mathbf{u}_h \cdot \text{div}(\mathbf{F}).$$

Therefore

$$\lim_{h \rightarrow 0} \int_{\Omega} G_h^{\text{sym}}(\mathbf{u}_h) : \mathbf{F} = - \int_{\Omega} \bar{\mathbf{u}} \cdot \text{div}(\mathbf{F}) = \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \mathbf{F} \quad \forall \mathbf{F} \in H_0^1(\Omega)^{d \times d}_{\text{sym}}.$$

A comparison with (5.59) and uniqueness of the limit yield

$$\mathbf{D}(\bar{\mathbf{u}}) = \bar{\mathbf{w}},$$

thus proving (5.58).  $\square$

**Remark 6.** The fact that  $\bar{\mathbf{u}}$  belongs to  $H_0^1(\Omega)^d$  is an easy consequence of the above proof.

A similar argument to the one in Proposition 4 gives that

$$\lim_{h \rightarrow 0} G_h^{\text{div}}(\mathbf{u}_h) = \text{div}(\bar{\mathbf{u}}) \quad \text{weakly in } L^2(\Omega). \quad (5.60)$$

Hence, by passing to the limit in the last equation of (5.12), we immediately deduce that  $\text{div}(\bar{\mathbf{u}}) = 0$ ; thus  $\bar{\mathbf{u}}$  belongs to  $\mathcal{V}$  and satisfies the third equation of (4.1).

In the next theorem, these results are used to show that the limit satisfies the remaining equations of (4.1).

**Theorem 5.** *Let the family of hexahedra  $\mathcal{T}_h$  be regular in the sense defined above. Then the triple  $(\bar{\mathbf{T}}, \bar{\mathbf{u}}, \bar{p})$  solves (4.1).*

*Proof.* The proof proceeds in two steps.

**Step 1.** Let us start with the first equation of (5.12). Take a function  $\mathbf{v} \in \mathcal{D}(\Omega)^d$  and let  $\mathbf{v}_h \in V_h$  be the  $L^2(\Omega)^d$  orthogonal projection of  $\mathbf{v}$  on  $\mathbb{P}_1^d$  in each element. It is easy to check that

$$\lim_{h \rightarrow 0} G_h^{\text{sym}}(\mathbf{v}_h) = \mathbf{D}(\mathbf{v}) \quad \text{strongly in } L^2(\Omega)^{d \times d}.$$

Therefore the weak convergence of  $\mathbf{T}_h$  and the definition of  $G_h^{\text{sym}}(\mathbf{v}_h)$  imply that

$$\lim_{h \rightarrow 0} b_{1h}(\mathbf{T}_h, \mathbf{v}_h) = \int_{\Omega} \mathbf{D}(\mathbf{v}) : \bar{\mathbf{T}}.$$

Similarly,

$$\lim_{h \rightarrow 0} b_{2h}(\mathbf{v}_h, p_h) = - \int_{\Omega} \bar{p} \operatorname{div}(\mathbf{v}).$$

Also

$$\lim_{h \rightarrow 0} J_h(\mathbf{u}_h, \mathbf{v}_h) = 0.$$

As the right-hand side tends to  $\int_{\Omega} \mathbf{f} \cdot \mathbf{v}$ , it remains to examine  $d_h(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h)$ . Recall that

$$d_h(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) = \sum_{E \in \mathcal{T}_h} \int_E [(\mathbf{u}_h \cdot \nabla) \mathbf{u}_h] \cdot \mathbf{v}_h - \frac{1}{2} b_{2h}(\mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{v}_h) - \sum_{e \in \Gamma_h^i} \int_e \{\mathbf{u}_h\}_e \cdot \mathbf{n}_e [\mathbf{u}_h]_e \cdot \{\mathbf{v}_h\}_e.$$

Thanks to the antisymmetry of  $d_h$ , we have

$$d_h(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) = - \sum_{E \in \mathcal{T}_h} \int_E [(\mathbf{u}_h \cdot \nabla) \mathbf{v}_h] \cdot \mathbf{u}_h + \frac{1}{2} b_{2h}(\mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{v}_h) + \sum_{e \in \Gamma_h^i} \int_e \{\mathbf{u}_h\}_e \cdot \mathbf{n}_e [\mathbf{v}_h]_e \cdot \{\mathbf{u}_h\}_e. \quad (5.61)$$

For the first term, the strong convergence of  $\mathbf{u}_h$  in  $L^4(\Omega)^d$  and the strong convergence of the broken gradient  $\nabla_h \mathbf{v}_h$  in  $L^2(\Omega)^{d \times d}$  imply that

$$- \lim_{h \rightarrow 0} \sum_{E \in \mathcal{T}_h} \int_E [(\mathbf{u}_h \cdot \nabla) \mathbf{v}_h] \cdot \mathbf{u}_h = - \int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \mathbf{v}] \cdot \bar{\mathbf{u}} = \int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}}] \cdot \mathbf{v},$$

since  $\bar{\mathbf{u}} \in \mathcal{V}$ . For the second term, take any piecewise constant approximation  $\bar{\mathbf{v}}_h$  of  $\mathbf{v}$ . Then

$$b_{2h}(\mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{v}_h) = b_{2h}(\mathbf{u}_h, \mathbf{u}_h \cdot (\mathbf{v}_h - \bar{\mathbf{v}}_h)) + b_{2h}(\mathbf{u}_h, \mathbf{u}_h \cdot \bar{\mathbf{v}}_h).$$

The boundedness of  $\mathbf{u}_h$  in  $V_h$  and the convergence to zero of  $\mathbf{v}_h - \bar{\mathbf{v}}_h$  in  $L^\infty(\Omega)^d$  imply that the first term tends to zero. For the second term, we deduce from the definition of  $G_h^{\text{div}}(\mathbf{u}_h)$  that

$$b_{2h}(\mathbf{u}_h, \mathbf{u}_h \cdot \bar{\mathbf{v}}_h) = \int_{\Omega} G_h^{\text{div}}(\mathbf{u}_h)(\mathbf{u}_h \cdot \bar{\mathbf{v}}_h). \quad (5.62)$$

As  $\operatorname{div}(\bar{\mathbf{u}}) = 0$ ,  $G_h^{\text{div}}(\mathbf{u}_h)$  tends to zero weakly in  $L^2(\Omega)$ . Then the strong convergence of  $\mathbf{u}_h$  in  $L^2(\Omega)^d$  and that of  $\bar{\mathbf{v}}_h$  in  $L^\infty(\Omega)^d$  show that this second term tends to zero. It remains to examine the last term of (5.61). Here we use the fact that, for any  $\mathbf{v} \in W^{2,\infty}(\Omega)^d$ ,

$$\|\mathbf{v}_h - \mathbf{v}\|_{L^\infty(e)} \leq Ch_e^2 |\mathbf{v}|_{W^{2,\infty}(\Omega)}.$$

This, with the boundedness of  $\mathbf{u}_h$  in  $V_h$ , gives that this last term tends to zero. Thus, we conclude that

$$\lim_{h \rightarrow 0} d_h(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) = \int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}}] \cdot \mathbf{v} \quad \forall \mathbf{v} \in W^{2,\infty}(\Omega)^d \cap H_0^1(\Omega)^d.$$

The conclusion of these limits and a density argument is that the triple  $(\bar{\mathbf{T}}, \bar{\mathbf{u}}, \bar{p})$  satisfies the first equation of (4.1)

$$\int_{\Omega} [(\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}}] \cdot \mathbf{v} + \int_{\Omega} \bar{\mathbf{T}} : \mathbf{D}(\mathbf{v}) - \int_{\Omega} \bar{p} \operatorname{div}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \quad \forall \mathbf{v} \in H_0^1(\Omega)^d. \quad (5.63)$$

**Step 2.** The argument for recovering the constitutive relation  $\bar{\mathbf{T}} = \mathcal{G}(\bar{\mathbf{u}})$  is close to that for the conforming case, up to some changes. On the one hand, we observe that

$$\lim_{h \rightarrow 0} (b_{1h}(\mathbf{T}_h, \mathbf{u}_h) + J_h(\mathbf{u}_h, \mathbf{u}_h)) = \int_{\Omega} \mathbf{f} \cdot \bar{\mathbf{u}}$$

and, since  $J_h(\mathbf{u}_h, \mathbf{u}_h)$  is positive and bounded, this implies that

$$\lim_{h \rightarrow 0} b_{1h}(\mathbf{T}_h, \mathbf{u}_h) \leq \int_{\Omega} \mathbf{f} \cdot \bar{\mathbf{u}}.$$

On the other hand, we infer from (5.63) that

$$\int_{\Omega} \bar{\mathbf{T}} : \mathbf{D}(\bar{\mathbf{u}}) = \int_{\Omega} \mathbf{f} \cdot \bar{\mathbf{u}}.$$

Hence

$$\lim_{h \rightarrow 0} b_{1h}(\mathbf{T}_h, \mathbf{u}_h) \leq \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \bar{\mathbf{T}}. \quad (5.64)$$

Next, we set

$$\tilde{\mathbf{T}}^{\mathbf{d}} = \mathcal{G}(\bar{\mathbf{u}})$$

and define  $\tilde{\mathbf{T}}_h = \mathcal{G}_{h,DG}(\bar{\mathbf{u}})$ , i.e.,

$$\alpha \int_{\Omega} \tilde{\mathbf{T}}_h : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\tilde{\mathbf{T}}_h|) \tilde{\mathbf{T}}_h : \mathbf{S}_h = b_{1h}(\mathbf{S}_h, \bar{\mathbf{u}}) = \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \mathbf{S}_h \quad \forall \mathbf{S}_h \in M_h,$$

where the second equality holds thanks to the fact that  $\bar{\mathbf{u}}$  belongs to  $H_0^1(\Omega)^d$ . The fact that  $\operatorname{div}(\bar{\mathbf{u}}) = 0$  implies that the trace of  $\tilde{\mathbf{T}}^{\mathbf{d}}$  is zero and justifies the above superscript. Therefore

$$\alpha \int_{\Omega} (\tilde{\mathbf{T}}_h - \tilde{\mathbf{T}}^{\mathbf{d}}) : \mathbf{S}_h + \gamma \int_{\Omega} (\mu(|\tilde{\mathbf{T}}_h|) \tilde{\mathbf{T}}_h - \mu(|\tilde{\mathbf{T}}^{\mathbf{d}}|) \tilde{\mathbf{T}}^{\mathbf{d}}) : \mathbf{S}_h = 0 \quad \forall \mathbf{S}_h \in M_h,$$

and, as in the conforming case, we conclude that

$$\lim_{h \rightarrow 0} \|\tilde{\mathbf{T}}_h - \tilde{\mathbf{T}}^{\mathbf{d}}\|_{L^2(\Omega)} = 0. \quad (5.65)$$

Finally, the difference between the equations satisfied by  $\mathbf{T}_h$  and  $\tilde{\mathbf{T}}_h$  yields

$$\alpha \int_{\Omega} (\mathbf{T}_h - \tilde{\mathbf{T}}_h) : \mathbf{S}_h + \gamma \int_{\Omega} (\mu(|\mathbf{T}_h|) \mathbf{T}_h - \mu(|\tilde{\mathbf{T}}_h|) \tilde{\mathbf{T}}_h) : \mathbf{S}_h = b_{1h}(\mathbf{S}_h, \mathbf{u}_h) - \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \mathbf{S}_h \quad \forall \mathbf{S}_h \in M_h.$$

By testing this equation with  $\mathbf{S}_h = \mathbf{T}_h - \tilde{\mathbf{T}}_h$  and using the monotonicity property (2.14), we deduce that

$$\alpha \|\mathbf{T}_h - \tilde{\mathbf{T}}_h\|_{L^2(\Omega)}^2 \leq b_{1h}(\mathbf{T}_h, \mathbf{u}_h) - b_{1h}(\tilde{\mathbf{T}}_h, \mathbf{u}_h) - \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : (\mathbf{T}_h - \tilde{\mathbf{T}}_h). \quad (5.66)$$

However, by (5.54),

$$b_{1h}(\tilde{\mathbf{T}}_h, \mathbf{u}_h) = \int_{\Omega} G_h^{\text{sym}}(\mathbf{u}_h) : \tilde{\mathbf{T}}_h,$$

and it follows from Proposition 4 and (5.65) that

$$\lim_{h \rightarrow 0} b_{1h}(\tilde{\mathbf{T}}_h, \mathbf{u}_h) = \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \tilde{\mathbf{T}}^{\mathbf{d}}.$$

Then, by passing to the limit in (5.66), we obtain in view of (5.64) the inequality

$$\alpha \lim_{h \rightarrow 0} \|\mathbf{T}_h - \tilde{\mathbf{T}}_h\|_{L^2(\Omega)}^2 \leq \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \bar{\mathbf{T}} - \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : \tilde{\mathbf{T}}^{\mathbf{d}} - \int_{\Omega} \mathbf{D}(\bar{\mathbf{u}}) : (\bar{\mathbf{T}} - \tilde{\mathbf{T}}^{\mathbf{d}}) = 0,$$

whence

$$\lim_{h \rightarrow 0} \|\mathbf{T}_h - \tilde{\mathbf{T}}_h\|_{L^2(\Omega)} = 0,$$

and uniqueness of the limit yields

$$\bar{\mathbf{T}} = \tilde{\mathbf{T}}^{\mathbf{d}} = \mathcal{G}(\bar{\mathbf{u}}).$$

This proves that  $(\bar{\mathbf{T}}, \bar{\mathbf{u}})$  satisfies the second equation of (4.1).  $\square$

## 5.2 The tetrahedral case

Here we study briefly two examples of finite element discretisations on tetrahedral meshes, the triangular case being simpler. Many of the details are skipped because they follow closely those in the previous subsection. The family of meshes  $\mathcal{T}_h$  is assumed to be regular as in (4.24). Let us start with the same spaces  $V_h$ ,  $Q_h$ , and  $M_h$  defined on  $\mathcal{T}_h$  by (5.1), (5.2), and (5.3), respectively, and the same bilinear forms  $b_{1h}(\mathbf{S}_h, \mathbf{v}_h)$ ,  $d_h(\mathbf{u}_h; \mathbf{v}_h, \mathbf{w}_h)$ , and  $b_{2h}(\mathbf{v}_h, q_h)$  defined by (5.4), (5.5), and (5.6), respectively. Then the scheme is again given by (5.12) and, under assumption (4.24), all proofs from the previous subsections are valid in this case, except possibly the proof of the inf-sup condition. In fact, Theorem 4 holds with a much simpler proof. Indeed, take any tetrahedron  $E$ . Recalling that the case of interest is  $k = 1$ , a polynomial of  $\mathbb{P}_1$  is uniquely determined in  $E$  by its values at the centre points  $\mathbf{b}_e$  of its four faces  $e$ . Then, instead of (5.36), we can use the sufficient condition

$$\mathbf{c}_h(\mathbf{b}_e)|_E = \frac{1}{|e|} \int_e (R_h(\mathbf{v}) - \mathbf{v})|_E \quad \forall E \in \mathcal{T}_h, \quad \forall e \in \partial E, \quad (5.67)$$

and this defines uniquely the correction  $\mathbf{c}_h$ . Furthermore, thanks to (4.24), the stability of this correction follows from the fact that  $E$  is the image of the unit tetrahedron  $\hat{E}$  by an invertible affine mapping whose matrix satisfies the same properties as the matrix  $\mathbf{B}$  used above. Thus the conclusion of Theorem 4 is valid in this case.

As a second example, it would be tempting to use the Crouzeix–Raviart element of degree one on tetrahedra; see [15]. This would be possible if the analysis did not invoke Korn’s inequality (with respect to the broken symmetric gradient), because it is not satisfied by the Crouzeix–Raviart element; cf [18]. Thus, the simplest way to bypass this difficulty is to introduce the jump penalty term  $J_h(\mathbf{u}_h, \mathbf{v}_h)$  defined in (5.8). Let us describe this discretisation. Again, we suppose that (4.24) holds. The discrete spaces  $Q_h$  and  $M_h$  are the same, with  $k = 1$ , as in (5.2) and (5.3), respectively. However, instead of  $V_h$ , we now use the space  $V_h^{CR}$  whose elements are also piecewise polynomials of degree one in each element, but in contrast with (5.1), they are continuous at the centre points of all interior faces  $e \in \Gamma_h^i$ , and are set to zero at the centre points of all boundary faces  $e \in \Gamma_h^b$ . Thanks to this pointwise continuity and boundary condition, the scheme now involves the following bilinear/trilinear forms, compare with (5.4), (5.5), (5.6):

$$\int_{\Omega} \mathbf{D}(\mathbf{v}) : \mathbf{S} \simeq b_{1h}^{CR}(\mathbf{S}_h, \mathbf{v}_h) := \sum_{E \in \mathcal{T}_h} \int_E \mathbf{D}(\mathbf{v}_h) : \mathbf{S}_h, \quad (5.68)$$

$$d_h^{CR}(\mathbf{u}_h; \mathbf{v}_h, \mathbf{w}_h) := \frac{1}{2} \left[ \sum_{E \in \mathcal{T}_h} \int_E [(\mathbf{u}_h \cdot \nabla) \mathbf{v}_h] \cdot \mathbf{w}_h - \sum_{E \in \mathcal{T}_h} \int_E [(\mathbf{u}_h \cdot \nabla) \mathbf{w}_h] \cdot \mathbf{v}_h \right], \quad (5.69)$$

$$b_{2h}^{CR}(\mathbf{v}_h, q_h) := - \sum_{E \in \mathcal{T}_h} \int_E q_h \operatorname{div}(\mathbf{v}_h). \quad (5.70)$$

With these new forms, analogously to (5.12), the finite element approximation of the problem reads as follows: find a triple  $(\mathbf{T}_h, \mathbf{u}_h, p_h) \in M_h \times V_h^{CR} \times Q_h$  such that

$$\begin{aligned} d_h^{CR}(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + b_{1h}^{CR}(\mathbf{T}_h, \mathbf{v}_h) + b_{2h}^{CR}(\mathbf{v}_h, p_h) + J_h(\mathbf{u}_h, \mathbf{v}_h) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h & \forall \mathbf{v}_h \in V_h^{CR}, \\ \alpha \int_{\Omega} \mathbf{T}_h : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\mathbf{T}_h|) \mathbf{T}_h : \mathbf{S}_h &= b_{1h}^{CR}(\mathbf{S}_h, \mathbf{u}_h) & \forall \mathbf{S}_h \in M_h, \\ b_{2h}^{CR}(\mathbf{u}_h, q_h) &= 0 & \forall q_h \in Q_h. \end{aligned} \quad (5.71)$$

Note that  $b_{1h}^{CR}(\mathbf{S}_h, \mathbf{v}_h)$  coincides with  $b_{1h}(\mathbf{S}_h, \mathbf{v}_h)$  and  $b_{2h}^{CR}(\mathbf{v}_h, q_h)$  coincides with  $b_{2h}(\mathbf{v}_h, q_h)$  because the additional face terms vanish for elements of the space  $V_h^{CR}$ . This is not necessarily the case with  $d_h^{CR}$  and  $d_h$ , but  $d_h^{CR}$  is obviously antisymmetric and is simpler. Although the norm of the broken gradient is a norm on  $V_h^{CR}$ , the mapping  $\mathbf{v}_h \mapsto \|\mathbf{D}(\mathbf{v}_h)\|_h$  is not a norm on  $V_h^{CR}$ . According to [6, 7], we have instead (5.14) and (5.15). That is why we use again the norm  $\|\mathbf{v}_h\|_{V_h}$  defined in (5.9) and keep the term  $J_h(\mathbf{u}_h, \mathbf{v}_h)$  in the first line of (5.71). Note however that the parameters  $\sigma_e$  need not be tuned by Proposition 3 since there

are no surface terms in  $b_{1h}^{CR}(\mathbf{T}_h, \mathbf{v}_h)$ ; thus it suffices for instance to take  $\sigma_e = 1$  for each face  $e$ . Moreover, the analysis used for the general discontinuous elements substantially simplifies here. First, as there are no surface terms in the bilinear forms, the bounds are simpler. Next, the operator  $\Pi_h$  satisfying the statement of Lemma 6 is constructed directly by setting, for  $\mathbf{v}$  in  $H_0^1(\Omega)^d$ ,

$$\Pi_h(\mathbf{v})(\mathbf{b}_e)|_E = \frac{1}{|e|} \int_e \mathbf{v} \quad \forall E \in \mathcal{T}_h, \forall e \in \partial E, \quad (5.72)$$

see [15]. Clearly, as  $\mathbf{v} \in H_0^1(\Omega)^d$ , (5.72) defines a piecewise polynomial function of degree one in  $V_h^{CR}$ . Finally, convergence of the scheme is derived without the discrete differential operators  $G_h^{\text{sym}}$  and  $G_h^{\text{div}}$ . Indeed, property (5.17) can be extended as is asserted in the proposition.

**Proposition 5.** *Let the family  $\mathcal{T}_h$  satisfy (4.24). If  $\mathbf{v}_h$  is a sequence in  $V_h^{CR}$  such that*

$$\|\mathbf{v}_h\|_{V_h} \leq C$$

*with a constant  $C$  independent of  $h$ , then there exists a function  $\bar{\mathbf{v}} \in H_0^1(\Omega)^d$  satisfying (5.17) and*

$$\lim_{h \rightarrow 0} \mathbf{D}_h(\mathbf{v}_h) = \mathbf{D}(\bar{\mathbf{v}}) \quad \text{weakly in } L^2(\Omega)^{d \times d}, \quad (5.73)$$

*where  $\mathbf{D}_h$  stands for the broken symmetric gradient.*

The proof, contained in [15], relies on the fact that the integral average of the jump  $[\mathbf{v}_h]_e$  vanishes on any face  $e$  and hence, for any tensor  $\mathbf{F}$  in  $H^1(\Omega)^{d \times d}$ ,

$$\int_e \mathbf{F} \mathbf{n}_e \cdot [\mathbf{v}_h]_e = \int_e (\mathbf{F} - \mathbf{C}) \mathbf{n}_e \cdot [\mathbf{v}_h]_e \quad \forall \mathbf{C} \in \mathbb{R}^{d \times d}.$$

Thus, there is no need for  $G_h^{\text{sym}}$ ; the same is true for  $G_h^{\text{div}}$ . This permits to pass directly to the limit in (5.71).

## 6 Numerical illustrations

We introduce two decoupled iterative algorithms. The first one is based on a Lions–Mercier decoupling strategy while the second one is a fixed point algorithm. All the algorithms are implemented using the *deall.i* library [1]. For simplicity, we focus on conforming finite element approximations for which an a priori error estimate has been derived in Subsection 4.1.2. Performing numerical experiments in the case of the nonconforming approximation scheme will be the subject of future work.

The general setup is the following:

- Dirichlet boundary conditions are imposed on the entire domain boundary (not necessarily homogeneous);
- A sequence of uniformly refined meshes with square elements of diameter  $h = \sqrt{2}/2^n$ ,  $n = 2, \dots, 6$  (level of refinement) are considered for the mesh refinement analysis;
- The finite element spaces  $M_h$ ,  $V_h$ , and  $Q_h$  consist, respectively, of discontinuous piecewise polynomials of degree 2, continuous piecewise polynomials of degree 2, and continuous piecewise polynomials of degree 1 (see Subsection 4.2.2).

Following [5], we replace the constitutive relation

$$\alpha \mathbf{T}^d + \gamma \mu(|\mathbf{T}^d|) \mathbf{T}^d - \mathbf{D}(\mathbf{u}) = \mathbf{0}$$

by

$$\alpha \mathbf{T}^d + \gamma \mu(|\mathbf{T}^d|) \mathbf{T}^d - \mathbf{D}(\mathbf{u}) = \mathbf{g}$$

to design an exact solution. Then, given  $\mathbf{T}^d$ ,  $\mathbf{u}$  and  $p$ , we compute the corresponding right-hand sides  $\mathbf{g}$  and  $\mathbf{f}$  (forcing term), where we recall that

$$\mathbf{f} = (\mathbf{u} \cdot \nabla) \mathbf{u} - \frac{1}{\alpha} \operatorname{div}(\mathbf{D}(\mathbf{u})) + \nabla p + \frac{\gamma}{\alpha} \operatorname{div}(\mu(|\mathbf{T}^d|) \mathbf{T}^d).$$

Finally, we choose  $\mu(s) = \frac{1}{\sqrt{1+s^2}}$  which corresponds to (1.10) with  $\beta = 1$  and  $n = -1/2$ .

## 6.1 Lions–Mercier decoupled iterative algorithm

We present here an iterative algorithm to compute approximately the solution to problem (3.5), which is based on the formulation (3.4): find  $(\mathbf{T}_h, \mathbf{u}_h, p_h) \in M_h \times V_h \times Q_h$  such that

$$\begin{aligned} d(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\mathbf{u}_h) : \mathbf{D}(\mathbf{v}_h) - \int_{\Omega} p_h \operatorname{div}(\mathbf{v}_h) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}_h|) \mathbf{T}_h : \mathbf{D}(\mathbf{v}_h), \\ \alpha \int_{\Omega} \mathbf{T}_h : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\mathbf{T}_h|) \mathbf{T}_h : \mathbf{S}_h &= \int_{\Omega} \mathbf{D}(\mathbf{u}_h) : \mathbf{S}_h, \\ \int_{\Omega} q_h \operatorname{div}(\mathbf{u}_h) &= 0 \end{aligned} \quad (6.1)$$

for all  $(\mathbf{S}_h, \mathbf{v}_h, q_h) \in M_h \times V_h \times Q_h$ , where  $d : V \times V \times V \rightarrow \mathbb{R}$  is defined in (4.7). Note that problem (6.1) is equivalent to problem (4.9) analysed in Section 4.

To compute the solution to problem (6.1), we propose a decoupled algorithm based on a Lions–Mercier splitting algorithm [25] (alternating-direction method of the Peaceman–Rachford type [31]) applied to the unknown  $\mathbf{T}_h$ . Following the discussion in [5, Section 7], the algorithm reads, for a pseudo-time step  $\tau > 0$ :

*Initialisation:* find  $(\mathbf{T}_h^{(0)}, \mathbf{u}_h^{(0)}, p_h^{(0)}) \in M_h \times V_h \times Q_h$  such that

$$\begin{aligned} d(\mathbf{u}_h^{(0)}; \mathbf{u}_h^{(0)}, \mathbf{v}_h) + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\mathbf{u}_h^{(0)}) : \mathbf{D}(\mathbf{v}_h) - \int_{\Omega} p_h^{(0)} \operatorname{div}(\mathbf{v}_h) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h & \forall \mathbf{v}_h \in V_h, \\ \alpha \int_{\Omega} \mathbf{T}_h^{(0)} : \mathbf{S}_h &= \int_{\Omega} \mathbf{D}(\mathbf{u}_h^{(0)}) : \mathbf{S}_h & \forall \mathbf{S}_h \in M_h, \\ \int_{\Omega} q_h \operatorname{div}(\mathbf{u}_h^{(0)}) &= 0 & \forall q_h \in Q_h. \end{aligned} \quad (6.2)$$

Then, for  $k = 0, 1, \dots$ , perform the following two steps:

*Step 1:* Find  $\mathbf{T}_h^{(k+\frac{1}{2})} \in M_h$  such that

$$\frac{1}{\tau} \int_{\Omega} (\mathbf{T}_h^{(k+\frac{1}{2})} - \mathbf{T}_h^{(k)}) : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\mathbf{T}_h^{(k+\frac{1}{2})}|) \mathbf{T}_h^{(k+\frac{1}{2})} : \mathbf{S}_h = \int_{\Omega} \mathbf{D}(\mathbf{u}_h^{(k)}) : \mathbf{S}_h - \alpha \int_{\Omega} \mathbf{T}_h^{(k)} : \mathbf{S}_h \quad \forall \mathbf{S}_h \in M_h.$$

*Step 2:* Find  $(\mathbf{T}_h^{(k+1)}, \mathbf{u}_h^{(k+1)}, p_h^{(k+1)}) \in M_h \times V_h \times Q_h$  such that

$$\begin{aligned} d(\mathbf{u}_h^{(k+1)}; \mathbf{u}_h^{(k+1)}, \mathbf{v}_h) + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\mathbf{u}_h^{(k+1)}) : \mathbf{D}(\mathbf{v}_h) - \int_{\Omega} p_h^{(k+1)} \operatorname{div}(\mathbf{v}_h) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}_h^{(k+\frac{1}{2})}|) \mathbf{T}_h^{(k+\frac{1}{2})} : \mathbf{D}(\mathbf{v}_h), \\ \frac{1}{\tau} \int_{\Omega} (\mathbf{T}_h^{(k+1)} - \mathbf{T}_h^{(k+\frac{1}{2})}) : \mathbf{S}_h + \alpha \int_{\Omega} \mathbf{T}_h^{(k+1)} : \mathbf{S}_h &= \int_{\Omega} \mathbf{D}(\mathbf{u}_h^{(k+1)}) : \mathbf{S}_h - \gamma \int_{\Omega} \mu(|\mathbf{T}_h^{(k+\frac{1}{2})}|) \mathbf{T}_h^{(k+\frac{1}{2})} : \mathbf{S}_h, \\ \int_{\Omega} q_h \operatorname{div}(\mathbf{u}_h^{(k+1)}) &= 0 \end{aligned} \quad (6.3)$$

for all  $(\mathbf{S}_h, \mathbf{v}_h, q_h) \in M_h \times V_h \times Q_h$ .

The solution to (6.2) is obtained by first determining  $\mathbf{u}_h^{(0)}$  and  $p_h^{(0)}$  as the solution to a standard steady-state Navier–Stokes equation (first and third equations in (6.2)) and then by setting  $\mathbf{T}_h^{(0)} = \frac{1}{\alpha} \mathbf{D}(\mathbf{u}_h^{(0)})$ . Similarly, the solution to problem (6.3) can be obtained by first solving the first and third equations for  $\mathbf{u}_h^{(k+1)}$  and  $p_h^{(k+1)}$  and then solving the second equation for  $\mathbf{T}_h^{(k+1)}$ . A standard argument shows that the above algorithm generates uniformly bounded sequences. Thus they converge up to subsequences. However, the identification of a unique limit for the entire sequence is currently unclear.

Regarding the implementation, we make the following comments:

- *Stopping criterion:* For the main loop (Lions–Mercier algorithm), the stopping criterion is

$$\frac{\|\mathbf{T}_h^{(k+1)} - \mathbf{T}_h^{(k)}\|_{L^2(\Omega)} + \|\nabla(\mathbf{u}_h^{(k+1)} - \mathbf{u}_h^{(k)})\|_{L^2(\Omega)} + \|p_h^{(k+1)} - p_h^{(k)}\|_{L^2(\Omega)}}{\|\mathbf{T}_h^{(k+1)}\|_{L^2(\Omega)} + \|\nabla \mathbf{u}_h^{(k+1)}\|_{L^2(\Omega)} + \|p_h^{(k+1)}\|_{L^2(\Omega)}} \leq 10^{-5}; \quad (6.4)$$

- *Initialisation:* We solve the Navier–Stokes system associated to problem (6.2) using Newton’s method (the iterates are indexed by  $m$ ) until the following stopping criterion is met:

$$\frac{\|\nabla(\mathbf{u}_h^{(m+1)} - \mathbf{u}_h^{(m)})\|_{L^2(\Omega)} + \|p_h^{(m+1)} - p_h^{(m)}\|_{L^2(\Omega)}}{\|\nabla \mathbf{u}_h^{(m+1)}\|_{L^2(\Omega)} + \|p_h^{(m+1)}\|_{L^2(\Omega)}} \leq 10^{-6}.$$

As an initial guess, we take the solution of the associated Stokes system without the convective term.

The solution to each saddle-point system of the form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{P} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{G} \end{pmatrix}$$

is obtained using a Schur complement formulation

$$BA^{-1}B^T\mathbf{P} = BA^{-1}\mathbf{F} - \mathbf{G}, \quad A\mathbf{U} = \mathbf{F} - B^T\mathbf{P}.$$

To solve for  $\mathbf{P}$ , we use the conjugate gradient algorithm in the case of the Stokes problem and GMRES for the (linearised) Navier–Stokes problems. In both cases, the pressure mass matrix is used as preconditioner and the tolerance for the iterative algorithm is set to  $10^{-6}\|BA^{-1}\mathbf{F} - \mathbf{G}\|_{\ell_2}$ . A direct method is advocated for every occurrence of  $A^{-1}$  and also to obtain  $\mathbf{T}^{(0)}$ .

- *Step 1 (monotone part):*  $\mathbf{T}_h^{(k+\frac{1}{2})}$  is the zero of the functional

$$F(\mathbf{T}_h) := \mathbf{T}_h + \tau\gamma\mu(|\mathbf{T}_h|)\mathbf{T}_h - \tau\mathbf{D}(\mathbf{u}_h^{(k)}) - (1 - \alpha\tau)\mathbf{T}_h^{(k)}.$$

Recall that discontinuous piecewise polynomial approximations are used for the stress and so  $\mathbf{T}_h$  is determined locally on each element  $E \in \mathcal{T}_h$  as the solution to

$$\int_E F(\mathbf{T}_h) : \mathbf{S}_h = 0 \quad \forall \mathbf{S}_h \in \mathbb{Q}_2.$$

We again employ Newton’s method starting with  $\mathbf{T}_h^{(0)} = \mathbf{T}_h^{(k)}|_E$  and use the stopping criterion

$$\|F(\mathbf{T}_h^{(m)})\|_{L^2(E)} \leq 10^{-6} \frac{\sqrt{|E|}}{\sqrt{|\Omega|}}$$

so that the global residual is less than  $10^{-6}$ . Note that in this case, it might happen that no iteration is needed (e.g. when  $\gamma = 0$ ), in which case  $\mathbf{T}_h^{(k+\frac{1}{2})}|_E = \mathbf{T}_h^{(k)}|_E$ .

- *Step 2:* This step is similar to the initialisation step except that we take  $(\mathbf{u}_h^{(k)}, p_h^{(k)})$  as our initial guess for Newton’s method for solving the finite element approximation of the Navier–Stokes system.

### 6.1.1 Case 1: smooth solution

We consider the case  $\Omega = (0, 1)^2$  and

$$\mathbf{T}^d = \begin{pmatrix} \frac{\cos(2\pi x) - \cos(2\pi y)}{4} & 0 \\ 0 & \frac{\cos(2\pi y) - \cos(2\pi x)}{4} \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} -\cos(\pi x) \sin(\pi y) \\ \sin(\pi x) \cos(\pi y) \end{pmatrix}, \quad p = -\frac{\cos(2\pi x) + \cos(2\pi y)}{4}.$$

Note that  $\mathbf{T}^d$  is the deviatoric part of  $\mathbf{T}$  defined by

$$\mathbf{T} = \begin{pmatrix} \frac{\cos(2\pi x)}{2} & 0 \\ 0 & \frac{\cos(2\pi y)}{2} \end{pmatrix},$$

and in particular it has vanishing trace. We observe that  $\mathbf{u}$  is divergence-free. Moreover, the pressure satisfies  $p = -\frac{1}{2}\text{tr}(\mathbf{T})$  and has zero mean. We report in Table 1 the error for each component of the solution for the case  $\alpha = 1$  and  $\gamma = 0$ , while Table 2 contains the results for  $\alpha = \gamma = 1$ . Note that we use the  $H^1$  semi-norm for the velocity and not the (equivalent)  $L^2(\Omega)^{2 \times 2}$  norm of the symmetric gradient. We observe in Table 2

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$6.04199 \times 10^{-2}$	$7.51266 \times 10^{-2}$	$3.02263 \times 10^{-2}$	1
3	0.177	$1.44750 \times 10^{-2}$	$1.82293 \times 10^{-2}$	$6.18331 \times 10^{-3}$	1
4	0.088	$3.58096 \times 10^{-3}$	$4.52460 \times 10^{-3}$	$1.46371 \times 10^{-3}$	1
5	0.044	$8.92901 \times 10^{-4}$	$1.12913 \times 10^{-3}$	$3.60874 \times 10^{-4}$	1
6	0.022	$2.23079 \times 10^{-4}$	$2.82155 \times 10^{-4}$	$8.99041 \times 10^{-5}$	1

Table 1: Case 1,  $\alpha = 1$ ,  $\gamma = 0$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.01$ .

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.57579 \times 10^{-2}$	$8.21275 \times 10^{-2}$	$3.01953 \times 10^{-2}$	183
3	0.177	$7.78829 \times 10^{-3}$	$1.86706 \times 10^{-2}$	$6.18695 \times 10^{-3}$	182
4	0.088	$2.00882 \times 10^{-3}$	$4.55378 \times 10^{-3}$	$1.50017 \times 10^{-3}$	182
5	0.044	$8.86597 \times 10^{-4}$	$1.13687 \times 10^{-3}$	$4.91418 \times 10^{-4}$	182
6	0.022	$7.66438 \times 10^{-4}$	$3.05389 \times 10^{-4}$	$3.45733 \times 10^{-4}$	182

Table 2: Case 1,  $\alpha = \gamma = 1$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.01$ .

that all three errors are  $\mathcal{O}(h^2)$ . The deterioration of the convergence rate we observe for  $\mathbf{T}^d$  and  $p$  in Table 2 is due to the stopping criterion. Indeed, if we use  $10^{-6}$  instead of  $10^{-5}$  in the stopping criterion (6.4) for the main loop, then for  $h = 0.044$  ( $n = 5$ ) we need 250 iterations and we get

$$\|\mathbf{T}^d - \mathbf{T}_h\|_{L^2(\Omega)} = 4.66898 \times 10^{-4}, \quad \|\mathbf{u} - \mathbf{u}_h\|_{L^2(\Omega)} = 1.13118 \times 10^{-3} \quad \text{and} \quad \|p - p_h\|_{L^2(\Omega)} = 3.62581 \times 10^{-4},$$

compare with the fourth row of Table 2.

We give in Tables 3, 4 and 5 the results obtained when a larger pseudo-time step is used. We see that

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.57161 \times 10^{-2}$	$8.21200 \times 10^{-2}$	$3.02043 \times 10^{-2}$	47
3	0.177	$7.74372 \times 10^{-3}$	$1.86697 \times 10^{-2}$	$6.18194 \times 10^{-3}$	47
4	0.088	$1.86276 \times 10^{-3}$	$4.55240 \times 10^{-3}$	$1.46465 \times 10^{-3}$	47
5	0.044	$4.77556 \times 10^{-4}$	$1.13167 \times 10^{-3}$	$3.65463 \times 10^{-4}$	47
6	0.022	$1.72916 \times 10^{-4}$	$2.85469 \times 10^{-4}$	$1.06995 \times 10^{-4}$	47

Table 3: Case 1,  $\alpha = \gamma = 1$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.05$ .

the larger the pseudo-time step, the fewer the number of iterations. Moreover, for all cases  $\tau = 0.05$ ,  $\tau = 0.1$  and  $\tau = 0.5$ , there is no deterioration of the convergence rate in contrast to what we observed in Table 2 (due to the stopping criterion).



$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.57060 \times 10^{-2}$	$8.21133 \times 10^{-2}$	$3.02054 \times 10^{-2}$	26
3	0.177	$7.74150 \times 10^{-3}$	$1.86693 \times 10^{-2}$	$6.18213 \times 10^{-3}$	26
4	0.088	$1.85887 \times 10^{-3}$	$4.55234 \times 10^{-3}$	$1.46384 \times 10^{-3}$	26
5	0.044	$4.63124 \times 10^{-4}$	$1.13153 \times 10^{-3}$	$3.61829 \times 10^{-4}$	26
6	0.022	$1.27963 \times 10^{-4}$	$2.84928 \times 10^{-4}$	$9.37426 \times 10^{-5}$	26

Table 4: Case 1,  $\alpha = \gamma = 1$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.1$ .

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.57028 \times 10^{-2}$	$8.21057 \times 10^{-2}$	$3.02063 \times 10^{-2}$	10
3	0.177	$7.73342 \times 10^{-3}$	$1.86606 \times 10^{-2}$	$6.18238 \times 10^{-3}$	7
4	0.088	$1.85742 \times 10^{-3}$	$4.55172 \times 10^{-3}$	$1.46368 \times 10^{-3}$	7
5	0.044	$4.59753 \times 10^{-4}$	$1.13121 \times 10^{-3}$	$3.60876 \times 10^{-4}$	7
6	0.022	$1.15437 \times 10^{-4}$	$2.83829 \times 10^{-4}$	$8.99203 \times 10^{-5}$	7

Table 5: Case 1,  $\alpha = \gamma = 1$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.5$ .

### 6.1.2 Case 2: non-smooth velocity

We consider now the  $L$ -shaped domain  $\Omega = (-1, 1)^2 \setminus [0, 1)^2$ ; we take  $\mathbf{T}^d$  and  $p$  as above, but here

$$\mathbf{u} = \begin{pmatrix} y(x^2 + y^2)^{\frac{1}{3}} \\ -x(x^2 + y^2)^{\frac{1}{3}} \end{pmatrix},$$

which is divergence-free. The results when  $\alpha = 1$  and  $\gamma = 0$  are given in Table 6 while Tables 7 and 8 contain the results for the case  $\alpha = \gamma = 1$  with  $\tau = 0.01$  and  $\tau = 0.5$ , respectively.

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.65187 \times 10^{-2}$	$3.80529 \times 10^{-2}$	$5.22497 \times 10^{-2}$	1
3	0.177	$5.61550 \times 10^{-3}$	$6.85310 \times 10^{-3}$	$1.07102 \times 10^{-2}$	1
4	0.088	$1.32332 \times 10^{-3}$	$1.86233 \times 10^{-3}$	$2.53671 \times 10^{-3}$	1
5	0.044	$3.95496 \times 10^{-4}$	$5.79343 \times 10^{-4}$	$6.25652 \times 10^{-4}$	1
6	0.022	$1.24435 \times 10^{-4}$	$1.83765 \times 10^{-4}$	$1.55951 \times 10^{-4}$	1

Table 6: Case 2,  $\alpha = 1$ ,  $\gamma = 0$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.5$ .

## 6.2 A fixed-point algorithm

Instead of the Lions–Mercier type algorithm introduced in Subsection 6.1, we explore the following fixed-point strategy.

*Initialisation:*  $(\mathbf{T}_h^{(0)}, \mathbf{u}_h^{(0)}, p_h^{(0)}) = \mathbf{0}$ .

Then for  $k = 0, 1, \dots$ , do the following two steps.

*Step 1:* Find  $(\mathbf{u}_h^{(k+1)}, p_h^{(k+1)}) \in V_h \times Q_h$  such that

$$\begin{aligned} d(\mathbf{u}_h^{(k+1)}; \mathbf{u}_h^{(k+1)}, \mathbf{v}_h) + \frac{1}{\alpha} \int_{\Omega} \mathbf{D}(\mathbf{u}_h^{(k+1)}) : \mathbf{D}(\mathbf{v}_h) - \int_{\Omega} p_h^{(k+1)} \operatorname{div}(\mathbf{v}_h) &= \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_h + \frac{\gamma}{\alpha} \int_{\Omega} \mu(|\mathbf{T}_h^{(k)}|) \mathbf{T}_h^{(k)} : \mathbf{D}(\mathbf{v}_h), \\ \int_{\Omega} q_h \operatorname{div}(\mathbf{u}_h^{(k+1)}) &= 0 \end{aligned}$$

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.51269 \times 10^{-2}$	$6.79039 \times 10^{-2}$	$5.22609 \times 10^{-2}$	198
3	0.177	$4.65311 \times 10^{-3}$	$9.59150 \times 10^{-3}$	$1.07129 \times 10^{-2}$	198
4	0.088	$1.10623 \times 10^{-3}$	$2.04375 \times 10^{-3}$	$2.55916 \times 10^{-3}$	198
5	0.044	$7.88026 \times 10^{-4}$	$6.03974 \times 10^{-4}$	$7.14457 \times 10^{-4}$	198
6	0.022	$7.63053 \times 10^{-4}$	$2.28218 \times 10^{-4}$	$3.79151 \times 10^{-4}$	198

Table 7: Case 2,  $\alpha = \gamma = 1$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.01$ .

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.50999 \times 10^{-2}$	$6.78690 \times 10^{-2}$	$5.22585 \times 10^{-2}$	11
3	0.177	$4.56878 \times 10^{-3}$	$9.56288 \times 10^{-3}$	$1.07075 \times 10^{-2}$	8
4	0.088	$8.00090 \times 10^{-4}$	$2.03639 \times 10^{-3}$	$2.53571 \times 10^{-3}$	7
5	0.044	$2.08125 \times 10^{-4}$	$5.91247 \times 10^{-4}$	$6.25280 \times 10^{-4}$	7
6	0.022	$6.85220 \times 10^{-5}$	$1.92531 \times 10^{-4}$	$1.55842 \times 10^{-4}$	7

Table 8: Case 2,  $\alpha = \gamma = 1$ ,  $\delta = 10^{-5}$ ,  $\tau = 0.5$ .

for all  $(\mathbf{u}_h, q_h) \in V_h \times Q_h$ .

Step 2: Find  $\mathbf{T}_h^{(k+1)} \in M_h$  such that

$$\alpha \int_{\Omega} \mathbf{T}_h^{(k+1)} : \mathbf{S}_h + \gamma \int_{\Omega} \mu(|\mathbf{T}_h^{(k+1)}|) \mathbf{T}_h^{(k+1)} : \mathbf{S}_h = \int_{\Omega} \mathbf{D}(\mathbf{u}_h^{(k+1)}) : \mathbf{S}_h \quad \forall \mathbf{S}_h \in M_h.$$

It is easy to show that this algorithm produces uniformly bounded sequences.

The solvers used for these two steps are similar to those described in Subsection 6.1. In particular, we take  $(\mathbf{u}_h^{(k)}, p_h^{(k)})$  as initial guess for Newton's method for the finite element approximation of the Navier–Stokes system, except when  $k = 0$ , in which case we use the solution of the associated Stokes problem.

The results obtained using the stopping criterion (6.4) are given in Table 9. There are similar to those obtained in Table 5.

$n$	$h$	$\ \mathbf{T}^d - \mathbf{T}_h\ _{L^2(\Omega)}$	$\ \nabla(\mathbf{u} - \mathbf{u}_h)\ _{L^2(\Omega)}$	$\ p - p_h\ _{L^2(\Omega)}$	iter
2	0.354	$3.57082 \times 10^{-2}$	$8.21052 \times 10^{-2}$	$3.02063 \times 10^{-2}$	10
3	0.177	$7.73745 \times 10^{-3}$	$1.86629 \times 10^{-2}$	$6.18241 \times 10^{-3}$	8
4	0.088	$1.85777 \times 10^{-3}$	$4.55234 \times 10^{-3}$	$1.46369 \times 10^{-3}$	8
5	0.044	$4.60268 \times 10^{-4}$	$1.13344 \times 10^{-3}$	$3.60885 \times 10^{-4}$	8
6	0.022	$1.17442 \times 10^{-4}$	$2.92590 \times 10^{-4}$	$8.99455 \times 10^{-5}$	8

Table 9: Case 1,  $\alpha = \gamma = 1$ ,  $\delta = 10^{-5}$ .

Concerning the computational cost when similar results are obtained, i.e., when  $\tau = 0.5$  for the Lions–Mercier type algorithm, we note that the latter requires the solution of one more equation per iteration, namely the linear equation for  $\mathbf{T}_h^{(k+1)}$  in Step 2.

## References

- [1] W. BANGERTH, R. HARTMANN, AND G. KANSCHAT, *deal.II – a general purpose object oriented finite element library*, ACM Trans. Math. Softw., 33 (2007).

$h$	Lions–Mercier, $\tau = 0.5$			Alt. Algo		
	iter	CPU time [s]	wall time [s]	iter	CPU time [s]	wall time [s]
0.354	10	10.33	14.76	10	7.30	10.22
0.177	7	32.56	39.92	8	25.57	32.70
0.088	7	134.28	162.05	8	94.76	111.77
0.044	7	474.41	565.44	8	364.79	414.95
0.022	7	1542.63	1695.55	8	1307.07	1539.07

Table 10: Time (in seconds) needed to meet the stopping criteria (6.4) for the Lions–Mercier type algorithm (setup of Table 5) and the algorithm of Section 6.2

- [2] S. BARTELS, M. JENSEN, AND R. MÜLLER, *Discontinuous Galerkin finite element convergence for incompressible miscible displacement problems of low regularity*, SIAM J. Numer. Anal., 47 (2009), pp. 3720–3743.
- [3] D. BOFFI, *Three Dimensional Finite Element Methods for the Stokes Problem*, SIAM J. Numer. Anal., 34 (1997), pp. 664–670.
- [4] P. BOLTENHAGEN, Y. HU, E. MATTHYS, AND D. PINE, *Observation of bulk phase separation and coexistence in a sheared micellar solution*, Phys. Rev. Lett., 79 (1997), pp. 2359–2362.
- [5] A. BONITO, V. GIRAULT, AND E. SÜLI, *Finite element approximation of a strain-limiting elastic model*, IMA J. Numer. Anal., 40 (2020), pp. 29–86.
- [6] S. BRENNER, *Poincaré–Friedrichs inequalities for piecewise  $H^1$  functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324 (electronic).
- [7] ———, *Korn’s inequalities for piecewise  $H^1$  vector fields*, Math. Comp., 73 (2004), pp. 1067–1087 (electronic).
- [8] S. BRENNER AND L. SCOTT, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.
- [9] P. BRIDGMAN, *The Physics of High Pressure*, MacMillan, 1931.
- [10] A. BUFFA AND C. ORTNER, *Compact embeddings of broken Sobolev spaces and applications*, IMA J. Numer. Anal., 29 (2009), pp. 827–855.
- [11] M. BULÍČEK, P. GWIAZDA, J. MÁLEK, K. RAJAGOPAL, AND A. ŚWIERCZEWSKA-GWIAZDA, *On flows of fluids described by an implicit constitutive equation characterized by a maximal monotone graph*, in Mathematical Aspects of Fluid Mechanics, J. Robinson, J. Rodrigo, and W. Sadowski, eds., London Mathematical Society Lecture Note Series, Cambridge University Press, 2012, pp. 23–51.
- [12] M. BULÍČEK, J. MÁLEK, K. RAJAGOPAL, AND E. SÜLI, *On elastic solids with limiting small strain: modelling and analysis*, EMS Surv. Math. Sci., 1 (2014), pp. 283–332.
- [13] J. BURGERS, *Mechanical considerations-Model systems-Phenomenological theories of relaxation and of viscosity, First report on viscosity and plasticity, Prepared by the committee for the study of viscosity, 2nd Edition*, tech. rep., Academy of Sciences at Amsterdam, 1939.
- [14] P. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991, pp. 17–351.
- [15] M. CROUZEIX AND P.-A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary stokes equations*, R.A.I.R.O., 7 (1973), pp. 33–75.
- [16] D. DIPIETRO AND A. ERN, *Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier–Stokes equations*, Math. Comput., 79 (2010), pp. 1303–1330.

- [17] L. EVANS, *Partial Differential Equations*, Graduate Studies in Mathematics, 19, American Mathematical Society, 2010.
- [18] R. FALK, *Nonconforming finite element methods for the equations of linear elasticity*, Math. Comput., 57 (1991), pp. 529–550.
- [19] V. GIRAULT, J. LI, AND B. RIVIÈRE, *Strong convergence of discrete DG solutions to the heat equation*, J. Numer. Math., 24 (2016), pp. 235–252.
- [20] V. GIRAULT AND P.-A. RAVIART, *Finite element methods for Navier–Stokes equations: Theory and algorithms*, vol. 5 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1986.
- [21] T. HU, P. BOLTENHAGEN, E. MATTHYS, AND D. PINE, *Shear thickening in low-concentration solutions of wormlike micelles. II. Slip, fracture, and stability of the shear-induced phase*, J. Rheol., 42 (1998), pp. 1209–1226.
- [22] A. JOHNEN, J.-C. WEILL, AND J.-F. REMACLE, *Robust and efficient validation of the linear hexahedral element*, Procedia Eng., 203 (2017), pp. 271–283.
- [23] P. KNABNER, S. KOROTOV, AND G. SUMM, *Conditions for the invertibility of the isoparametric mapping for hexahedral finite elements*, Finite. Elem. Anal. Des., 40 (2003), pp. 159–172.
- [24] A. LASIS AND E. SÜLI, *Poincaré-type inequalities for broken Sobolev spaces*, Internal Report 03/10, Oxford University Computing Laboratory, Numerical Analysis Group, Oxford, England, OX1 3QD, 2003.
- [25] P.-L. LIONS AND B. MERCIER, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
- [26] D. LOPEZ-DIAZ, E. SARMIENTO-GOMEZ, C. GARZA, AND R. CASTILLO, *A rheological study in the dilute regime of the worm-micelle fluid made of zwitterionic surfactant (TDPS), anionic surfactant (SDS), and brine*, J. Colloid Interface Sci., 348 (2010), pp. 152–158.
- [27] J. MÁLEK AND K. RAJAGOPAL, *Mathematical properties of the equations governing the flow of fluids with pressure and shear rate dependent viscosities*, in Handbook of Mathematical Fluid Dynamics, S. Friedlander and D. Serre, eds., vol. 4, Elsevier, 2006.
- [28] J. MAXWELL, *On the dynamical theory of gases*, Philos. Trans. R. Soc. A, 157 (1866), pp. 26–78.
- [29] G. MINTY, *On a “monotonicity” method for the solution of nonlinear equations in Banach spaces*, Proc. Natl. Acad. Sci. U.S.A., 50 (1963), pp. 1038–1041.
- [30] J. OLDROYD, *On the formulation of rheological equations of state*, Proc. R. Soc. London Ser. A, 200 (1950), pp. 523–591.
- [31] D. PEACEMAN AND H. RACHFORD, *The numerical solution of parabolic elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28–41.
- [32] T. PERLÁCOVÁ AND V. PRŮŠA, *Tensorial implicit constitutive relations in mechanics of incompressible non-Newtonian fluids*, J. Non-Newtonian Fluid Mech., 216 (2015), pp. 13–21.
- [33] V. PRŮŠA, J. MÁLEK, AND K. RAJAGOPAL, *Generalizations of the Navier–Stokes fluid from a new perspective*, Int. J. Eng. Sci., 48 (2010), pp. 1907–1924.
- [34] K. RAJAGOPAL, *On implicit constitutive theories*, Appl. Math., 48 (2003), pp. 279–319.
- [35] ———, *On implicit constitutive theories for fluids*, J. Fluid Mech., 550 (2006), pp. 243–249.
- [36] ———, *A new development and interpretation of the Navier–Stokes fluid which reveals why the Stokes Assumption is inapt*, Int. J. Non-Linear Mech., 50 (2013), pp. 141–151.

- [37] ———, *Remarks on the notion of “pressure”*, Int. J. Non-Linear Mech., 71 (2015), pp. 165–172.
- [38] B. RIVIÈRE, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*, SIAM, 2008.
- [39] C. L. ROUX AND K. RAJAGOPAL, *Shear flows of a new class of power law fluids*, Appl. Math., 58 (2013), pp. 153–177.
- [40] G. STOKES, *On the theories of internal friction of fluids in motion, and of the equilibrium and motion of elastic solids*, Trans. Cambridge Phil. Soc., 8 (1845), pp. 287–341.
- [41] R. TEMAM, *Theory and Numerical Analysis of the Navier–Stokes Equations*, North-Holland, Amsterdam, The Netherlands, 1977.
- [42] S. ZHANG, *Subtetrahedral test for the positive Jacobian of hexahedral elements*. report.