Mathematical Models and Methods in Applied Sciences
Vol. 31, No. 5 (2021) 979–1020
© World Scientific Publishing Company
DOI: 10.1142/S021820252150024X

World Scientific
www.worldscientific.com

New primal and dual-mixed finite element methods for stable image registration with singular regularization

Nicolás Barnafi

MOX - Modellistica e Calcolo Scientifico,
Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, via Bonardi 9, 20133 Milano, Italy
nicolas.barnafi@polimi.it

Gabriel N. Gatica*

CI^{2}MA *and Departamento de Ingeniería Matemática,*
Universidad de Concepción, Casilla 160-C, Concepciòn, Chile
ggatica@ci2ma.udec.cl

Daniel E. Hurtado

Department of Structural and Geotechnical Engineering,
School of Engineering, and Institute for Biological and Medical Engineering,
Schools of Engineering, Medicine and Biological Sciences,
Pontificia Universidad Católica de Chile,
Vicuña Mackenna 4860, Santiago, Chile
dhurtado@ing.puc.cl

Willian Miranda

CI^{2}MA *and Departamento de Ingeniería Matemática,*
Universidad de Concepción, Casilla 160-C, Concepciòn, Chile
wmiranda@ci2ma.udec.cl

Ricardo Ruiz-Baier

School of Mathematics, Monash University,
9 Rainforest Walk, Melbourne 3800 VIC, Australia
and
Institute of Computer Science and Mathematical Modelling,
Sechenov University, Moscow, Russian Federation
and
Universidad Adventista de Chile, Casilla 7-D, Chillán, Chile
ricardo.ruizbaier@monash.edu

Received 17 July 2020
Revised 28 January 2021
Accepted 4 February 2021
Published 29 April 2021
Communicated by L. Beirao da Veiga

*Corresponding author.

This work introduces and analyzes new primal and dual-mixed finite element methods for deformable image registration, in which the regularizer has a nontrivial kernel, and constructed under minimal assumptions of the registration model: Lipschitz continuity of the similarity measure and ellipticity of the regularizer on the orthogonal complement of its kernel. The aforementioned singularity of the regularizer suggests to modify the original model by incorporating the additional degrees of freedom arising from its kernel, thus granting ellipticity of the former on the whole solution space. In this way, we are able to prove well-posedness of the resulting extended primal and dual-mixed continuous formulations, as well as of the associated Galerkin schemes. *A priori* error estimates and corresponding rates of convergence are also established for both discrete methods. Finally, we provide numerical examples confronting our formulations with the standard ones, which prove our finite element methods to be particularly more efficient on the registration of translations and rotations, in addition for the dual-mixed approach to be much more suitable for the quasi-incompressible case, and all the above without losing the flexibility to solve problems arising from more realistic scenarios such as the image registration of the human brain.

Keywords: Deformable image registration; finite elements; mixed finite elements.

AMS Subject Classification: 68U10, 65N30, 65N15, 74B05

1. Introduction

Deformable image registration (DIR) is a challenging process where a given set of images are aligned by means of a transformation that warps one or more of these images. It arises in numerous applications and particularly in medical imaging.[32] Its formulation requires three inputs: a transformation model (composed by a family of mappings that warp the target images), a function that measures the differences between images known as similarity measure, and a regularizer that renders the problem well-posed. In addition to the many variants of these components, different modeling approaches exist, between which we highlight: traditional variational minimization,[23,28] L^2-optimal mass transport[22,35] (which does not require regularization), and level-set modeling.[33] The solution strategy in general considers the incorporation of an auxiliary time variable, which can be seen as a semi-implicit formulation of the proximal point algorithm[31] recently extended to a more general class of proximal operators by using forward–backward splitting.[17] Also, machine learning techniques have been recently developed for the solution of this problem, which do not depend on the existence of ground truth solutions and support large deformations.[6] This last work proved competitive against the well-established software ANTs.[5]

For a more mathematical explanation of DIR, let us now consider a domain $\Omega \subset \mathrm{R}^{d=2,3}$, and two fields $R : \Omega \to \mathrm{R}$ and $T : \Omega \to \mathrm{R}$ referred to as *reference* and *target* images, where $R(\boldsymbol{x})$ and $T(\boldsymbol{x})$ denote the *image intensity* at point $\boldsymbol{x}$. Then, the objective of DIR is to find a mapping of T onto R by means of a warping $\boldsymbol{u}$ such that — ideally — it holds that

$$T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) = R(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \Omega. \tag{1.1}$$

This problem is ill-posed in general, so one formulates it as a minimization problem by considering a *similarity measure* $\mathcal{D}$ (a functional which attains its minimum when (1.1) holds), a regularizer $\mathcal{S}$ (which provides smoothness to the problem), a family of deformations $\mathcal{V}$ (such that $\boldsymbol{u} \in \mathcal{V}$) and a positive constant α (which balances $\mathcal{D}$ and $\mathcal{S}$). Putting everything together, the following minimization problem arises:

$$\min_{\boldsymbol{v}\in\mathcal{V}}\{\alpha\mathcal{D}(\boldsymbol{v}) + \mathcal{S}(\boldsymbol{v})\}. \tag{1.2}$$

The choices of $\mathcal{V}$ and $\mathcal{S}$ are not independent. For example, it would not make sense to consider $\mathcal{V} = \boldsymbol{L}^2(\Omega)$ together with a regularizer $\mathcal{S}(\boldsymbol{u}) = \int_\Omega |\nabla \boldsymbol{u}|^2\, dx$ which penalizes steep gradients, as $\mathcal{S}$ would not be well-defined in all $\mathcal{V}$. It is common practice to consider $\mathcal{S}$ to be a quadratic term of the form $\mathcal{S}(\boldsymbol{v}) = \frac{1}{2}a(\boldsymbol{v}, \boldsymbol{v})$, where a is a suitable bounded bilinear form. One common example is given by considering the L^2 error as a similarity measure together with the $\boldsymbol{H}_0^1$ norm as a regularizer (equivalently, using $a(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \nabla \boldsymbol{u} \cdot \nabla \boldsymbol{v}$), which yields the following problem:

$$\min_{\boldsymbol{v}\in\boldsymbol{H}_0^1(\Omega)} \left\{ \alpha \int_\Omega |T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) - R(\boldsymbol{x})|^2\, dx + \int_\Omega |\nabla \boldsymbol{u}|^2\, dx \right\}, \tag{1.3}$$

with first-order conditions given by: Find $\boldsymbol{u} \in \boldsymbol{H}_0^1(\Omega)$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = -\langle \nabla \mathcal{D}(\boldsymbol{u}), \boldsymbol{v}\rangle = -\int_\Omega \nabla T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}))(T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) - R(\boldsymbol{x}))\, dx \quad \forall\, \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega),$$

where $\nabla\mathcal{D}$ stands for the Fréchet derivative of $\mathcal{D}$. Further details and examples beyond this brief overview can be found in Ref. 28.

Our present work has been mainly motivated by the study of lung regional deformation computed from tomography images of the thorax.[14,25] However, as we will illustrate later on, it is also applicable to related problems such as the image registration of the human brain. The optimal warping, $\boldsymbol{u}$, can be interpreted as a displacement field, from which the gradient $\nabla\boldsymbol{u}$ can be calculated to obtain the strain tensor required to characterize the continuum mechanics framework. The study of deformation from one side has revealed the lungs to present a highly heterogeneous and anisotropic behavior,[2,24] thus providing new deformation-based markers to understand lung diseases.[13,30] The proposal of the optical flow formulation by Horn and Schunk[23] gave origin to much mathematical analysis at the continuous level, with an increasing interest towards the discrete analysis in an algorithm-specific fashion in Ref. 29, in the optimal-control setting within a more classical Galerkin framework,[26] and more recently the variational problem was tackled in its primal and mixed formulation in Ref. 7.

In fact, the mixed finite element method (MFEM) is a well-established technique which allows to incorporate unknowns of physical interest, such as stress and rotation, and also delivers locking-free schemes in the context of incompressible elasticity (see, e.g. Refs. 12 and 19). It also introduces additional difficulties: (i) the new variables increase the dimension of the numerical scheme, making its computational

solution more expensive, (ii) the mixed formulation may now possess a saddle-point structure, which results in linear systems of equations that are harder to solve numerically and (iii) only discrete spaces that satisfy the required inf-sup conditions grant a stable scheme, therefore restricting the choices for approximations and also demanding more attention in the analysis of the finite element scheme. For a mixed formulation of DIR with elastic regularization and a target image with Lipschitz gradient, it has been shown that classical existence of solutions is independent of the regularization parameter in the primal case. Furthermore, both primal and mixed schemes give existence and uniqueness for a sufficiently small regularization, and PEERS elements, as well as BDM-$\mathbb{P}_0$ for stress-displacement, are inf-sup stable.[7] In addition, the drawback mentioned in (iii) is alternatively overcome in Ref. 7 by using an augmented mixed variational formulation whose discrete analysis does not require the verification of any inf-sup condition, and hence arbitrary finite element subspaces can be employed. More precisely, in this last work a complete numerical analysis of the method was presented, in the particular case of an elastic regularizer and a sum-of-squared-differences similarity measure with Neumann boundary conditions. Using such conditions is usually physically desirable, as other ones present artificial stress accumulation on the boundaries, thus yielding the difficulty of non-uniqueness to iterative schemes.

In this paper, we aim to generalize the analysis presented in Ref. 7 to regularizers that may present a kernel, and to Lipschitz similarity measures. This is performed by splitting weakly the warping with respect to the kernel of the regularizer so that such kernel remains present in the formulation throughout the model, under the assumption of a relationship between the regularizer and the similarity measure commonly known in the inverse problems community as *source condition*.[34] Numerical experiments validating our aforedescribed extended model and showing how it compares to a more traditional formulation are also presented. The rest of the work is organized as follows. At the end of this section, we collect some notations to be employed in the paper. In Sec. 2, we derive the new model and analyze its primal formulation at both continuous and discrete levels. The main results, which are obtained by using the Babuška–Brezzi theory and duality arguments, include well-posedness of the continuous and discrete formulations, *a priori* error estimates, and the respective rates of convergence. Then, in Sec. 3, we introduce and analyze, using basically the same theoretical tools from Sec. 2, an extended dual-mixed formulation in the particular (though very common and useful) case of an elastic energy. Next, in Sec. 4, we explain how to use the traditional time regularization to implement the methods, and provide a suitable bound of the time step guaranteeing convergence. In Sec. 5, we present several numerical experiments illustrating convergence, the capability of the methods to capture translations and rotations, the effect of the added degrees of freedom, the advantage of using the dual-mixed approach in the quasi-incompressible case, and the application to the image registration of the human brain. We conclude the paper with a brief discussion section.

Notation

Throughout the paper, $\Omega \subseteq \mathrm{R}^2$ is a given open and bounded domain with polyhedral boundary Γ, whose outward unit normal vector on Γ is denoted $\boldsymbol{\nu}$. Standard notation is adopted for the Lebesgue space $\mathrm{L}^2(\Omega)$ and for the Sobolev spaces $\mathrm{H}^m(\Omega)$, $m \in \mathrm{N}$. In particular, their corresponding norms, either for the scalar, vectorial, or tensorial case, are denoted by $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_{m,\Omega}$, respectively. In turn, given a generic scalar functional space M, we let $\mathbf{M}$ and $\mathbb{M}$ be the corresponding vectorial and tensorial counterparts, whereas $\|\cdot\|$, with no subscripts, will be employed for the norm of any element or operator whenever there is no confusion about the space to which they belong. Also, $\mathbb{I}$ stands for the identity tensor in $\mathrm{R}^{2\times 2}$, and $|\cdot|$ denotes the Euclidean norm in R^2. On the other hand, for any tensor field $\boldsymbol{\tau} = (\tau_{ij})_{i,j=1,2}$, we let $\mathbf{div}\,\boldsymbol{\tau}$ be the divergence operator div acting along the rows of $\boldsymbol{\tau}$, and denote by $\boldsymbol{\tau}^{\mathtt{t}}$, $\mathrm{tr}(\boldsymbol{\tau})$, and $\boldsymbol{\tau}^{\mathtt{d}}$, the transpose, the trace, and the deviatoric tensor of $\boldsymbol{\tau}$, respectively. In addition, given other tensor field $\boldsymbol{\zeta} = (\zeta_{ij})_{i,j=1,2}$, we define the tensor inner product between $\boldsymbol{\tau}$ and $\boldsymbol{\zeta}$ as $\boldsymbol{\tau} : \boldsymbol{\zeta} := \sum_{i,j=1}^{2} \tau_{ij}\zeta_{ij}$. Finally, we introduce $\mathbb{H}(\mathbf{div};\Omega) := \{\boldsymbol{\tau} \in \mathbb{L}^2(\Omega) : \mathbf{div}\,\boldsymbol{\tau} \in \mathbf{L}^2(\Omega)\}$, which, equipped with the norm $\|\boldsymbol{\tau}\|_{\mathbf{div};\Omega} := \{\|\boldsymbol{\tau}\|_{0,\Omega}^2 + \|\mathbf{div}\,\boldsymbol{\tau}\|_{0,\Omega}^2\}^{1/2}$, is a Hilbert space.

2. Extended Primal Formulation in Abstract Form

In this section, we derive an abstract extended model and analyze its continuous and discrete primal formulations.

2.1. *Setting of the problem*

As briefly commented in Sec. 1, our problem is posed in the following framework: a Hilbert space $(\mathcal{V}, \langle\cdot,\cdot\rangle)$, a similarity measure $\mathcal{D} : \mathcal{V} \to \mathrm{R}$, a symmetric bounded bilinear form $a : \mathcal{V}\times\mathcal{V} \to \mathrm{R}$ acting as the regularizer, and a positive scalar α. Then, we look for minimizers of the following problem:

$$\min_{v\in\mathcal{V}} \left\{ \alpha\mathcal{D}(v) + \frac{1}{2}a(v,v) \right\}. \tag{2.1}$$

The first-order conditions yield the following nonlinear problem: Find $u \in \mathcal{V}$ such that

$$a(u,v) = \alpha\, F_u(v) \quad \forall\, v \in \mathcal{V}, \tag{2.2}$$

where given $w \in \mathcal{V}$, $F_w : \mathcal{V} \to \mathrm{R}$ is the linear functional defined as

$$F_w(v) := -\langle \nabla\mathcal{D}(w), v\rangle \quad \forall\, v \in \mathcal{V}, \tag{2.3}$$

which is clearly bounded with $\|F_w\|_{\mathcal{V}'} = \|\nabla\mathcal{D}(w)\|_{\mathcal{V}}$. Next, denoting by Q the kernel of the adjoint of the bounded operator induced by a, which we assume to be nontrivial and finite-dimensional, and splitting $\mathcal{V}$ as $Q^{\perp} \oplus Q$, we can rewrite (2.1) equivalently as

$$\min_{(v,\eta)\in Q^{\perp}\times Q} \left\{ \alpha\mathcal{D}(v+\eta) + \frac{1}{2}a(v,v) \right\},$$

and then impose the condition $v \in Q^\perp$ as $\langle v, \xi \rangle = 0 \; \forall \xi \in Q$, to obtain

$$\min_{(v,\eta)\in\mathcal{V}\times Q} \max_{\xi\in Q} \left\{ \alpha\mathcal{D}(v+\eta) + \frac{1}{2}a(v,v) + \langle v, \xi \rangle \right\}. \tag{2.4}$$

Finally, to avoid having the nonlinear term $\mathcal{D}$ in more than one equation, we perform the change of variables $v \leftarrow v + \eta$, whence (2.4) becomes

$$\min_{(v,\eta)\in\mathcal{V}\times Q} \max_{\xi\in Q} \left\{ \alpha\mathcal{D}(v) + \frac{1}{2}a(v,v) + \langle v-\eta, \xi \rangle \right\}. \tag{2.5}$$

In this formulation a is not elliptic, which gives difficulties in proving the well-posedness of the weak problem. If we consider the form (2.4) with solution $(u, \lambda) \in \mathcal{V} \times Q$, we get that $F_{u+\lambda}(\xi) = 0$ for all ξ in Q, which is fully nonlinear and does not give the required control over λ, but on the other hand, form (2.5) gives rise to a non-invertible linear operator. This hints the requirement of controlling the component of u in Q, for which, given a positive constant β, we consider the problem

$$\min_{(v,\eta)\in\mathcal{V}\times Q} \max_{\xi\in Q} \left\{ \alpha\mathcal{D}(v) + \frac{1}{2}a(v,v) + \langle v-\eta, \xi \rangle + \frac{\beta}{2}\|\eta\|_{\mathcal{V}}^2 \right\}. \tag{2.6}$$

We call (2.6) the extended formulation of (2.1). Equivalently, this setting can be obtained by splitting $\mathcal{V}$ in the Euler–Lagrange equations (2.2). First write them as finding $(u, \lambda) \in \mathcal{V} \times Q$ such that

$$\begin{aligned} a(u,v) &= \alpha\, F_u(v) \quad && \forall\, v \in Q^\perp, \\ \langle \lambda, \xi \rangle &= \langle u, \xi \rangle && \forall\, \xi \in Q, \end{aligned} \tag{2.7}$$

and then impose the weak orthogonality by adding a Lagrange multiplier ρ together with the compact perturbation $\beta\langle\lambda, \eta\rangle$ to obtain the extended weak form: Find $(u, \lambda, \rho) \in \mathcal{V} \times Q \times Q$ such that

$$\begin{aligned} a(u,v) + \beta\langle\lambda,\eta\rangle + \langle v-\eta,\rho\rangle &= \alpha F_u(v) \quad && \forall\,(v,\eta) \in \mathcal{V}\times Q, \\ \langle u-\lambda,\xi\rangle &= 0 && \forall\,\xi \in Q. \end{aligned} \tag{2.8}$$

The extended formulation presents two advantages:

- The standard formulation gives rise to a nonlinear compatibility condition for the solution u, namely $0 = F_u(\xi) \forall\, \xi \in Q$, which arises after testing (2.2) against elements in Q. Thus, the new variable λ does not affect the compatibility condition. The existence of functions such that this holds is known as the source condition, and is usually stated in the inverse problems community as $\partial\mathcal{D} \perp Q$,[34] which we assume true throughout the paper.
- Fixed-point schemes arising from such problems impose an undesired orthogonality to the solution, which we refer to as kernel locking. If we let u_n in $\mathcal{V}$ be a previous solution, we get the fixed-point problem of finding u_{n+1} in $\mathcal{V}$ such that

$$a(u_{n+1}, v) = F_{u_n}(v) \quad \forall\, v \in \mathcal{V}\,.$$

This problem does not have a unique solution, so it is common in practice to choose u_{n+1} such that $u_{n+1} \perp Q$. The orthogonal space is closed, and thus if the sequence $\{u_n\}_n$ converges to a solution u, such solution is also orthogonal to Q.

The interpretation of λ in the overall context of the problem is crucial to understand the extent to which it regularizes the problem. For it, we first focus on the nonlinear compatibility condition $F_u(\xi) = 0$, also written as $\Pi_Q \nabla \mathcal{D}(u) = 0$, where $\Pi_Q : \mathcal{V} \to Q$ is the orthogonal projection on Q. This condition rises naturally from the extended formulation, and thus it is a necessary condition for the differentiability of $\mathcal{D}$. If the functional does not comply with it, then it is unrelated to a variational principle, so we can add a function $\tilde{\lambda} = \Pi_Q \nabla \mathcal{D}(u) \in Q$ to (2.8) — without λ — such that the compatibility condition holds, that is,

$$\begin{aligned} a(u,v) + \langle v - \eta, \rho \rangle &= \alpha(F_u - \tilde{\lambda})(v) \quad \forall\, (v,\eta) \in \mathcal{V} \times Q, \\ \langle u - \lambda, \xi \rangle &= 0 \qquad\qquad\qquad\quad \forall\, \xi \in Q. \end{aligned} \tag{2.9}$$

We can see that λ indeed takes the desired values by testing the first equation with $v = \eta \in Q$, which gives $F_u(\eta) = \langle \tilde{\lambda}, \eta \rangle \ \forall\, \eta \in Q$. Note that the same holds if we take the term $\langle \tilde{\lambda}, v \rangle$ to the left-hand side and replace it with $\langle \tilde{\lambda}, \eta \rangle$, which means that the compatibilized problem is equivalent to (2.8) if we take $\beta\tilde{\lambda} = \lambda$. In what follows, we show that such choice gives a well-posed problem with many numerical advantages, for which we will make the following assumptions:

(A1) There exist two positive constants $\widetilde{c}_a$ and $\widetilde{C}_a$ such that

$$\widetilde{c}_a \|v\|_{\mathcal{V}}^2 \leq a(v,v) \quad \forall\, v \in Q^{\perp},$$

and

$$|a(w,v)| \leq \widetilde{C}_a \|w\|_{\mathcal{V}} \|v\|_{\mathcal{V}} \quad \forall\, w, v \in \mathcal{V}.$$

(A2) There exists a positive constant $L_{\mathcal{D}}$ and a space $\mathcal{W}$ containing $\mathcal{V}$, such that the embedding $i_{\mathcal{W}} : \mathcal{V} \hookrightarrow \mathcal{W}$ is compact and there holds

$$\|\nabla \mathcal{D}(z_1) - \nabla \mathcal{D}(z_2)\|_{\mathcal{V}} \leq L_{\mathcal{D}} \|z_1 - z_2\|_{\mathcal{W}} \quad \forall\, z_1, z_2 \in \mathcal{V}.$$

(A3) There exists a positive constant $M_{\mathcal{D}}$ such that $\|\nabla \mathcal{D}(w)\|_{\mathcal{V}} \leq M_{\mathcal{D}}$ for all $w \in \mathcal{V}$.

2.2. *Analysis of the continuous formulation*

We now show that the extended problem (2.8) has at least one solution, which is stable with respect to the data. For this, we first set the product space $H := \mathcal{V} \times Q$, and let $A : H \times H \to \mathrm{R}$ and $B : H \times Q \to \mathrm{R}$ be the bilinear forms involved in (2.8), that is

$$A((w,\vartheta),(v,\eta)) := a(w,v) + \beta\langle \vartheta, \eta \rangle \quad \forall\, (w,\vartheta),(v,\eta) \in H, \tag{2.10}$$

and

$$B((v,\eta),\xi) := \langle v - \eta, \xi \rangle \quad \forall\, (v,\eta) \in H, \quad \forall\, \xi \in Q. \tag{2.11}$$

In addition, for each $z \in \mathcal{V}$, we denote by $G_z : H \to \mathrm{R}$ the linear functional given by (cf. (2.3))

$$G_z(v,\eta) := \alpha\, F_z(v) \quad \forall\, (v,\eta) \in H. \tag{2.12}$$

Note here that A, B, and G_z are bounded. In fact, considering the corresponding euclidean norm for the product space H, and denoting the constants $\|A\| := \max\{\widetilde{C}_a, \beta\}$ (cf. (A1)) and $\|B\| := \sqrt{2}$, we easily find, using the Cauchy–Schwarz inequality, that

$$|A((w,\vartheta),(v,\eta))| \leq \|A\|\|(w,\vartheta)\|_H\|(v,\eta)\|_H \quad \text{and}$$

$$|B((v,\eta),\xi)| \leq \|B\|\|(v,\eta)\|_H\|\xi\|_{\mathcal{V}},$$

for all $(w,\vartheta),(v,\eta) \in H, \forall\, \xi \in Q$. In turn, it is clear from the above definition of G_z and the fact that $F_z \in \mathcal{V}'$ (cf. (2.3)) that $G_z \in H'$ and $\|G_z\| = \alpha\|F_z\| = \alpha\|\nabla D(z)\|$. According to the previous notations, (2.8) can be rewritten as: Find $((u,\lambda),\rho) \in H \times Q$ such that

$$\begin{aligned} A((u,\lambda),(v,\eta)) + B((v,\eta),\rho) &= G_u(v,\eta) \quad \forall\, (v,\eta) \in H, \\ B((u,\lambda),\xi) &= 0 \qquad\qquad \forall\, \xi \in Q. \end{aligned} \tag{2.13}$$

Then, we introduce the operator $T : \mathcal{V} \to \mathcal{V}$ defined by $T(z) := \widetilde{u}$ for each $z \in \mathcal{V}$, where $\widetilde{u} \in \mathcal{V}$ is the first component of the solution to the problem: Find $((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho}) \in H \times Q$ such that

$$\begin{aligned} A((\widetilde{u},\widetilde{\lambda}),(v,\eta)) + B((v,\eta),\widetilde{\rho}) &= G_z(v,\eta) \quad \forall\, (v,\eta) \in H, \\ B((\widetilde{u},\widetilde{\lambda}),\xi) &= 0 \qquad\qquad \forall\, \xi \in Q. \end{aligned} \tag{2.14}$$

We stress here that solving (2.13) is equivalent to seeking a fixed point of T, that is: Find $u \in \mathcal{V}$ such that $T(u) = u$. In the following lemma we show that, for any $z \in \mathcal{V}$, the linear problem (2.14) is well-posed, whence the operator T is well-defined.

Lemma 2.1. *Given z in $\mathcal{V}$, there exists a unique $((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho}) \in H \times Q$ solution to* (2.14)*. Moreover, there exists a positive constant C_T, independent of $((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho})$, such that the following a priori estimate holds*:

$$\|T(z)\|_{\mathcal{V}} \leq \|((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho})\|_{H\times Q} \leq C_T\|G_z\|_{H'} = \alpha C_T\|\nabla\mathcal{D}(z)\|_{\mathcal{V}}. \tag{2.15}$$

Proof. In what follows, we apply the Babuška–Brezzi theory (cf. Chap. 2 of Ref. 19). To this end, we first let N be the kernel of the operator induced by B, that is

$$N = \{(v,\eta) \in H : B((v,\eta),\xi) = 0\ \forall\, \xi \in Q\},$$

which, according to (2.11), yields $N = \{(v,\eta) \in H : \eta = \Pi_Q v\}$. Then, given $(v,\eta) = (v,\Pi_Q v) \in N$, we split $v = v^\perp + \eta \in Q^\perp \oplus Q$ and use assumption (A1) to obtain

$$A((v,\eta),(v,\eta)) = a(v^\perp,v^\perp) + \beta\|\eta\|^2_{\mathcal{V}} \geq \widetilde{c}_a\|v^\perp\|^2_{\mathcal{V}} + \beta\|\eta\|^2_{\mathcal{V}} \geq c_a\|(v,\eta)\|^2_H, \tag{2.16}$$

with $c_a := \min\{\widetilde{c}_a, \frac{\beta}{2}\}$, which gives the N-ellipticity of A. On the other hand, given an arbitrary $\xi \in Q$, we easily see that

$$\sup_{\substack{(v,\eta)\in H\\(v,\eta)\neq(0,0)}} \frac{B((v,\eta),\xi)}{\|(v,\eta)\|_H} \geq \frac{B((0,-\xi),\xi)}{\|(0,-\xi)\|_H} = c_b\|\xi\|_{\mathcal{V}}, \tag{2.17}$$

with $c_b = 1$, which proves the continuous inf-sup condition for B. In this way, a straightforward application of Theorem 2.3 in Ref. 19 implies the existence of a unique solution to (2.14) and the corresponding stability estimate (2.15) with a constant C_T depending on c_a, c_b, and $\|A\|$. □

Now, given $r > 0$, we let $\bar{B}(r)$ be the closed ball of $\mathcal{V}$ centered at the origin with radius r. Then, as a consequence of the previous lemma, we have the following additional result.

Lemma 2.2. *Let $L_{\mathcal{D}}$, $M_{\mathcal{D}}$, and C_T be the constants specified in* (A2), (A3), *and Lemma* 2.1, *respectively, and define $r_0 := \alpha C_T M_{\mathcal{D}}$. Then, there hold $T(\mathcal{V}) \subseteq \bar{B}(r_0)$ and*

$$\|T(z_1) - T(z_2)\|_{\mathcal{V}} \leq \alpha C_T L_{\mathcal{D}}\|z_1 - z_2\|_{\mathcal{W}} \quad \forall\, z_1, z_2 \in \mathcal{V}. \tag{2.18}$$

Proof. Given $z \in \mathcal{V}$, it readily follows from (2.15) and (A3) that $\|T(z)\|_{\mathcal{V}} \leq \alpha C_T M_{\mathcal{D}} := r_0$, which proves the required inclusion for T. In turn, the fact that (2.14) is a linear problem guarantees that, given $z_1, z_2 \in \mathcal{V}$, the difference $T(z_1) - T(z_2)$ is the first component of the unique solution of (2.14) when G_z is replaced there by the functional $G_{z_1} - G_{z_2}$. Thus, from the stability estimate (2.15) again, and the Lipschitz continuity provided by (A2), we deduce that

$$\|T(z_1) - T(z_2)\|_{\mathcal{V}} \leq \alpha C_T\|\nabla\mathcal{D}(z_1) - \nabla\mathcal{D}(z_2)\|_{\mathcal{V}} \leq \alpha C_T L_{\mathcal{D}}\|z_1 - z_2\|_{\mathcal{W}},$$

which completes the proof. □

Having established the above properties of T, we are now in position to provide the main result of this section.

Theorem 2.1. *Let r_0 be the radius defined in the statement of Lemma* 2.2. *Then, problem* (2.13) *admits at least one solution $((u,\lambda),\rho) \in H \times Q$, with $u \in \bar{B}(r_0)$. Moreover, under the additional assumption $\alpha C_T L_{\mathcal{D}}\|i_{\mathcal{W}}\| < 1$, this solution is unique.*

Proof. We begin by noticing from Lemma 2.2 that certainly $T(\bar{B}(r_0)) \subseteq \bar{B}(r_0)$. Next, it is easy to see from the Lipschitz continuity of T (cf. (2.18)) and the compactness of the embedding $i_{\mathcal{W}} : \mathcal{V} \to \mathcal{W}$ (cf. (A2)) that $\overline{T(\bar{B}(r_0))}$ is compact. Hence, Schauder's fixed-point theorem (cf. Theorem 9.12-1(b) in Ref. 15) implies the existence of a fixed point $u \in \bar{B}(r_0)$ for T, and hence of a solution $((u,\lambda),\rho) \in H \times Q$ to problem (2.13). Furthermore, it also follows from (2.18) and (A2) that

$$\|T(z_1) - T(z_2)\|_{\mathcal{V}} \leq \alpha C_T L_{\mathcal{D}}\|i_{\mathcal{W}}\|\|z_1 - z_2\|_{\mathcal{V}} \quad \forall\, z_1, z_2 \in \mathcal{V},$$

whence the uniqueness in $\mathcal{V}$ is imposed by forcing T to be a contraction and then using the Banach fixed-point theorem, which happens precisely when $\alpha C_T L_{\mathcal{D}} \|i_{\mathcal{W}}\| < 1$. □

2.3. *Analysis of the discrete scheme*

In this section, we consider the Galerkin scheme approximating the solutions of (2.13), establish its well-posedness, derive the associated Céa estimate, and provide the corresponding rates of convergence. For this purpose, we now let $\{\mathcal{V}_h\}_{h>0}$ be a sequence of finite dimensional subspaces of $\mathcal{V}$, where $h > 0$ is an index thought as a characteristic meshsize. Then, bearing in mind that Q is finite-dimensional, and defining $H_h := \mathcal{V}_h \times Q$, our discrete extended problem reduces to: Find $((u_h, \lambda_h), \rho_h) \in H_h \times Q$ such that

$$\begin{aligned} A((u_h, \lambda_h), (v_h, \eta_h)) + B((v_h, \eta_h), \rho_h) &= G_{u_h}(v_h, \eta_h) \quad \forall\, (v_h, \eta_h) \in H_h, \\ B((u_h, \lambda_h), \xi_h) &= 0 \qquad\qquad\quad \forall\, \xi_h \in Q. \end{aligned} \tag{2.19}$$

In turn, we introduce the discrete operator $T_h : \mathcal{V}_h \to \mathcal{V}_h$ given by $T(z_h) := \widetilde{u}_h$ $\forall\, z_h \in \mathcal{V}_h$, where $\widetilde{u}_h$ is the first component of the solution $((\widetilde{u}_h, \widetilde{\lambda}_h), \widetilde{\rho}_h) \in H_h \times Q$ to (2.19) with G_{z_h} instead of G_{u_h}, that is:

$$\begin{aligned} A((\widetilde{u}_h, \widetilde{\lambda}_h), (v_h, \eta_h)) + B((v_h, \eta_h), \widetilde{\rho}_h) &= G_{z_h}(v_h, \eta_h) \quad \forall\, (v_h, \eta_h) \in H_h, \\ B((\widetilde{u}_h, \widetilde{\lambda}_h), \xi_h) &= 0 \qquad\qquad\quad \forall\, \xi_h \in Q. \end{aligned} \tag{2.20}$$

As for the continuous case, we emphasize here that solving (2.19) is equivalent to finding $u_h \in \mathcal{V}_h$ such that $T_h(u_h) = u_h$. We start our discrete analysis by proving the well-posedness of (2.20), thus confirming that T_h is well-defined.

Lemma 2.3. *Given $z_h \in \mathcal{V}_h$, there exists a unique $((\widetilde{u}_h, \widetilde{\lambda}_h), \widetilde{\rho}_h) \in H_h \times Q$ solution to* (2.20)*. Moreover, with the same constant C_T from Lemma* 2.1*, there holds*

$$\begin{aligned} \|T_h(z_h)\|_{\mathcal{V}} \le \|((\widetilde{u}_h, \widetilde{\lambda}_h), \widetilde{\rho}_h)\|_{H\times Q} &\le C_T \|G_{z_h}\|_{H'} = \alpha C_T \|\nabla \mathcal{D}(z_h)\|_{\mathcal{V}} \\ &\le \alpha C_T M_{\mathcal{D}} =: r_0. \end{aligned} \tag{2.21}$$

Proof. The proof is analogous to the one shown for the well-posedness of problem (2.14) (cf. Lemma 2.1). In fact, we first observe that the discrete kernel N_h of B becomes

$$N_h = \{(v_h, \eta_h) \in H_h : \eta_h = \Pi_Q v_h\},$$

which is clearly contained in N, and hence the N_h-ellipticity of A follows from that of N, and certainly with the same ellipticity constant c_a. In turn, given $\xi_h \in Q$, the discrete inf-sup condition for B is obtained as in (2.17) by bounding below the involved supremum with $(v_h, \eta_h) = (0, -\xi_h)$, which yields the same resulting constant c_b. In this way, applying now the discrete version of the Babuška–Brezzi theory (cf. Theorem 2.4 in Ref. 19), and using from (A3) that $\|\nabla \mathcal{D}(z_h)\| \le M_{\mathcal{D}}$, we conclude the proof. □

Next, given $r > 0$, we let $\bar{B}_h(r)$ be the closed ball of $\mathcal{V}_h$ centered at the origin with radius r. Then, the main result concerning the solvability of (2.19), which summarizes the discrete analogues of Lemma 2.2 and Theorem 2.1, is established as follows.

Theorem 2.2. *The discrete problem* (2.19) *has at least one solution* $((u_h, \lambda_h), \rho_h) \in H_h \times Q$, *with* $u_h \in \bar{B}_h(r_0)$. *Moreover, under the assumption* $\alpha C_T L_{\mathcal{D}} \|i_W\| < 1$, *this solution is unique.*

Proof. We first notice from (2.21) (cf. Lemma 2.3) that $T_h(\mathcal{V}_h) \subseteq \bar{B}_h(r_0)$, which obviously yields, in particular, $T_h(\bar{B}_h(r_0)) \subseteq \bar{B}_h(r_0)$. In addition, proceeding as in the proofs of Lemma 2.2 and Theorem 2.1, but certainly using now the linear character of problem (2.20), and employing the stability estimate (2.21), the assumption (A2), and the boundedness of i_W, we easily find that

$$\|T_h(z_{1,h}) - T_h(z_{2,h})\|_{\mathcal{V}} \leq \alpha C_T L_{\mathcal{D}} \|i_W\| \|z_{1,h} - z_{2,h}\|_{\mathcal{V}} \quad \forall\, z_{1,h}, z_{2,h} \in \mathcal{V}_h. \tag{2.22}$$

In this way, the fact that $\bar{B}_h(r_0)$ is clearly a compact and convex subset of $\mathcal{V}_h$, the continuity of $T_h : \bar{B}_h(r_0) \to \bar{B}_h(r_0)$, and a straightforward application of Brouwer's theorem (cf. Theorem 9.9-2 in Ref. 15) implies the existence of a fixed point $u_h \in \bar{B}_h(r_0)$ for T_h, and therefore of a solution $((u_h, \lambda_h), \rho_h) \in H_h \times Q$ to (2.19). Finally, uniqueness in $\mathcal{V}_h$ follows again by forcing T_h to be a contraction. □

Having proved the existence of solutions for the discrete and continuous problems, we now provide the Céa estimate for the corresponding error. In what follows, given a subspace X_h of a generic Banach space $(X, \|\cdot\|_X)$, we set

$$\operatorname{dist}(x, X_h) := \inf_{x_h \in X_h} \|x - x_h\|_X \quad \forall\, x \in X.$$

Theorem 2.3. *Assume that* $\alpha C_T L_{\mathcal{D}} \|i_W\| \leq 1 - \delta$, *with* $\delta \in]0, 1[$, *and let* $((u, \lambda), \rho) \in H \times Q$ *and* $((u_h, \lambda_h), \rho_h) \in H_h \times Q$ *be the unique solutions of* (2.13) *and* (2.19), *respectively. Then, there exists a positive constant* $\widehat{C}$, *depending only on* c_a, c_b, $\|A\|$, *and* $\|B\|$, *and hence independent of* h, *such that*

$$\|((u, \lambda), \rho) - ((u_h, \lambda_h), \rho_h)\|_{H \times Q} \leq \delta^{-1} \widehat{C} \operatorname{dist}(u, \mathcal{V}_h). \tag{2.23}$$

Proof. Let $((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h) \in H_h \times Q$ be the resulting unique solution of the discrete scheme (2.19) when the functional G_{u_h} is replaced there by G_u. In this way, $((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h) \in H_h \times Q$ constitutes a conforming Galerkin approximation of the unique solution $((u, \lambda), \rho) \in H \times Q$ to (2.13), and hence the Céa estimate provided by the discrete Babuška–Brezzi theory (cf. Theorems 2.5 and 2.6 in Ref. 19) gives the existence of a positive constant $\widehat{C}$, depending only on c_a, c_b, $\|A\|$, and $\|B\|$, such that

$$\|((u, \lambda), \rho) - ((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h)\|_{H \times Q} \leq \widehat{C} \operatorname{dist}(((u, \lambda), \rho), H_h \times Q) = \widehat{C} \operatorname{dist}(u, \mathcal{V}_h), \tag{2.24}$$

where the last equality arises from the fact that λ and ρ belong to Q. On the other hand, the linear character of the discrete problem (2.20) readily implies that the difference $((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h) - ((u_h, \lambda_h), \rho_h)$ is the unique solution of it when G_{z_h} is replaced there by $G_u - G_{u_h}$, and therefore, the *a priori* estimate (2.21) and the assumption (A2) yield

$$\begin{aligned}\|((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h) - ((u_h, \lambda_h), \rho_h)\| &\leq C_T \|G_u - G_{u_h}\|_{H'} \\ &= \alpha C_T \|\nabla D(u) - \nabla D(u_h)\|_{\mathcal{V}} \leq \alpha C_T L_{\mathcal{D}} \|i_W\| \|u - u_h\|_{\mathcal{V}}. \end{aligned} \tag{2.25}$$

Finally, the required estimate (2.23) follows easily from triangle inequality, (2.24), (2.25), and the hypothesis $\alpha C_T L_{\mathcal{D}} \|i_W\| \leq 1 - \delta$. □

We end this section by stressing that the main assumption in Theorem 2.3 is handled by choosing a particular value of δ. Certainly, the closer to 1, the smaller the constant $\delta^{-1}\widehat{C}$ in the Céa estimate, but then the hypothesis $\alpha C_T L_{\mathcal{D}} \|i_W\| \leq 1 - \delta$, with $1 - \delta$ approaching 0, is more demanding on the constants involved. Conversely, the closer to 0, the hypothesis is less restrictive, but then the constant in the Céa estimate blows up. According to the above, it seems more reasonable to consider the midpoint of the range for δ, that is $\delta = 1/2$, which yields the assumption $\alpha C_T L_{\mathcal{D}} \|i_W\| \leq 1/2$, and the corresponding Céa estimate

$$\|((u, \lambda), \rho) - ((u_h, \lambda_h), \rho_h)\|_{H \times Q} \leq 2\widehat{C} \mathrm{dist}(u, \mathcal{V}_h). \tag{2.26}$$

2.4. *The rates of convergence*

For the sake of exposition and clearness, we now assume $\mathcal{V} = \mathbf{H}^1(\Omega) := [\mathrm{H}^1(\Omega)]^2$, which is precisely the case of the application to an elastic energy that we report later on in Sec. 5. In there, the unknown u of the abstract analyses from Secs. 2.1–2.3, and 4, becomes the respective displacement vector $\mathbf{u}$ of the elastic material.

Now, let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular triangulations of $\bar{\Omega}$ made of triangles K with diameter h_K, and define the meshsize $h := \max\{h_K : K \in \mathcal{T}_h\}$, which also acts as the index of $\mathcal{T}_h$. Then, given an integer $k \geq 1$, we denote by $\mathbf{P}_k(K) := [\mathrm{P}_k(K)]^2$ the space of polynomial vectors of degree $\leq k$ on K, introduce the Lagrange finite element subspace of $\mathcal{V}$ of order k

$$\mathcal{V}_h := \{\mathbf{v}_h \in \mathbf{H}^1(\Omega) : \mathbf{v}_h|_K \in \mathbf{P}_k(K) \ \forall\, K \in \mathcal{T}_h\}, \tag{2.27}$$

and let $\mathcal{L}_h : \mathbf{C}(\bar{\Omega}) := [C(\bar{\Omega})]^2 \to \mathcal{V}_h$ be its associated interpolation operator. It is well known that there holds the following approximation property (cf. Ref. 10):

($\mathbf{AP}_h^{\mathbf{u}}$) for each $m \in \{1, \ldots, k+1\}$ there exists a positive constant C_m such that

$$(\mathbf{v}, \mathcal{V}_h) \leq \|\mathbf{v} - \mathcal{L}_h(\mathbf{v})\|_{1,\Omega} \leq C_m h^{m-1} |\mathbf{v}|_{m,\Omega} \quad \forall\, \mathbf{v} \in \mathbf{H}^m(\Omega) := [\mathrm{H}^m(\Omega)]^2. \tag{2.28}$$

Then, as a straightforward consequence of Theorem 2.3, (2.26), and ($\mathbf{AP}_h^{\mathbf{u}}$), and analogously to Ref. 7, we obtain the following convergence result.

Theorem 2.4. *Assume that* $\alpha C_T L_{\mathcal{D}} \|i_W\| \leq 1/2$, *and let* $((\mathbf{u},\lambda),\rho) \in H \times Q$ *and* $((\mathbf{u}_h,\lambda_h),\rho_h) \in H_h \times Q$ *be the unique solutions of* (2.13) *and* (2.19), *respectively. In addition, suppose that* $\mathbf{u} \in \mathbf{H}^m(\Omega)$, *for some* $m \in \{1,\ldots,k+1\}$. *Then, there holds*

$$\|((\mathbf{u},\lambda),\rho) - ((\mathbf{u}_h,\lambda_h),\rho_h)\|_{H\times Q} \leq 2\widehat{C} C_m h^{m-1} |\mathbf{u}|_{m,\Omega}. \tag{2.29}$$

Furthermore, in what follows, we apply usual duality arguments to derive the rate of convergence for the error $\mathbf{u}-\mathbf{u}_h$, but measured in the weaker norm $\|\cdot\|_{0,\Omega}$. For this purpose, we now simplify the writing of the vector versions of (2.13) and (2.19) by introducing the bilinear form arising after adding the expressions on the left-hand side of either one, that is we let $\mathcal{A} : (H\times Q)\times(H\times Q) \to \mathrm{R}$ be defined as

$$\mathcal{A}((\vec{\mathbf{w}},\chi),(\vec{\mathbf{v}},\xi)) := A(\vec{\mathbf{w}},\vec{\mathbf{v}}) + B(\vec{\mathbf{v}},\chi) + B(\vec{\mathbf{w}},\xi),$$

for all $\vec{\mathbf{w}} := (\mathbf{w},\vartheta)$, $\vec{\mathbf{v}} := (\mathbf{v},\eta) \in H := \mathcal{V}\times Q$, for all $\chi,\xi \in Q$. In this way, (2.13) and (2.19) can be rewritten, respectively, as: Find $(\vec{\mathbf{u}},\rho) := ((\mathbf{u},\lambda),\rho) \in H\times Q$ such that

$$\mathcal{A}((\vec{\mathbf{u}},\rho),(\vec{\mathbf{v}},\xi)) = G_{\mathbf{u}}(\vec{\mathbf{v}}) \quad \forall\,(\vec{\mathbf{v}},\xi) := ((\mathbf{v},\eta),\xi) \in H\times Q, \tag{2.30}$$

and: Find $(\vec{\mathbf{u}}_h,\rho_h) := ((\mathbf{u}_h,\lambda_h),\rho_h) \in H_h \times Q$ such that

$$\mathcal{A}((\vec{\mathbf{u}}_h,\rho_h),(\vec{\mathbf{v}}_h,\xi_h)) = G_{\mathbf{u}_h}(\vec{\mathbf{v}}_h) \quad \forall\,(\vec{\mathbf{v}}_h,\xi_h) := ((\mathbf{v}_h,\eta_h),\xi_h) \in H_h\times Q. \tag{2.31}$$

Note that $\mathcal{A}$ is obviously bounded with a corresponding constant $\|\mathcal{A}\|$ depending on $\|A\|$ and $\|B\|$.

Next, we let $(\vec{\mathbf{w}},\chi) := ((\mathbf{w},\vartheta),\chi) \in H\times Q$ be the unique solution, guaranteed by Lemma 2.1 and the symmetry of $\mathcal{A}$, of the continuous problem

$$\mathcal{A}((\vec{\mathbf{v}},\xi),(\vec{\mathbf{w}},\chi)) = \int_\Omega (\mathbf{u}-\mathbf{u}_h)\cdot\mathbf{v} \quad \forall\,(\vec{\mathbf{v}},\xi) := ((\mathbf{v},\eta),\xi) \in H\times Q, \tag{2.32}$$

and consider the following regularity assumption:

($\mathbf{RA^w}$) there holds $\mathbf{w} \in \mathrm{H}^2(\Omega)$ and there exists a positive constant $C_{\mathtt{reg}}$, independent of $\mathbf{w}$ and h, such that

$$\|\mathbf{w}\|_{2,\Omega} \leq C_{\mathtt{reg}} \|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}. \tag{2.33}$$

In addition, throughout the rest of the section, we assume $\mathcal{W} = \mathbf{L}^2(\Omega)$ in (A2). Then, we are able to prove the following result, which establishes an extra $O(h)$ for the rate of convergence of $\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}$.

Theorem 2.5. *In addition to the hypotheses of Theorem* 2.3 *with* $\delta = 1/2$, *assume* ($\mathbf{RA^w}$) *and that* $\alpha L_{\mathcal{D}}(C_2+1)C_{\mathtt{reg}} \leq 1/2$. *Then, there exists a positive constant* C_0, *depending only on* $\|\mathcal{A}\|$, $\widehat{C}$, C_2 (*cf.* (2.28)), *and* $C_{\mathtt{reg}}$ (*cf.* (2.33)), *and hence independent of* h, *such that*

$$\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega} \leq C_0 h \mathrm{dist}(\mathbf{u},\mathcal{V}_h). \tag{2.34}$$

In particular, if $\mathbf{u} \in \mathbf{H}^m(\Omega)$, *with* $m \in \{1, \ldots, k+1\}$, *there holds*

$$\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \leq \widetilde{C}_0 h^m |\mathbf{u}|_{m,\Omega}, \tag{2.35}$$

with $\widetilde{C}_0 := C_m C_0$.

Proof. We begin by taking $(\vec{\mathbf{v}}, \xi) = (\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h)$ in (2.32), which yields

$$\|\mathbf{u} - \mathbf{u}_h\|^2_{0,\Omega} = \mathcal{A}((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi)),$$

and by recalling from the Sobolev embedding theorem that $\mathbf{H}^2(\Omega) \subseteq \mathbf{C}(\bar{\Omega})$, which implies, according to $(\mathbf{RA^w})$, that $\mathbf{w} \in \mathbf{C}(\bar{\Omega})$. Thus, adding and subtracting $(\vec{\mathbf{w}}_h, \chi_h) := ((\mathcal{L}_h(\mathbf{w}), \vartheta), \chi) \in H_h \times Q$ in the second component of $\mathcal{A}$, and then using (2.30), (2.31), and the definition of the functional $G_{\mathbf{z}}$ (cf. vector version of (2.3) and (2.12)), we obtain from the foregoing equation

$$\begin{aligned}
\|\mathbf{u} - \mathbf{u}_h\|^2_{0,\Omega} &= \mathcal{A}((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi) - (\vec{\mathbf{w}}_h, \chi_h)) + \mathcal{A}((\vec{\mathbf{u}}, \rho) \\
&\quad - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}_h, \chi_h)) \\
&= \mathcal{A}((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi) - (\vec{\mathbf{w}}_h, \chi_h)) + G_{\mathbf{u}}(\vec{\mathbf{w}}_h) - G_{\mathbf{u}_h}(\vec{\mathbf{w}}_h) \\
&= \mathcal{A}((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi) - (\vec{\mathbf{w}}_h, \chi_h)) + \alpha \langle \nabla \mathcal{D}(\mathbf{u}_h) \\
&\quad - \nabla \mathcal{D}(\mathbf{u}), \mathbf{w}_h \rangle.
\end{aligned} \tag{2.36}$$

Next, employing now the boundedness of $\mathcal{A}$, the assumption (A2), the estimate (2.26), the approximation property (2.28) for $\mathcal{L}_h$, and the regularity bound (2.33), we deduce from (2.36) that

$$\begin{aligned}
\|\mathbf{u} - \mathbf{u}_h\|^2_{0,\Omega} &\leq \|\mathcal{A}\| \|(\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h)\| \|\mathbf{w} - \mathcal{L}_h(\mathbf{w})\|_{1,\Omega} \\
&\quad + \alpha L_{\mathcal{D}} \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \|\mathbf{w}_h\|_{1,\Omega} \\
&\leq \|\mathcal{A}\| 2\widehat{C} \mathrm{dist}(\mathbf{u}, \mathcal{V}_h) C_2 h |\mathbf{w}|_{2,\Omega} + \alpha L_{\mathcal{D}} \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \|\mathbf{w}_h\|_{1,\Omega} \\
&\leq Ch \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \mathrm{dist}(\mathbf{u}, \mathcal{V}_h) + \alpha L_{\mathcal{D}} \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \|\mathbf{w}_h\|_{1,\Omega},
\end{aligned} \tag{2.37}$$

with $C := 2\|\mathcal{A}\| \widehat{C} C_2 C_{\mathtt{reg}}$, which yields

$$\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \leq Ch \, \mathrm{dist}(\mathbf{u}, \mathcal{V}_h) + \alpha L_{\mathcal{D}} \|\mathbf{w}_h\|_{1,\Omega}. \tag{2.38}$$

In turn, applying again (2.28) and (2.33), and assuming for sake of simplicity that $h \leq 1$, we find that

$$\begin{aligned}
\|\mathbf{w}_h\|_{1,\Omega} &\leq \|\mathbf{w} - \mathbf{w}_h\|_{1,\Omega} + \|\mathbf{w}\|_{1,\Omega} \leq (C_2 h + 1) \|\mathbf{w}\|_{2,\Omega} \\
&\leq (C_2 + 1) C_{\mathtt{reg}} \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega},
\end{aligned}$$

which, replaced back into (2.38), leads to (2.34) with $C_0 = 2C$. Finally, it is straightforward to see that (2.28) and (2.34) imply (2.35), which completes the proof. □

As a particular case of (2.35), we notice that for $k = 1$ and $\mathbf{u} \in \mathbf{H}^2(\Omega)$ there holds the error estimate $\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \leq \widetilde{C}_0 h^2 |\mathbf{u}|_{2,\Omega}$, that is $\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} = O(h^2)$.

This rate of convergence will be illustrated below in Sec. 5 with some numerical results.

3. Extended Mixed Formulation and Application to Elastic Energies

In this section, we present and analyze a dual-mixed formulation of problem (2.13) in the particular case of an elastic energy. In this regard, we find it important to remark in advance that the setting and analysis to be considered and developed, respectively, in what follows, do not correspond to a straightforward application of those from Secs. 2.1 and 2.2, which basically refer to a primal formulation, but to a modification of them yielding the associated extended mixed approach to be employed here. Still, the point of departure for this novel model is the use of an elastic regularizer with Neumann boundary conditions, which presents a nontrivial kernel.

3.1. *Setting of the problem*

Let $\mathcal{C} : \mathbb{L}^2(\Omega) \longrightarrow \mathbb{L}^2(\Omega)$ be the Hooke operator defined by

$$\mathcal{C}\boldsymbol{\tau} := \lambda_s \mathrm{tr}(\boldsymbol{\tau})\mathbb{I} + 2\mu_s \boldsymbol{\tau} \quad \forall\, \boldsymbol{\tau} \in \mathbb{L}^2(\Omega), \tag{3.1}$$

where λ_s and μ_s are the associated Lamé parameters, and let $\varepsilon(\mathbf{u}) := \frac{1}{2}\{(\nabla \mathbf{u}) + (\nabla \mathbf{u})^{\mathtt{t}}\}$ be the strain rate tensor, also known as the symmetric component of $\nabla \mathbf{u}$. Then, letting $\mathcal{V} := \mathbf{H}^1(\Omega)$, the bilinear form a from Sec. 2 is defined as

$$a(\mathbf{w}, \mathbf{v}) := \int_\Omega \mathcal{C}\varepsilon(\mathbf{w}) : \varepsilon(\mathbf{v}) \quad \forall\, \mathbf{w}, \mathbf{v} \in \mathcal{V}, \tag{3.2}$$

and its kernel Q is given by the subspace of $\mathcal{V}$ determined by the rigid motions, that is

$$Q := \left\langle \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix} \right\} \right\rangle. \tag{3.3}$$

Next, we introduce the auxiliary unknown $\boldsymbol{\sigma} := \mathcal{C}\varepsilon(\mathbf{u})$, and observe that there holds

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathtt{t}} \quad \text{and} \quad \mathcal{C}^{-1}\boldsymbol{\sigma} = \nabla \mathbf{u} - \boldsymbol{\Phi} \quad \text{in} \quad \Omega, \tag{3.4}$$

where the rotation $\boldsymbol{\Phi} := \frac{1}{2}\{(\nabla \mathbf{u}) - (\nabla \mathbf{u})^{\mathtt{t}}\}$ is considered as a further unknown as well. In addition, we look for rigid motions $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ such that

$$-\mathbf{div}\boldsymbol{\sigma} + \boldsymbol{\rho} = -\alpha \nabla \mathcal{D}(\mathbf{u}), \quad \boldsymbol{\lambda} = \Pi_Q \mathbf{u}, \quad \text{and} \quad \boldsymbol{\rho} = \beta \boldsymbol{\lambda} \quad \text{in} \quad \Omega, \tag{3.5}$$

where α and β are the analogue parameters from Sec. 2, and incorporate the Neumann boundary condition

$$\boldsymbol{\sigma}\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \Gamma. \tag{3.6}$$

We now proceed to derive the variational formulation of (3.4)–(3.6). In fact, recalling that the definition of $\mathbb{H}(\mathbf{div};\Omega)$ was provided in Sec. 1, we first define the spaces

$$\mathbb{H}_0(\mathbf{div};\Omega) := \{\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div};\Omega) : \boldsymbol{\tau}\boldsymbol{\nu} = \mathbf{0} \text{ on } \Gamma\},$$

and

$$\mathbb{L}^2_{\mathtt{skew}}(\Omega) := \{\boldsymbol{\Psi} \in \mathbb{L}^2(\Omega) : \boldsymbol{\Psi}^{\mathtt{t}} = -\boldsymbol{\Psi}\},$$

noting in advance that $\boldsymbol{\sigma}$ and $\boldsymbol{\Phi}$ will be sought in $\mathbb{H}_0(\mathbf{div};\Omega)$ and $\mathbb{L}^2_{\mathtt{skew}}(\Omega)$, respectively. Thus, performing the tensor inner product of the second equation in (3.4) with an arbitrary $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega)$, integrating by parts, and using the boundary condition that holds for $\boldsymbol{\tau}$, we obtain

$$\int_\Omega \mathcal{C}^{-1}\boldsymbol{\sigma} : \boldsymbol{\tau} + \int_\Omega \boldsymbol{\Phi} : \boldsymbol{\tau} + \int_\Omega \mathbf{u} \cdot \mathbf{div}\boldsymbol{\tau} = 0 \quad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega). \tag{3.7}$$

In addition, testing the first and third equations in (3.5) against $\mathbf{v} \in \mathbf{L}^2(\Omega)$ and $\boldsymbol{\xi} \in Q$, respectively, and rewriting the second equation in (3.5) as the equivalent orthogonality condition, we find that

$$\int_\Omega \mathbf{v} \cdot \mathbf{div}\boldsymbol{\sigma} - \int_\Omega \boldsymbol{\rho} \cdot \mathbf{v} = \alpha \int_\Omega \nabla\mathcal{D}(\mathbf{u}) \cdot \mathbf{v} \quad \forall\, \mathbf{v} \in \mathbf{L}^2(\Omega), \tag{3.8}$$

$$\int_\Omega (\boldsymbol{\rho} - \beta\boldsymbol{\lambda}) \cdot \boldsymbol{\xi} = 0 \quad \forall\, \boldsymbol{\xi} \in Q, \tag{3.9}$$

and

$$\int_\Omega (\boldsymbol{\lambda} - \mathbf{u}) \cdot \boldsymbol{\eta} = 0 \quad \forall\, \boldsymbol{\eta} \in Q. \tag{3.10}$$

Finally, the symmetry of $\boldsymbol{\sigma}$ (first equation in (3.4)) is imposed weakly as

$$\int_\Omega \boldsymbol{\Psi} : \boldsymbol{\sigma} = 0 \quad \forall\, \boldsymbol{\Psi} \in \mathbb{L}^2_{\mathtt{skew}}(\Omega). \tag{3.11}$$

Therefore, incorporating (3.10) into (3.7), and adding (3.8), (3.9), and (3.11), we arrive at the following dual-mixed variational formulation of (3.4)–(3.6): Find $\vec{\boldsymbol{\sigma}} := (\boldsymbol{\sigma}, \boldsymbol{\rho}) \in \mathbf{H} := \mathbb{H}_0(\mathbf{div};\Omega) \times Q$ and $\vec{\mathbf{u}} := (\mathbf{u}, \boldsymbol{\Phi}, \boldsymbol{\lambda}) \in \mathbf{Q} := \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathtt{skew}}(\Omega) \times Q$, such that

$$\begin{aligned} &\mathbf{a}(\vec{\boldsymbol{\sigma}}, \vec{\boldsymbol{\tau}}) + \mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{u}}) = 0 && \forall\, \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}, \\ &\mathbf{b}(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{v}}) - \mathbf{c}(\vec{\mathbf{u}}, \vec{\mathbf{v}}) = \alpha \mathbf{F}_{\mathbf{u}}(\vec{\mathbf{v}}) && \forall\, \vec{\mathbf{v}} := (\mathbf{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}, \end{aligned} \tag{3.12}$$

where $\mathbf{a} : \mathbf{H} \times \mathbf{H} \to \mathrm{R}$, $\mathbf{b} : \mathbf{H} \times \mathbf{Q} \to \mathrm{R}$, and $\mathbf{c} : \mathbf{Q} \times \mathbf{Q} \to \mathrm{R}$, are the bilinear forms defined as

$$\mathbf{a}(\vec{\boldsymbol{\zeta}}, \vec{\boldsymbol{\tau}}) := \int_\Omega \mathcal{C}^{-1}\boldsymbol{\zeta} : \boldsymbol{\tau}, \tag{3.13}$$

$$\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{v}}) := \int_\Omega \mathbf{v} \cdot \mathbf{div}\boldsymbol{\tau} + \int_\Omega \boldsymbol{\Psi} : \boldsymbol{\tau} + \int_\Omega (\boldsymbol{\xi} - \mathbf{v}) \cdot \boldsymbol{\eta}, \tag{3.14}$$

and

$$\mathbf{c}(\vec{\mathbf{w}},\vec{\mathbf{v}}) := \beta \int_\Omega \boldsymbol{\vartheta}\cdot\boldsymbol{\xi}, \tag{3.15}$$

for all $\vec{\boldsymbol{\zeta}} := (\boldsymbol{\zeta},\chi), \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H}$, for all $\vec{\mathbf{w}} := (\mathbf{w},\boldsymbol{\Upsilon},\boldsymbol{\vartheta}), \vec{\mathbf{v}} := (\mathbf{v},\boldsymbol{\Psi},\boldsymbol{\xi}) \in \mathbf{Q}$. In turn, given $\vec{\mathbf{w}} := (\mathbf{w},\boldsymbol{\Upsilon},\boldsymbol{\vartheta}) \in \mathbf{Q}$, the linear functional $\mathbf{F}_{\mathbf{w}} : \mathbf{Q} \to \mathbb{R}$ is defined by

$$\mathbf{F}_{\mathbf{w}}(\vec{\mathbf{v}}) := \int_\Omega \nabla\mathcal{D}(\mathbf{w})\cdot\mathbf{v} \quad \forall\,\vec{\mathbf{v}} := (\mathbf{v},\boldsymbol{\Psi},\boldsymbol{\xi}) \in \mathbf{Q}. \tag{3.16}$$

At this point, we stress that $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are all bounded bilinear forms with respect to the usual norms of the product spaces $\mathbf{H}$ and $\mathbf{Q}$, that is

$$\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}} := \{\|\boldsymbol{\tau}\|^2_{\mathbf{div};\Omega} + \|\boldsymbol{\eta}\|^2_{0,\Omega}\}^{1/2} \quad \forall\,\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H},$$

and

$$\|\vec{\mathbf{v}}\|_{\mathbf{Q}} := \{\|\mathbf{v}\|^2_{0,\Omega} + \|\boldsymbol{\Psi}\|^2_{0,\Omega} + \|\boldsymbol{\xi}\|^2_{0,\Omega}\}^{1/2} \quad \forall\,\vec{\mathbf{v}} := (\mathbf{v},\boldsymbol{\Psi},\boldsymbol{\xi}) \in \mathbf{Q}.$$

Moreover, $\mathbf{a}$ and $\mathbf{c}$ are both symmetric and positive semi-definite, that is

$$\mathbf{a}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{\tau}}) \geq 0 \quad \forall\,\vec{\boldsymbol{\tau}} \in \mathbf{H} \quad \text{and} \quad \mathbf{c}(\vec{\mathbf{v}},\vec{\mathbf{v}}) \geq 0 \quad \forall\,\vec{\mathbf{v}} \in \mathbf{Q}. \tag{3.17}$$

In addition, it is clear that $\mathbf{F}_{\mathbf{w}}$ is bounded for each $\mathbf{w} \in \mathbf{L}^2(\Omega)$.

3.2. *Analysis of the continuous formulation*

In order to study the solvability of (3.12), and similarly to the analysis in Sec. 2.2, we now introduce the operator $\mathbf{T} : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$ defined by $\mathbf{T}(\mathbf{z}) := \underline{\mathbf{u}}$ for each $\mathbf{z} \in \mathbf{L}^2(\Omega)$, where $\vec{\underline{\boldsymbol{\sigma}}} := (\underline{\boldsymbol{\sigma}},\underline{\boldsymbol{\rho}}) \in \mathbf{H}$ and $\vec{\underline{\mathbf{u}}} := (\underline{\mathbf{u}},\underline{\boldsymbol{\Phi}},\underline{\boldsymbol{\lambda}}) \in \mathbf{Q}$ are such that

$$\begin{aligned} &\mathbf{a}(\vec{\underline{\boldsymbol{\sigma}}},\vec{\boldsymbol{\tau}}) + \mathbf{b}(\vec{\boldsymbol{\tau}},\vec{\underline{\mathbf{u}}}) = 0 \qquad\qquad \forall\,\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H}, \\ &\mathbf{b}(\vec{\underline{\boldsymbol{\sigma}}},\vec{\mathbf{v}}) - \mathbf{c}(\vec{\underline{\mathbf{u}}},\vec{\mathbf{v}}) = \alpha\mathbf{F}_{\mathbf{z}}(\vec{\mathbf{v}}) \quad \forall\,\vec{\mathbf{v}} := (\mathbf{v},\boldsymbol{\Psi},\boldsymbol{\xi}) \in \mathbf{Q}. \end{aligned} \tag{3.18}$$

We remark here that solving (3.12) is equivalent to seeking a fixed point of $\mathbf{T}$, that is: Find $\mathbf{u} \in \mathbf{L}^2(\Omega)$ such that $\mathbf{T}(\mathbf{u}) = \mathbf{u}$. The following abstract result will allow us to show below that, given $\mathbf{z} \in \mathbf{L}^2(\Omega)$, the linear problem (3.18) is well-posed, thus confirming that the operator $\mathbf{T}$ is well-defined.

Theorem 3.1. *Let* $\mathbf{H}$ *and* $\mathbf{Q}$ *be real Hilbert spaces, and let* $\mathbf{a} : \mathbf{H}\times\mathbf{H} \to \mathrm{R}$, $\mathbf{b} : \mathbf{H}\times\mathbf{Q} \to \mathrm{R}$, *and* $\mathbf{c} : \mathbf{Q}\times\mathbf{Q} \to \mathrm{R}$ *be bounded bilinear forms with induced bounded linear operators* $\mathbf{A} : \mathbf{H} \to \mathbf{H}'$, $\mathbf{B} : \mathbf{H} \to \mathbf{Q}'$, $\mathbf{B}^{\mathtt{t}} : \mathbf{Q} \to \mathbf{H}'$, *and* $\mathbf{C} : \mathbf{Q} \to \mathbf{Q}'$, *defined, respectively, by the identities*

$$\begin{aligned} &\mathbf{A}(\boldsymbol{\zeta})(\boldsymbol{\tau}) := \mathbf{a}(\boldsymbol{\zeta},\boldsymbol{\tau}) \quad \forall\,\boldsymbol{\zeta},\boldsymbol{\tau} \in \mathbf{H}, \\ &\mathbf{B}(\boldsymbol{\tau})(\mathbf{v}) = \mathbf{B}^{\mathtt{t}}(\mathbf{v})(\boldsymbol{\tau}) := \mathbf{b}(\boldsymbol{\tau},\mathbf{v}) \quad \forall\,\boldsymbol{\tau} \in \mathbf{H}, \forall\,\mathbf{v} \in \mathbf{Q}, \\ &\mathbf{C}(\mathbf{w})(\mathbf{v}) := \mathbf{c}(\mathbf{w},\mathbf{v}) \quad \forall\,\mathbf{w},\mathbf{v} \in \mathbf{Q}. \end{aligned}$$

In turn, let $\mathbf{K} = N(\mathbf{B})$ *and* $\mathbf{V} = N(\mathbf{B}^{\mathtt{t}})$*, and assume the following hypotheses*:

(i) $\mathbf{a}$ *and* $\mathbf{c}$ *are symmetric and positive semi-definite.*

(ii) $\mathbf{a}$ *is* $\mathbf{K}$*-elliptic, that is there exists a positive constant* $\alpha_{\mathbf{K}}$ *such that*

$$\mathbf{a}(\boldsymbol{\tau}, \boldsymbol{\tau}) \geq \alpha_{\mathbf{K}} \|\boldsymbol{\tau}\|_{\mathbf{H}}^2 \quad \forall\, \boldsymbol{\tau} \in \mathbf{K}.$$

(iii) $R(\mathbf{B})$ *is closed, that is there exists a positive constant* $\beta_{\mathbf{B}}$ *such that*

$$\sup_{\substack{\boldsymbol{\tau} \in \mathbf{H} \\ \boldsymbol{\tau} \neq \mathbf{0}}} \frac{\mathbf{b}(\boldsymbol{\tau}, \mathbf{v})}{\|\boldsymbol{\tau}\|_{\mathbf{H}}} \geq \beta_{\mathbf{B}} \|\mathbf{v}\|_{\mathbf{Q}} \quad \forall\, \mathbf{v} \in \mathbf{V}^{\perp},$$

or equivalently

$$\sup_{\substack{\mathbf{v} \in \mathbf{Q} \\ \mathbf{v} \neq \mathbf{0}}} \frac{\mathbf{b}(\boldsymbol{\tau}, \mathbf{v})}{\|\mathbf{v}\|_{\mathbf{Q}}} \geq \beta_{\mathbf{B}} \|\boldsymbol{\tau}\|_{\mathbf{H}} \quad \forall\, \boldsymbol{\tau} \in \mathbf{K}^{\perp}.$$

(iv) $\mathbf{c}$ *is* $\mathbf{V}$*-elliptic, that is there exists a positive constant* $\gamma_{\mathbf{V}}$ *such that*

$$\mathbf{c}(\mathbf{v}, \mathbf{v}) \geq \gamma_{\mathbf{V}} \|\mathbf{v}\|_{\mathbf{Q}}^2 \quad \forall\, \mathbf{v} \in \mathbf{V}.$$

Then, for each pair $(\mathbf{F}, \mathbf{G}) \in \mathbf{H}' \times \mathbf{Q}'$ *there exists a unique* $(\boldsymbol{\sigma}, \mathbf{u}) \in \mathbf{H} \times \mathbf{Q}$ *solution to*

$$\begin{aligned} \mathbf{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}) + \mathbf{b}(\boldsymbol{\tau}, \mathbf{u}) &= \mathbf{F}(\boldsymbol{\tau}) \quad \forall\, \boldsymbol{\tau} \in \mathbf{H}, \\ \mathbf{b}(\boldsymbol{\sigma}, \mathbf{v}) - \mathbf{c}(\mathbf{u}, \mathbf{v}) &= \mathbf{G}(\mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{Q}. \end{aligned} \tag{3.19}$$

In addition, there exists a positive constant C*, depending only on* $\alpha_{\mathbf{K}}$*,* $\beta_{\mathbf{B}}$*,* $\gamma_{\mathbf{V}}$*,* $\|\mathbf{A}\|$*, and* $\|\mathbf{C}\|$*, such that*

$$\|\boldsymbol{\sigma}\|_{\mathbf{H}} + \|\mathbf{u}\|_{\mathbf{Q}} \leq C\{\|\mathbf{F}\|_{\mathbf{H}'} + \|\mathbf{G}\|_{\mathbf{Q}'}\}.$$

Proof. See Theorem 4.3.1 in Ref. 9. □

We now apply Theorem 3.1 to show the well-posedness of (3.18), and hence the well-definiteness of the associated operator $\mathbf{T}$. To this end, we first rewrite the bilinear form $\mathbf{b}$ (cf. (3.14)) as

$$\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{v}}) := \int_{\Omega} \mathbf{v} \cdot \{\mathbf{div}\boldsymbol{\tau} - \boldsymbol{\eta}\} + \int_{\Omega} \boldsymbol{\Psi} : \boldsymbol{\tau} + \int_{\Omega} \boldsymbol{\xi} \cdot \boldsymbol{\eta}, \tag{3.20}$$

for all $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}$, for all $\vec{\mathbf{v}} := (\mathbf{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}$, from which we deduce that the null space of its induced operator $\mathbf{B} : \mathbf{H} \to \mathbf{Q}'$ is given by

$$\mathbf{K} = N(\mathbf{B}) := \{\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H} : \mathbf{div}\boldsymbol{\tau} - \boldsymbol{\eta} = 0, \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathtt{t}}, \quad \text{and} \quad \boldsymbol{\eta} = \mathbf{0}\},$$

which yields

$$\mathbf{K} = \{\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H} : \mathbf{div}\boldsymbol{\tau} = \mathbf{0}, \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathtt{t}}, \quad \text{and} \quad \boldsymbol{\eta} = \mathbf{0}\}. \tag{3.21}$$

Similarly, looking at the original definition (3.14) of $\mathbf{b}$, we readily find that

$$\mathbf{V} = N(\mathbf{B}^{\mathtt{t}}) := \left\{ \vec{\mathbf{v}} := (\mathbf{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q} : \int_\Omega \mathbf{v} \cdot \mathbf{div}\boldsymbol{\tau} + \int_\Omega \boldsymbol{\Psi} : \boldsymbol{\tau} = 0 \right.$$
$$\left. \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \quad \text{and} \quad \boldsymbol{\xi} = \mathit{\Pi}_Q \mathbf{v} \right\},$$

from which, rewriting the expression involving $\boldsymbol{\tau}$ in the distributional sense, we are lead to

$$\mathbf{V} = \{ \vec{\mathbf{v}} := (\mathbf{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q} : \boldsymbol{\Psi} = \nabla \mathbf{v} \text{ in } \mathcal{D}'(\Omega) \text{ and } \boldsymbol{\xi} = \mathit{\Pi}_Q \mathbf{v} \}.$$

Moreover, the fact that $\nabla \mathbf{v} = \boldsymbol{\Psi} \in \mathbb{L}^2_{\mathtt{skew}}(\Omega)$ implies that $\varepsilon(\mathbf{v}) = \mathbf{0}$, that is $\mathbf{v}$ lies in the subspace of rigid motions Q, and therefore $\mathbf{V} \subseteq \mathbf{V}_0$, where

$$\mathbf{V}_0 := \{ \vec{\mathbf{q}} := (\mathbf{q}, \nabla \mathbf{q}, \mathbf{q}) \in \mathbf{Q} : \mathbf{q} \in Q \}. \tag{3.22}$$

Conversely, it is easy to see that, given $\vec{\mathbf{q}} \in \mathbf{V}_0$, there holds $\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{q}}) = 0$ for all $\vec{\boldsymbol{\tau}} \in \mathbf{H}$ (see also (3.26)), which shows that $\mathbf{V}_0 \subseteq \mathbf{V}$, and hence $\mathbf{V} = \mathbf{V}_0$.

We now aim to show the $\mathbf{K}$-ellipticity of $\mathbf{a}$, for which we first state two preliminary results that are based on the decomposition $\mathbb{H}(\mathbf{div}; \Omega) := \widetilde{\mathbb{H}}(\mathbf{div}; \Omega) \oplus \mathrm{R}\mathbb{I}$, where

$$\widetilde{\mathbb{H}}(\mathbf{div}; \Omega) := \left\{ \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega) : \int_\Omega \mathrm{tr}(\boldsymbol{\tau}) = 0 \right\}.$$

In fact, we have the following lemmas, in which we use that for each $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega)$ there exist unique $\boldsymbol{\tau}_0 \in \widetilde{\mathbb{H}}(\mathbf{div}; \Omega)$ and $d \in \mathrm{R}$ such that $\boldsymbol{\tau} = \boldsymbol{\tau}_0 + d\mathbb{I} \in \mathbb{H}(\mathbf{div}; \Omega)$.

Lemma 3.1. *There exists a positive constant c_1, depending only on Ω, such that*

$$\|\boldsymbol{\tau}^{\mathtt{d}}\|^2_{0,\Omega} + \|\mathbf{div}(\boldsymbol{\tau})\|^2_{0,\Omega} \geq c_1 \|\boldsymbol{\tau}_0\|^2_{0,\Omega} \quad \forall\, \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega). \tag{3.23}$$

Proof. See Proposition 3.1 of Chap. IV of Ref. 12 or Lemma 2.3 of Ref. 19. □

Lemma 3.2. *There exists a positive constant c_2, depending only on Ω, such that*

$$\|\boldsymbol{\tau}_0\|^2_{\mathbf{div};\Omega} \geq c_2 \|\boldsymbol{\tau}\|^2_{\mathbf{div};\Omega} \quad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega). \tag{3.24}$$

Proof. See Lemma 2.2 of Ref. 18 or Lemma 2.5 of Ref. 19. □

Then, the announced result for $\mathbf{a}$ is established as follows.

Lemma 3.3. *There exists a constant $\alpha_{\mathbf{K}} > 0$, independent of the Lamé parameter λ_s, such that*

$$\mathbf{a}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{\tau}}) \geq \alpha_{\mathbf{K}} \|\vec{\boldsymbol{\tau}}\|^2_{\mathbf{H}} \quad \forall\, \vec{\boldsymbol{\tau}} \in \mathbf{K}.$$

Proof. We begin by recalling from Sec. 2.4.3 of Ref. 19 that in the present 2D case the inverse $\mathcal{C}^{-1}$ of the Hooke tensor $\mathcal{C}$ becomes

$$\mathcal{C}^{-1}\boldsymbol{\tau} = \frac{1}{2\mu_s}\boldsymbol{\tau} - \frac{\lambda_s}{4\mu_s(\lambda_s + \mu_s)}\mathrm{tr}(\boldsymbol{\tau})\mathbb{I} \quad \forall\, \boldsymbol{\tau} \in \mathbb{L}^2(\Omega),$$

which, after some algebraic manipulations, yields (cf. Eqs. (2.48) and (2.52) of Ref. 19)

$$\mathbf{a}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{\tau}}) = \int_\Omega \mathcal{C}^{-1}\boldsymbol{\tau} : \boldsymbol{\tau} = \frac{1}{2\mu_s}\|\boldsymbol{\tau}^{\mathtt{d}}\|^2_{0,\Omega} + \frac{1}{4(\lambda_s+\mu_s)}\|\mathrm{tr}(\boldsymbol{\tau})\|^2_{0,\Omega} \geq \frac{1}{2\mu_s}\|\boldsymbol{\tau}^{\mathtt{d}}\|^2_{0,\Omega}$$

for all $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H}$. In particular, given $\vec{\boldsymbol{\tau}} \in \mathbf{K}$, that is $\boldsymbol{\eta} = \mathbf{0}$ and $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega)$ such that $\mathbf{div}(\boldsymbol{\tau}) = 0$ and $\boldsymbol{\tau} = \boldsymbol{\tau}^{\mathtt{t}}$, it follows from the foregoing inequality and straightforward applications of Lemmas 3.1 and 3.2, that

$$\mathbf{a}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{\tau}}) \geq \frac{c_1}{2\mu_s}\|\boldsymbol{\tau}_0\|^2_{0,\Omega} = \frac{c_1}{2\mu_s}\|\boldsymbol{\tau}_0\|^2_{\mathbf{div};\Omega} \geq \frac{c_1c_2}{2\mu_s}\|\boldsymbol{\tau}\|^2_{\mathbf{div};\Omega} = \frac{c_1c_2}{2\mu_s}\|\vec{\boldsymbol{\tau}}\|^2_{\mathbf{H}},$$

which completes the proof with the constant $\alpha_{\mathbf{K}} := \frac{c_1c_2}{2\mu_s}$. □

A preliminary continuous inf-sup condition for the bilinear form $\mathbf{b}$ (cf. (3.14)), in which the space $\mathbf{V}_0$ as such (cf. (3.22)) plays a key role, is established next.

Lemma 3.4. *There exists a positive constant* $\beta_{\mathbf{B}}$, *independent of the Lamé parameters, such that*

$$\sup_{\substack{\vec{\boldsymbol{\tau}}\in\mathbf{H}\\ \vec{\boldsymbol{\tau}}\neq\mathbf{0}}} \frac{\vec{\boldsymbol{\tau}},\vec{\mathbf{v}}}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \beta_{\mathbf{B}}\mathrm{dist}(\vec{\mathbf{v}},\mathbf{V}_0) \quad \forall\,\vec{\mathbf{v}} \in \mathbf{Q}. \tag{3.25}$$

Proof. While we already know that $\mathbf{V}_0 = \mathbf{V}$, the inclusion $\mathbf{V}_0 \subseteq \mathbf{V} = N(\mathbf{B}^{\mathtt{t}})$ suffices to realize that (3.25) trivially holds for $\vec{\mathbf{v}} \in \mathbf{V}_0$, and therefore in what follows, we prove for $\vec{\mathbf{v}} := (\mathbf{v},\boldsymbol{\Psi},\boldsymbol{\xi}) \in \mathbf{Q}\backslash\mathbf{V}_0$. Indeed, given an arbitrary $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H}$, we first use the orthogonal decomposition $\mathbf{v} = (\mathbf{v} - \boldsymbol{\Pi}_Q\mathbf{v}) + \boldsymbol{\Pi}_Q\mathbf{v} \in Q^\perp \oplus Q$, and then integrate by parts the expression $\int_\Omega \boldsymbol{\Pi}_Q\mathbf{v}\cdot\mathbf{div}\boldsymbol{\tau}$, to deduce from (3.14) that there holds

$$\mathbf{b}(\vec{\boldsymbol{\tau}},\vec{\mathbf{v}}) := \int_\Omega (\mathbf{v}-\boldsymbol{\Pi}_Q\mathbf{v})\cdot\mathbf{div}\boldsymbol{\tau} + \int_\Omega (\boldsymbol{\Psi}-\nabla\boldsymbol{\Pi}_Q\mathbf{v}) : \boldsymbol{\tau} + \int_\Omega (\boldsymbol{\xi}-\boldsymbol{\Pi}_Q\mathbf{v})\cdot\boldsymbol{\eta}. \tag{3.26}$$

Next, we proceed as in the proof of Lemma 3.4 in Ref. 20. In fact, assuming that $\mathbf{v} - \boldsymbol{\Pi}_Q\mathbf{v} \neq \mathbf{0}$, we let $\boldsymbol{\zeta} := \varepsilon(\mathbf{z})$ in Ω, where $\mathbf{z} \in \mathbf{H}^1(\Omega)$ is the unique solution, up to an element in Q, of the problem

$$\mathbf{div}(\varepsilon(\mathbf{z})) = \mathbf{v} - \boldsymbol{\Pi}_Q\mathbf{v} \quad \text{in} \quad \Omega, \quad \varepsilon(\mathbf{z})\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \Gamma. \tag{3.27}$$

Note that the compatibility condition required by this Neumann problem is satisfied thanks to the orthogonality relation $\int_\Omega(\mathbf{v}-\boldsymbol{\Pi}_Q\mathbf{v})\cdot\mathbf{q} = 0\ \forall\,\mathbf{q}\in Q$. Thus, it is clear that $\boldsymbol{\zeta} \in \mathbb{H}_0(\mathbf{div};\Omega)$ with $\mathbf{div}(\boldsymbol{\zeta}) = \mathbf{v} - \boldsymbol{\Pi}_Q\mathbf{v}$ and $\boldsymbol{\zeta} = \boldsymbol{\zeta}^{\mathtt{t}}$ in Ω. In addition, the corresponding continuous dependence result for (3.27) guarantees the existence of a positive constant C_N, independent of $\mathbf{v} - \boldsymbol{\Pi}_Q\mathbf{v}$, such that $\|\boldsymbol{\zeta}\|_{\mathbf{div};\Omega} \leq C_N\|\mathbf{v} - \boldsymbol{\Pi}_Q\mathbf{v}\|_{0,\Omega}$. In this way, defining $\vec{\boldsymbol{\zeta}} := (\boldsymbol{\zeta},\mathbf{0}) \in \mathbf{H}$, it readily follows that

$$\sup_{\substack{\vec{\boldsymbol{\tau}}\in\mathbf{H}\\ \vec{\boldsymbol{\tau}}\neq\mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}},\vec{\mathbf{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}},\vec{\mathbf{v}})}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}} = \frac{\|\mathbf{v}-\boldsymbol{\Pi}_Q\mathbf{v}\|^2_{0,\Omega}}{\|\boldsymbol{\zeta}\|_{\mathbf{div};\Omega}} \geq \frac{1}{C_N}\|\mathbf{v}-\boldsymbol{\Pi}_Q\mathbf{v}\|_{0,\Omega}. \tag{3.28}$$

In turn, if $\boldsymbol{\Psi} - \nabla \Pi_Q \mathbf{v} \neq \mathbf{0}$, a slight variation of the proof of Lemma 4.4 in Ref. 21 allows us to show that there exists $\boldsymbol{\zeta} \in \mathbb{H}_0(\mathbf{div}; \Omega)$ such that $\frac{1}{2}(\boldsymbol{\zeta} - \boldsymbol{\zeta}^{\mathtt{t}}) = \boldsymbol{\Psi} - \nabla \Pi_Q \mathbf{v}$ and $\|\boldsymbol{\zeta}\|_{\mathbf{div};\Omega} \leq c_N \|\boldsymbol{\Psi} - \nabla \Pi_Q \mathbf{v}\|_{0,\Omega}$, with a positive constant c_N, independent of $\boldsymbol{\Psi} - \nabla \Pi_Q \mathbf{v}$. Hence, setting $\vec{\boldsymbol{\zeta}} := (\boldsymbol{\zeta}, \mathbf{0}) \in \mathbf{H}$, we see that

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}}, \vec{\mathbf{v}})}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}} = \frac{\|\boldsymbol{\Psi} - \nabla \Pi_Q \mathbf{v}\|_{0,\Omega}^2 + \int_\Omega (\mathbf{v} - \Pi_Q \mathbf{v}) \cdot \mathbf{div} \boldsymbol{\zeta}}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}}$$

$$\geq \frac{1}{c_N} \|\boldsymbol{\Psi} - \nabla \Pi_Q \mathbf{v}\|_{0,\Omega} - \|\mathbf{v} - \Pi_Q \mathbf{v}\|_{0,\Omega}. \tag{3.29}$$

Furthermore, assuming that $\boldsymbol{\xi} - \Pi_Q \mathbf{v} \neq \mathbf{0}$, we define $\vec{\boldsymbol{\zeta}} := (\mathbf{0}, \boldsymbol{\xi} - \Pi_Q \mathbf{v}) \in \mathbf{H}$ and readily observe that

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}}, \vec{\mathbf{v}})}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}} = \|\boldsymbol{\xi} - \Pi_Q \mathbf{v}\|_{0,\Omega}. \tag{3.30}$$

In this way, since at least one of the components of $(\mathbf{v} - \Pi_Q \mathbf{v}, \boldsymbol{\Psi} - \nabla \Pi_Q \mathbf{v}, \boldsymbol{\xi} - \Pi_Q \mathbf{v})$ does not vanish, which follows from the fact that $\vec{\mathbf{v}} \notin \mathbf{V}_0$, a suitable linear combination of (3.28)–(3.30) implies the existence of a positive constant $\beta_{\mathbf{B}}$, depending on C_N and c_N, such that

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \beta_{\mathbf{B}} \|\vec{\mathbf{v}} - (\Pi_Q \mathbf{v}, \nabla \Pi_Q \mathbf{v}, \Pi_Q \mathbf{v})\|_{\mathbf{Q}}. \tag{3.31}$$

Finally, (3.31) and the fact that $(\Pi_Q \mathbf{v}, \nabla \Pi_Q \mathbf{v}, \Pi_Q \mathbf{v}) \in \mathbf{V}_0$ yield (3.25) and complete the proof. □

We remark here that the inf-sup condition (3.25) provides an alternative proof of the inclusion $\mathbf{V} \subseteq \mathbf{V}_0$, and hence of the identity $\mathbf{V} = \mathbf{V}_0$. In fact, for each $\vec{\mathbf{v}} \in \mathbf{V}$ there necessarily holds, due to (3.25), $\mathrm{dist}(\vec{\mathbf{v}}, \mathbf{V}_0) = 0$, which is obviously equivalent to saying $\vec{\mathbf{v}} \in \mathbf{V}_0$. Furthermore, as a direct corollary of Lemma 3.4, we now state the continuous inf-sup condition for $\mathbf{b}$ required by item (iii) of Theorem 3.1.

Lemma 3.5. *With the same constant $\beta_{\mathbf{B}}$ from Lemma* 3.4 *there holds*

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \beta_{\mathbf{B}} \|\vec{\mathbf{v}}\|_{\mathbf{Q}} \quad \forall\, \vec{\mathbf{v}} \in \mathbf{V}^{\perp}. \tag{3.32}$$

Proof. It suffices to use in (3.25) that $\mathrm{dist}(\vec{\mathbf{v}}, \mathbf{V}_0) = \mathrm{dist}(\vec{\mathbf{v}}, \mathbf{V}) = \|\vec{\mathbf{v}}\|_{\mathbf{Q}}$ for all $\vec{\mathbf{v}} \in \mathbf{V}^{\perp}$. □

Next, having in mind that $\mathbf{V} = \mathbf{V}_0$ (cf. (3.22)), we prove the $\mathbf{V}$-ellipticity of the bilinear form $\mathbf{c}$ (cf. (3.15)).

Lemma 3.6. *There exists a positive constant $\gamma_{\mathbf{V}}$ such that*

$$\mathbf{c}(\vec{\mathbf{v}}, \vec{\mathbf{v}}) \geq \gamma_{\mathbf{V}} \|\vec{\mathbf{v}}\|_{\mathbf{Q}}^2 \quad \forall\, \vec{\mathbf{v}} \in \mathbf{V}.$$

Proof. Given $\vec{\mathbf{v}} := (\mathbf{q}, \nabla\mathbf{q}, \mathbf{q}) \in \mathbf{V}$ (cf. (3.22)), it follows from (3.15) and the fact that all the norms in Q are equivalent, that there exists a positive constant c_E, depending only on Q, such that

$$\mathbf{c}(\vec{\mathbf{v}}, \vec{\mathbf{v}}) = \beta\|\mathbf{q}\|_{0,\Omega}^2 \geq \frac{\beta}{2}\{\|\mathbf{q}\|_{0,\Omega}^2 + c_E\|\mathbf{q}\|_{1,\Omega}^2\} \geq \gamma_{\mathbf{V}}\|\vec{\mathbf{v}}\|_{\mathbf{Q}}^2 \quad \forall\, \vec{\mathbf{v}} \in \mathbf{V},$$

with $\gamma_{\mathbf{V}} = \frac{\beta}{2}\min\{1, c_E\}$. □

Hence, thanks to (3.17), and Lemmas 3.3, 3.5, and 3.6, we are able to prove the following result.

Lemma 3.7. *For each pair* $(\mathbf{F}, \mathbf{G}) \in \mathbf{H}' \times \mathbf{Q}'$ *there exist unique* $\underline{\vec{\boldsymbol{\sigma}}} := (\underline{\boldsymbol{\sigma}}, \underline{\boldsymbol{\rho}}) \in \mathbf{H}$ *and* $\underline{\vec{\mathbf{u}}} := (\underline{\mathbf{u}}, \underline{\boldsymbol{\Phi}}, \underline{\boldsymbol{\lambda}}) \in \mathbf{Q}$ *such that*

$$\begin{aligned} &\mathbf{a}(\underline{\vec{\boldsymbol{\sigma}}}, \vec{\boldsymbol{\tau}}) + \mathbf{b}(\vec{\boldsymbol{\tau}}, \underline{\vec{\mathbf{u}}}) = \mathbf{F}(\vec{\boldsymbol{\tau}}) \quad \forall\, \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}, \\ &\mathbf{b}(\underline{\vec{\boldsymbol{\sigma}}}, \vec{\mathbf{v}}) - \mathbf{c}(\underline{\vec{\mathbf{u}}}, \vec{\mathbf{v}}) = \mathbf{G}(\vec{\mathbf{v}}) \quad \forall\, \vec{\mathbf{v}} := (\mathbf{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}. \end{aligned} \tag{3.33}$$

Moreover, there exists a positive constant $\underline{C}$*, depending only on* $\alpha_{\mathbf{K}}$, $\beta_{\mathbf{B}}$, $\gamma_{\mathbf{V}}$, *and the norms of the operators induced by* $\mathbf{a}$ *and* $\mathbf{b}$*, such that*

$$\|(\underline{\vec{\boldsymbol{\sigma}}}, \underline{\vec{\mathbf{u}}})\|_{\mathbf{H}\times\mathbf{Q}} \leq \underline{C}\{\|\mathbf{F}\|_{\mathbf{H}'} + \|\mathbf{G}\|_{\mathbf{Q}'}\}. \tag{3.34}$$

Proof. It follows from a straightforward application of Theorem 3.1. □

Next, given an arbitrary $\mathbf{z} \in \mathbf{L}^2(\Omega)$, we consider the particular pair $(\mathbf{F}, \mathbf{G}) := (\mathbf{0}, \alpha\mathbf{F}_{\mathbf{z}}) \in \mathbf{H}' \times \mathbf{Q}'$, and conclude, thanks to Lemma 3.7, that the problem defining $\mathbf{T}(\mathbf{z})$ (cf. (3.18)) is well-posed, thus confirming that the operator $\mathbf{T} : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$ is well-defined. Moreover, by noticing from (3.16) that $\|\alpha\mathbf{F}_{\mathbf{z}}\|_{\mathbf{Q}'} = \alpha\|\nabla\mathcal{D}(\mathbf{z})\|_{0,\Omega}$, we deduce from (3.34) that there holds

$$\|\mathbf{T}(\mathbf{z})\|_{0,\Omega} \leq \|(\underline{\vec{\boldsymbol{\sigma}}}, \underline{\vec{\mathbf{u}}})\|_{\mathbf{H}\times\mathbf{Q}} \leq \underline{C}\alpha\|\nabla\mathcal{D}(\mathbf{z})\|_{0,\Omega} \quad \forall\, \mathbf{z} \in \mathbf{L}^2(\Omega). \tag{3.35}$$

The Lipschitz continuity of the operator $\mathbf{T}$ is established in the following lemma.

Lemma 3.8. *Assume* (A2) *and let* $\underline{C}$ *be the constant provided by the continuous dependence estimate* (3.34)*. Then, there holds*

$$\|\mathbf{T}(\mathbf{z}_1) - \mathbf{T}(\mathbf{z}_2)\|_{0,\Omega} \leq \alpha\underline{C}L_{\mathcal{D}}\|\mathbf{z}_1 - \mathbf{z}_2\|_{0,\Omega} \quad \forall\, \mathbf{z}_1, \mathbf{z}_2 \in \mathbf{L}^2(\Omega).$$

Proof. We proceed analogously to Lemma 11 in Ref. 7. In this way, given $\mathbf{z}_j \in \mathbf{L}^2(\Omega)$, $j \in \{1, 2\}$, we let $\underline{\vec{\boldsymbol{\sigma}}}_j := (\underline{\boldsymbol{\sigma}}_j, \underline{\boldsymbol{\rho}}_j) \in \mathbf{H}$ and $\underline{\vec{\mathbf{u}}}_j := (\underline{\mathbf{u}}_j, \underline{\boldsymbol{\Phi}}_j, \underline{\boldsymbol{\lambda}}_j) \in \mathbf{Q}$ be the unique solution to (3.18) with $\mathbf{z} = \mathbf{z}_j$, so that $\mathbf{T}(\mathbf{z}_j) = \underline{\mathbf{u}}_j$. Subtracting the respective rows of the resulting systems (3.18), we easily find that $(\underline{\vec{\boldsymbol{\sigma}}}_1 - \underline{\vec{\boldsymbol{\sigma}}}_2, \underline{\vec{\mathbf{u}}}_1 - \underline{\vec{\mathbf{u}}}_2) \in \mathbf{H}\times\mathbf{Q}$ is solution of (3.33) with $\mathbf{F} := 0$ and $\mathbf{G} := \alpha(\mathbf{F}_{\mathbf{z}_1} - \mathbf{F}_{\mathbf{z}_2})$, and hence the corresponding

estimate (3.34) and the Lipschitz continuity of $\nabla\mathcal{D}$ (cf. (A2)) yield

$$\begin{aligned}\|\mathbf{T}(\mathbf{z}_1)-\mathbf{T}(\mathbf{z}_2)\|_{0,\Omega} &\leq \|\vec{\mathbf{u}}_1-\vec{\mathbf{u}}_2\|_{\mathbf{Q}'} \leq \underline{C}\|\alpha(\mathbf{F}_{\mathbf{z}_1}-\mathbf{F}_{\mathbf{z}_2})\|_{\mathbf{Q}'}\\ &= \underline{C}\alpha\|\nabla\mathcal{D}(\mathbf{z}_1)-\nabla\mathcal{D}(\mathbf{z}_2)\|_{0,\Omega} \leq \underline{C}\alpha L_{\mathcal{D}}\|\mathbf{z}_1-\mathbf{z}_2\|_{0,\Omega},\end{aligned}$$

which finishes the proof. □

We are now in position to establish the existence of a unique fixed-point for the operator $\mathbf{T}$, or equivalently, the well-posedness of problem (3.12). More precisely, we have the following result.

Theorem 3.2. *Assume* (A2), (A3) *and* $\alpha\underline{C}L_{\mathcal{D}} < 1$. *Then, the mixed problem* (3.12) *has a unique solution* $(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}}) \in \mathbf{H}\times\mathbf{Q}$. *Moreover, the following a priori estimate holds*:

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})\|_{\mathbf{H}\times\mathbf{Q}} \leq \underline{C}\alpha M_D.$$

Proof. It follows straightforwardly from Lemma 3.8 and the present hypothesis involving the constants α, $\underline{C}$, and $L_{\mathcal{D}}$ that $\mathbf{T}$ is a contraction, and hence the classical Banach theorem implies the existence of a unique fixed point of $\mathbf{T}$. Equivalently, the mixed problem (3.12) has a unique solution $(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}}) \in \mathbf{H}\times\mathbf{Q}$, which, according to the estimate (3.35) and the assumption (A3), satisfies

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})\|_{\mathbf{H}\times\mathbf{Q}} \leq \underline{C}\alpha\|\nabla\mathcal{D}(\mathbf{u})\|_{0,\Omega} \leq \underline{C}\alpha M_{\mathcal{D}},$$

thus completing the proof. □

3.3. *Analysis of the discrete scheme*

In this section, we introduce and analyze a Galerkin scheme for problem (3.12). As in Sec. 2.4, we first let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular triangulations of $\bar{\Omega}$ made of triangles K with diameter h_K, and define the meshsize $h := \max\{h_K : K \in \mathcal{T}_h\}$, which also serves as the index of $\mathcal{T}_h$. In turn, we recall that, given a non-negative integer k, $\mathrm{P}_k(K)$ stands for the space of polynomials of degree $\leq k$ on K, whose vector and tensor versions are denoted by $\mathbf{P}_k(K)$ and $\mathbb{P}_k(K)$, respectively. Then, noting that certainly the space of rigid motions Q is already of finite dimension, we propose next two possible sets of finite element subspaces of $\mathbb{H}_0(\mathbf{div};\Omega)$, $\mathbf{L}^2(\Omega)$, and $\mathbb{L}^2_{\mathtt{skew}}(\Omega)$, which, in order to make clear the unknowns they are approximating, are denoted by $\mathrm{H}_h^{\boldsymbol{\sigma}}$, $\mathrm{H}_h^{\mathbf{u}}$ and $\mathrm{H}_h^{\boldsymbol{\Phi}}$, respectively. The first choice, employed in Sec. 4.2 of Ref. 7 and Sec. 3.4 of Ref. 8 for previous related results, consists of the Brezzi–Douglas–Marini (BDM) space of order 1 for the stress (cf. Ref. 11) and the rest as in Theorem 7.2 of Ref. 4, that is

$$\begin{aligned}\mathrm{H}_h^{\boldsymbol{\sigma}} &:= \{\boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \boldsymbol{\tau}_h|_K \in \mathbb{P}_1(K)\ \forall\, K \in \mathcal{T}_h\},\\ \mathrm{H}_h^{\mathbf{u}} &:= \{\mathbf{v}_h \in \mathbf{L}^2(\Omega) : \mathbf{v}_h|_K \in \mathbf{P}_0(K)\ \forall\, K \in \mathcal{T}_h\},\\ \mathrm{H}_h^{\boldsymbol{\Phi}} &:= \left\{\boldsymbol{\Psi}_h := \begin{pmatrix} 0 & \psi_h\\ -\psi_h & 0\end{pmatrix} \in \mathbb{L}^2_{\mathtt{skew}}(\Omega) : \psi_h|_K \in \mathrm{P}_0(K)\ \forall\, K \in \mathcal{T}_h\right\}.\end{aligned} \tag{3.36}$$

In addition, we also consider the classical PEERS space of order 0, originally introduced in Ref. 3 for linear elasticity as well, which is given by

$$\begin{aligned}
\mathrm{H}_h^{\boldsymbol{\sigma}} &:= \{\boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \boldsymbol{\tau}_{h,i}|_K \in \mathbf{RT}_0(K) \oplus \mathrm{P}_0(K)\mathrm{curl}^{\mathtt{t}} b_K \\
&\qquad \forall\, i \in \{1,2\},\ \forall\, K \in \mathcal{T}_h\}, \\
\mathrm{H}_h^{\mathbf{u}} &:= \{\mathbf{v}_h \in \mathbf{L}^2(\Omega) : \mathbf{v}_h|_K \in \mathbf{P}_0(K)\ \forall\, K \in \mathcal{T}_h\}, \\
\mathrm{H}_h^{\boldsymbol{\Phi}} &:= \left\{ \boldsymbol{\Psi}_h := \begin{pmatrix} 0 & \psi_h \\ -\psi_h & 0 \end{pmatrix} \in \mathbb{C}(\bar{\Omega}) : \psi_h|_K \in \mathrm{P}_1(K)\ \forall\, K \in \mathcal{T}_h \right\},
\end{aligned} \tag{3.37}$$

where $\boldsymbol{\tau}_{h,i}$ denotes the ith row of $\boldsymbol{\tau}_h$, $\mathbf{RT}_0(K)$ is the local Raviart–Thomas space of order 0 (cf. Refs. 12 and 19), b_K is the usual cubic bubble function on K, and $\mathrm{curl}^{\mathtt{t}} b_K = (\frac{\partial b_K}{\partial x_2}, -\frac{\partial b_K}{\partial x_1})$. Nevertheless, for stability purposes to be discussed later on (see Lemma 3.9), we need that the space of rigid motions Q be contained in the finite element subspace approximating $\mathbf{u}$, reason why we now enrich this space with the $\mathbf{P}_1(\Omega)$-component of Q, thus yielding the introduction of

$$\widetilde{\mathrm{H}}_h^{\mathbf{u}} := \mathrm{H}_h^{\mathbf{u}} \oplus \left\langle \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix} \right\rangle. \tag{3.38}$$

Then, letting $\mathbf{H}_h := \mathrm{H}_h^{\boldsymbol{\sigma}} \times Q$ and $\mathbf{Q}_h := \widetilde{\mathrm{H}}_h^{\mathbf{u}} \times \mathrm{H}_h^{\boldsymbol{\Phi}} \times Q$, the Galerkin scheme of (3.12) reads: Find $\vec{\boldsymbol{\sigma}}_h := (\boldsymbol{\sigma}_h, \boldsymbol{\rho}_h) \in \mathbf{H}_h$ and $\vec{\mathbf{u}}_h := (\mathbf{u}_h, \boldsymbol{\Phi}_h, \boldsymbol{\lambda}_h) \in \mathbf{Q}_h$ such that

$$\begin{aligned}
&\mathbf{a}(\vec{\boldsymbol{\sigma}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{u}}_h) = 0 && \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h, \\
&\mathbf{b}(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{v}}_h) - \mathbf{c}(\vec{\mathbf{u}}_h, \vec{\mathbf{v}}_h) = \alpha \mathbf{F}_{\mathbf{u}_h}(\vec{\mathbf{v}}_h) && \forall\, \vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h.
\end{aligned} \tag{3.39}$$

Analogously to the analysis from Sec. 3.2, we now introduce the discrete operator $\mathbf{T}_h : \widetilde{\mathrm{H}}_h^{\mathbf{u}} \to \widetilde{\mathrm{H}}_h^{\mathbf{u}}$ defined by $\mathbf{T}_h(\mathbf{z}_h) := \underline{\mathbf{u}}_h$ for each $\mathbf{z}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}$, where $\underline{\vec{\boldsymbol{\sigma}}}_h := (\underline{\boldsymbol{\sigma}}_h, \underline{\boldsymbol{\rho}}_h) \in \mathbf{H}_h$ and $\underline{\vec{\mathbf{u}}}_h := (\underline{\mathbf{u}}_h, \underline{\boldsymbol{\Phi}}_h, \underline{\boldsymbol{\lambda}}_h) \in \mathbf{Q}_h$ satisfy

$$\begin{aligned}
&\mathbf{a}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \underline{\vec{\mathbf{u}}}_h) = 0 && \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h, \\
&\mathbf{b}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\mathbf{v}}_h) - \mathbf{c}(\underline{\vec{\mathbf{u}}_h}, \vec{\mathbf{v}}_h) = \alpha \mathbf{F}_{\mathbf{z}_h}(\vec{\mathbf{v}}_h) && \forall\, \vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h.
\end{aligned} \tag{3.40}$$

As for the continuous problem, it is easy to see that solving (3.39) is equivalent to looking for a fixed point of $\mathbf{T}_h$, that is: Find $\mathbf{u}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}$ such that $\mathbf{T}_h(\mathbf{u}_h) = \mathbf{u}_h$, for whose solvability analysis we need to show first that $\mathbf{T}_h$ is well-defined, equivalently that (3.40) is well-posed. For this purpose, in what follows, we apply Theorem 3.1 to the discrete setting provided by the spaces $\mathbf{H}_h$ and $\mathbf{Q}_h$, the bilinear forms $\mathbf{a}|_{\mathbf{H}_h \times \mathbf{H}_h}$ and $\mathbf{b}|_{\mathbf{H}_h \times \mathbf{Q}_h}$, and the discrete kernels of $\mathbf{B}$ and $\mathbf{B}^{\mathtt{t}}$, which are given, respectively, by

$$\mathbf{K}_h := \{\vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h : \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{v}}_h) = 0\ \forall\, \vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h\}, \tag{3.41}$$

and

$$\mathbf{V}_h := \{\vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h : \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{v}}_h) = 0\ \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h\}. \tag{3.42}$$

Thus, employing the expression for $\mathbf{b}$ given by (3.20), we can redefine $\mathbf{K}_h$ as

$$\begin{aligned}\mathbf{K}_h := \Big\{ \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h : \int_\Omega \mathbf{v}_h \cdot \{\mathbf{div}\boldsymbol{\tau}_h - \boldsymbol{\eta}_h\} = 0 \ \forall\, \mathbf{v}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}, \\ \int_\Omega \boldsymbol{\Psi}_h : \boldsymbol{\tau}_h = 0 \ \forall\, \boldsymbol{\Psi}_h \in \mathrm{H}_h^{\boldsymbol{\Phi}}, \int_\Omega \boldsymbol{\xi}_h \cdot \boldsymbol{\eta}_h = 0 \ \forall\, \boldsymbol{\xi}_h \in Q \Big\},\end{aligned} \tag{3.43}$$

from which, noticing that the pair $(\mathrm{H}_h^{\boldsymbol{\sigma}}, \widetilde{\mathrm{H}}_h^{\mathbf{u}})$, taken either from (3.36)–(3.38) or (3.37)–(3.38), satisfies the inclusion $\mathbf{div}\mathrm{H}_h^{\boldsymbol{\sigma}} \subseteq \widetilde{\mathrm{H}}_h^{\mathbf{u}}$, it readily follows that

$$\mathbf{K}_h := \left\{ \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h : \mathbf{div}\boldsymbol{\tau}_h = 0, \boldsymbol{\eta}_h = 0, \int_\Omega \boldsymbol{\Psi}_h : \boldsymbol{\tau}_h = 0 \ \forall\, \boldsymbol{\Psi}_h \in \mathrm{H}_h^{\boldsymbol{\Phi}} \right\}.$$

In this way, due to the first two identities characterizing $\mathbf{K}_h$ in the foregoing equation, we deduce that the $\mathbf{K}_h$-ellipticity of $\mathbf{a}$ can be proved exactly as we did for its $\mathbf{K}$-ellipticity, and hence with the same constant $\alpha_{\mathbf{K}} := \frac{c_1 c_2}{2\mu_s}$ from Lemma 3.3 there holds

$$\mathbf{a}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{\tau}}_h) \geq \alpha_{\mathbf{K}} \|\vec{\boldsymbol{\tau}}_h\|_{\mathbf{H}}^2 \quad \forall\, \vec{\boldsymbol{\tau}}_h \in \mathbf{K}_h. \tag{3.44}$$

We now aim to establish the discrete analogue of Lemma 3.4, for which we first highlight that, thanks to the enriched space $\widetilde{\mathrm{H}}_h^{\mathbf{u}}$ (cf. (3.38)), one guarantees that $\mathbf{V}_0$ (cf. (3.22)) is a subspace of $\mathbf{Q}_h$. Then, we have the following result.

Lemma 3.9. *There exists a positive constant $\widetilde{\beta}_{\mathbf{B}}$, independent of h, such that*

$$S_h(\vec{\mathbf{v}}_h) := \sup_{\substack{\vec{\boldsymbol{\tau}}_h \in \mathbf{H}_h \\ \vec{\boldsymbol{\tau}}_h \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \mathbf{v}_h)}{\|\vec{\boldsymbol{\tau}}_h\|_{\mathbf{H}}} \geq \widetilde{\beta}_{\mathbf{B}} \mathrm{dist}(\vec{\mathbf{v}}_h, \mathbf{V}_0) \quad \forall\, \vec{\mathbf{v}}_h \in \mathbf{Q}_h. \tag{3.45}$$

Proof. We proceed analogously to the proof of Lemma 3.4. However, because of the similarities involved, we simplify our reasoning by using the results already available along the proof of Lemma 4.1 in Ref. 20, which in turn is an adaptation of the proof of Theorem 4.5 in Ref. 27. We begin by recalling from (3.26) that, given $\vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h$ and $\vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h$, we can rewrite $\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{v}}_h)$ as

$$\begin{aligned}\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{v}}_h) := \int_\Omega (\mathbf{v}_h - \mathit{\Pi}_Q \mathbf{v}_h) \cdot \mathbf{div}\boldsymbol{\tau}_h + \int_\Omega (\boldsymbol{\Psi}_h - \nabla \mathit{\Pi}_Q \mathbf{v}_h) : \boldsymbol{\tau}_h \\ + \int_\Omega (\boldsymbol{\xi}_h - \mathit{\Pi}_Q \mathbf{v}_h) \cdot \boldsymbol{\eta}_h,\end{aligned} \tag{3.46}$$

from which one easily deduces that $\mathbf{V}_0 \subseteq \mathbf{V}_h$, and hence (3.45) is trivially satisfied for $\vec{\mathbf{v}}_h \in \mathbf{V}_0$. According to this, it only remains to prove for $\vec{\mathbf{v}}_h \in \mathbf{Q}_h \backslash \mathbf{V}_0$. Indeed, if $\mathbf{v}_h - \mathit{\Pi}_Q \mathbf{v}_h \neq \mathbf{0}$, we know from the first part of the proof of Lemma 4.1 in Ref. 20 that there exists $\boldsymbol{\zeta}_h \in \mathrm{H}_h^{\boldsymbol{\sigma}}$ such that $\mathbf{div}(\boldsymbol{\zeta}_h) = \mathcal{P}_h(\mathbf{v}_h - \mathit{\Pi}_Q \mathbf{v}_h)$ and $\|\boldsymbol{\zeta}_h\|_{\mathbf{div};\Omega} \leq \widetilde{C}_N \|\mathbf{v}_h - \mathit{\Pi}_Q \mathbf{v}_h\|_{0,\Omega}$, where $\mathcal{P}_h : \mathbf{L}^2(\Omega) \to \mathrm{H}_h^{\mathbf{u}}$ is the orthogonal projection, and $\widetilde{C}_N$ is a positive constant independent of h. In turn, decomposing $\mathbf{v}_h = \bar{\mathbf{v}}_h + \mathbf{q}_h$, with $\bar{\mathbf{v}}_h \in \mathrm{H}_h^{\mathbf{u}}$ and $\mathbf{q}_h \in \langle \binom{x_2}{-x_1} \rangle$, we obtain $\mathit{\Pi}_Q \mathbf{v}_h = \mathit{\Pi}_Q \bar{\mathbf{v}}_h + \mathbf{q}_h$, and thus

$\mathbf{v}_h - \boldsymbol{\Pi}_Q\mathbf{v}_h = \bar{\mathbf{v}}_h - \boldsymbol{\Pi}_Q\bar{\mathbf{v}}_h$. In particular, this latter identity obviously implies $\mathbf{div}(\boldsymbol{\zeta}_h) = \mathcal{P}_h(\bar{\mathbf{v}}_h - \boldsymbol{\Pi}_Q\bar{\mathbf{v}}_h)$. Then, setting $\vec{\boldsymbol{\zeta}}_h := (\boldsymbol{\zeta}_h, \mathbf{0})$, using the original definition of $\mathbf{b}$ (cf. (3.14)), integrating by parts similarly as done for the derivation of (3.26), and applying the properties of the orthogonal projections $\mathcal{P}_h$ and $\boldsymbol{\Pi}_Q$, we find that

$$\begin{aligned}
\mathbf{b}(\vec{\boldsymbol{\zeta}}_h, \vec{\mathbf{v}}_h) &= \int_\Omega (\bar{\mathbf{v}}_h + \mathbf{q}_h) \cdot \mathbf{div}(\boldsymbol{\zeta}_h) + \int_\Omega \boldsymbol{\Psi}_h : \boldsymbol{\zeta}_h \\
&= \int_\Omega \bar{\mathbf{v}}_h \cdot \mathcal{P}_h(\bar{\mathbf{v}}_h - \boldsymbol{\Pi}_Q\bar{\mathbf{v}}_h) + \int_\Omega (\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h) : \boldsymbol{\zeta}_h \\
&= \int_\Omega \bar{\mathbf{v}}_h \cdot (\bar{\mathbf{v}}_h - \boldsymbol{\Pi}_Q\bar{\mathbf{v}}_h) + \int_\Omega (\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h) : \boldsymbol{\zeta}_h \\
&= \|\bar{\mathbf{v}}_h - \boldsymbol{\Pi}_Q\bar{\mathbf{v}}_h\|^2_{0,\Omega} + \int_\Omega (\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h) : \boldsymbol{\zeta}_h,
\end{aligned}$$

which readily yields

$$\begin{aligned}
S_h(\vec{\mathbf{v}}_h) \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}}_h, \vec{\mathbf{v}}_h)}{\|\vec{\boldsymbol{\zeta}}_h\|_{\mathbf{H}}} &= \frac{\|\mathbf{v}_h - \boldsymbol{\Pi}_Q\mathbf{v}_h\|^2_{0,\Omega} + \displaystyle\int_\Omega (\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h) : \boldsymbol{\zeta}_h}{\|\boldsymbol{\zeta}_h\|_{\mathbf{div};\Omega}} \\
&\geq \frac{1}{\widetilde{C}_N}\|\mathbf{v}_h - \boldsymbol{\Pi}_Q\mathbf{v}_h\|_{0,\Omega} - \|\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h\|_{0,\Omega}. \qquad (3.47)
\end{aligned}$$

Next, assuming that $\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h \neq \mathbf{0}$ and appealing now to the second half of the proof of Lemma 4.1 in Ref. 20, there exists another $\boldsymbol{\zeta}_h \in \mathrm{H}^{\boldsymbol{\sigma}}_h$ such that $\mathbf{div}(\boldsymbol{\zeta}_h) = \mathbf{0}$, $\displaystyle\int_\Omega (\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h) : \boldsymbol{\zeta}_h = \|\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h\|^2_{0,\Omega}$, and $\|\boldsymbol{\zeta}_h\|_{\mathbf{div};\Omega} \leq \widetilde{c}_N\|\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h\|_{0,\Omega}$, where $\widetilde{c}_N$ is a positive constant independent of h. Hence, defining $\vec{\boldsymbol{\zeta}}_h := (\boldsymbol{\zeta}_h, \mathbf{0})$, and employing again (3.14), we obtain

$$\mathbf{b}(\vec{\boldsymbol{\zeta}}_h, \vec{\mathbf{v}}_h) = \|\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h\|^2_{0,\Omega},$$

which, similarly as before, gives

$$S_h(\vec{\mathbf{v}}_h) \geq \frac{1}{\widetilde{c}_N}\|\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h\|_{0,\Omega}. \qquad (3.48)$$

In this way, a suitable linear combination of (3.47) and (3.48) implies the existence of a positive constant $\widetilde{\beta}_1$, depending only on $\widetilde{C}_N$ and $\widetilde{c}_N$, such that

$$S_h(\vec{\mathbf{v}}_h) \geq \widetilde{\beta}_1\{\|\mathbf{v}_h - \boldsymbol{\Pi}_Q\mathbf{v}_h\|_{0,\Omega} + \|\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h\|_{0,\Omega}\}. \qquad (3.49)$$

In addition, proceeding exactly as for the derivation of (3.48), but now considering $\boldsymbol{\Psi}_h - \nabla\boldsymbol{\Pi}_Q\mathbf{v}_h$ in place of $\boldsymbol{\Psi}_h - \nabla\mathbf{q}_h$, and utilizing the expression (3.46) for $\mathbf{b}$, we are able to show that

$$S_h(\vec{\mathbf{v}}_h) \geq \frac{1}{\widehat{c}_N}\|\boldsymbol{\Psi}_h - \nabla\boldsymbol{\Pi}_Q\mathbf{v}_h\|_{0,\Omega}, \qquad (3.50)$$

with a positive constant $\widehat{c}_N$ independent of h. Furthermore, if $\boldsymbol{\xi}_h - \mathit{\Pi}_Q \mathbf{v}_h \neq \mathbf{0}$, we do as in the continuous case (cf. (3.30) in the proof of Lemma 3.4) and choose $\vec{\boldsymbol{\zeta}}_h := (\mathbf{0}, \boldsymbol{\xi}_h - \mathit{\Pi}_Q \mathbf{v}_h)$ to prove, according to (3.46), that

$$S_h(\vec{\mathbf{v}}_h) \geq \|\boldsymbol{\xi}_h - \mathit{\Pi}_Q \mathbf{v}_h\|_{0,\Omega}. \tag{3.51}$$

The rest of the proof follows analogously to the one of Lemma 3.4 by considering now the inequalities (3.49)–(3.51), and after discarding the expression $\|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|_{0,\Omega}$ in the first one of them. We omit further details. □

As a first straightforward consequence of (3.45) we have that $\mathbf{V}_h \subseteq \mathbf{V}_0$, and hence $\mathbf{V}_h = \mathbf{V}_0$. Moreover, since $\mathrm{dist}(\vec{\mathbf{v}}_h, \mathbf{V}_h) = \|\vec{\mathbf{v}}_h\|_{\mathbf{Q}}$ for all $\vec{\mathbf{v}}_h \in \mathbf{V}_h^{\perp}$, we conclude the discrete inf-sup condition for $\mathbf{b}$, that is

$$\sup_{\substack{\vec{\boldsymbol{\tau}}_h \in \mathbf{H}_h \\ \vec{\boldsymbol{\tau}}_h \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{v}}_h)}{\|\vec{\boldsymbol{\tau}}_h\|_{\mathbf{H}}} \geq \widetilde{\beta}_{\mathbf{B}} \|\vec{\mathbf{v}}_h\|_{\mathbf{Q}} \quad \forall\, \vec{\mathbf{v}}_h \in \mathbf{V}_h^{\perp} \cap \mathbf{Q}_h, \tag{3.52}$$

with certainly the same constant $\widetilde{\beta}_{\mathbf{B}}$ from Lemma 3.9. On the other hand, since the continuous and discrete kernels $\mathbf{V}$ and $\mathbf{V}_h$, respectively, coincide, the $\mathbf{V}_h$-ellipticity of the bilinear form $\mathbf{c}$ is already proved by Lemma 3.6.

Therefore, bearing in mind (3.44), (3.52), and Lemma 3.6, a straightforward application of Theorem 3.1 allows us to establish the following result.

Lemma 3.10. *For each pair* $(\mathbf{F}, \mathbf{G}) \in \mathbf{H}' \times \mathbf{Q}'$ *there exist unique* $\underline{\vec{\boldsymbol{\sigma}}}_h := (\underline{\boldsymbol{\sigma}}_h, \underline{\boldsymbol{\rho}}_h) \in \mathbf{H}_h$ *and* $\underline{\vec{\mathbf{u}}}_h := (\underline{\mathbf{u}}_h, \underline{\boldsymbol{\Phi}}_h, \underline{\boldsymbol{\lambda}}_h) \in \mathbf{Q}_h$ *such that*

$$\begin{aligned} &\mathbf{a}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \underline{\vec{\mathbf{u}}}_h) = \mathbf{F}(\vec{\boldsymbol{\tau}}_h) \qquad \forall \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h, \\ &\mathbf{b}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\mathbf{v}}_h) - \mathbf{c}(\underline{\vec{\mathbf{u}}}_h, \vec{\mathbf{v}}_h) = \mathbf{G}(\vec{\mathbf{v}}_h) \qquad \forall \vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h. \end{aligned} \tag{3.53}$$

Moreover, there exists a positive constant $\underline{\widetilde{C}}$, *depending only on* $\alpha_{\mathbf{K}}$, $\widetilde{\beta}_{\mathbf{B}}$, $\gamma_{\mathbf{V}}$, *and the norms of the operators induced by* $\mathbf{a}$ *and* $\mathbf{b}$, *such that*

$$\|(\underline{\vec{\boldsymbol{\sigma}}}_h, \underline{\vec{\mathbf{u}}}_h)\|_{\mathbf{H} \times \mathbf{Q}} \leq \underline{\widetilde{C}} \Big\{ \|\mathbf{F}\|_{\mathbf{H}'} + \|\mathbf{G}\|_{\mathbf{Q}'} \Big\}. \tag{3.54}$$

Next, we proceed analogously to the continuous case (cf. (3.35) and the last part of Sec. 3.2) by applying now Lemma 3.10 to the pair of functionals $(\mathbf{F}, \mathbf{G}) := (\mathbf{0}, \alpha \mathbf{F}_{\mathbf{z}_h})$, with an arbitrary $\mathbf{z}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}$. In this way, we conclude that $\mathbf{T}_h : \widetilde{\mathrm{H}}_h^{\mathbf{u}} \to \widetilde{\mathrm{H}}_h^{\mathbf{u}}$ is well-posed, and that

$$\|\mathbf{T}_h(\mathbf{z}_h)\|_{0,\Omega} \leq \|(\underline{\vec{\boldsymbol{\sigma}}}_h, \underline{\vec{\mathbf{u}}}_h)\|_{\mathbf{H} \times \mathbf{Q}} \leq \underline{\widetilde{C}} \alpha \|\nabla \mathcal{D}(\mathbf{z}_h)\|_{0,\Omega} \quad \forall\, \mathbf{z}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}. \tag{3.55}$$

Moreover, adopting the same arguments from Lemma 3.8, and employing the *a priori* estimate (3.54) and the Lipschitz continuity of $\nabla \mathcal{D}$ (cf. (A2)), we arrive at the same property for the operator $\mathbf{T}_h$, that is

$$\|\mathbf{T}_h(\mathbf{z}_h) - \mathbf{T}_h(\mathbf{w}_h)\|_{0,\Omega} \leq \underline{\widetilde{C}} \alpha L_{\mathcal{D}} \|\mathbf{z}_h - \mathbf{w}_h\|_{0,\Omega} \quad \forall\, \mathbf{z}_h, \mathbf{w}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}.$$

Consequently, we are now in position to establish the well-posedness of our MFEM (3.39), by appealing to its equivalence with the existence of a unique fixed point

of $\mathbf{T}_h$, and applying again the respective Banach theorem. We omit further details and state the corresponding result as follows.

Theorem 3.3. *Assume* (A2), (A3) *and* $\alpha\widetilde{\underline{C}}L_{\mathcal{D}} < 1$. *Then, the discrete scheme* (3.39) *has a unique solution* $(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h) \in \mathbf{H}_h \times \mathbf{Q}_h$. *Moreover, the following a priori estimate holds*:

$$\|(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widetilde{\underline{C}}\alpha M_D.$$

3.4. *A priori error analysis*

Given $(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) \in \mathbf{H} \times \mathbf{Q}$ and $(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h) \in \mathbf{H}_h \times \mathbf{Q}_h$, the unique solutions of the continuous and discrete problems (3.12) and (3.39), respectively, we now aim to estimate the corresponding error given by $\|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) - (\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}}$. To this end, we first let $(\underline{\vec{\boldsymbol{\sigma}}}_h, \underline{\vec{\mathbf{u}}}_h) \in \mathbf{H}_h \times \mathbf{Q}_h$ be the solution to (3.53) with $\mathbf{F} = \mathbf{0}$ and $\mathbf{G} = \alpha\mathbf{F}_{\mathbf{u}}$, equivalently the solution to (3.40) with $\mathbf{u}$ in place of $\mathbf{z}_h$, that is

$$\begin{aligned} &\mathbf{a}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \underline{\vec{\mathbf{u}}}_h) = 0 \qquad && \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h, \\ &\mathbf{b}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\mathbf{v}}_h) - \mathbf{c}(\underline{\vec{\mathbf{u}}}_h, \vec{\mathbf{v}}_h) = \alpha\mathbf{F}_{\mathbf{u}}(\vec{\mathbf{v}}_h) \qquad && \forall\, \vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h, \end{aligned} \tag{3.56}$$

which certainly can be seen as the classical Galerkin approximation of (3.12). Then, invoking the corresponding Céa estimate (see, e.g. Proposition 5.5.2. in Ref. 9), we have the preliminary estimate

$$\|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) - (\underline{\vec{\boldsymbol{\sigma}}}_h, \underline{\vec{\mathbf{u}}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widehat{C}\{\mathrm{dist}(\vec{\boldsymbol{\sigma}}, \mathbf{H}_h) + \mathrm{dist}(\vec{\mathbf{u}}, \mathbf{Q}_h)\}, \tag{3.57}$$

where $\widehat{C}$ is a positive constant independent of h. Next, subtracting (3.56) from (3.39), we find that $(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h) - (\underline{\vec{\boldsymbol{\sigma}}}_h, \underline{\vec{\mathbf{u}}}_h)$ solves

$$\mathbf{a}(\vec{\boldsymbol{\sigma}}_h - \underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{u}}_h - \underline{\vec{\mathbf{u}}}_h)) = 0 \quad \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h,$$

$$\mathbf{b}(\vec{\boldsymbol{\sigma}}_h - \underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\mathbf{v}}_h) - \mathbf{c}(\vec{\mathbf{u}}_h - \underline{\vec{\mathbf{u}}}_h, \vec{\mathbf{v}}_h) = \alpha(\mathbf{F}_{\mathbf{u}_h} - \mathbf{F}_{\mathbf{u}})(\vec{\mathbf{v}}_h) \quad \forall\, \vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h,$$

and hence, thanks to the *a priori* estimate (3.54), the fact that $\|\mathbf{F}_{\mathbf{u}_h} - \mathbf{F}_{\mathbf{u}}\|_{\mathbf{Q}'} = \|\nabla\mathcal{D}(\mathbf{u}_h) - \nabla\mathcal{D}(\mathbf{u})\|_{0,\Omega}$ (cf. (3.16)), and the Lipschitz continuity of $\nabla\mathcal{D}$ (cf. (A2)), there holds

$$\|(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h) - (\underline{\vec{\boldsymbol{\sigma}}}_h, \underline{\vec{\mathbf{u}}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widetilde{\underline{C}}\alpha L_{\mathcal{D}}\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}. \tag{3.58}$$

In this way, employing the triangle inequality together with the estimates (3.57) and (3.58), and then realizing that $\mathrm{dist}(\vec{\boldsymbol{\sigma}}, \mathbf{H}_h) = \mathrm{dist}(\boldsymbol{\sigma}, \mathrm{H}_h^{\boldsymbol{\sigma}})$ and that $\mathrm{dist}(\vec{\mathbf{u}}, \mathbf{Q}_h) = \mathrm{dist}(\mathbf{u}, \widetilde{\mathrm{H}}_h^{\mathbf{u}}) + \mathrm{dist}(\boldsymbol{\Phi}, \mathrm{H}_h^{\boldsymbol{\Phi}})$, we get

$$\begin{aligned} \|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) - (\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq\ & \widehat{C}\{\mathrm{dist}(\boldsymbol{\sigma}, \mathrm{H}_h^{\boldsymbol{\sigma}}) + \mathrm{dist}(\mathbf{u}, \widetilde{\mathrm{H}}_h^{\mathbf{u}}) + \mathrm{dist}(\boldsymbol{\Phi}, \mathrm{H}_h^{\boldsymbol{\Phi}})\} \\ & + \widetilde{\underline{C}}\alpha L_{\mathcal{D}}\|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) - (\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}}. \end{aligned} \tag{3.59}$$

The foregoing inequality readily implies the following main result.

Theorem 3.4. *Assume* (A2), (A3) *and that* $\underline{\widetilde{C}}\alpha L_{\mathcal{D}} \leq 1-\delta$, *with* $\delta \in]0,1[$. *Then, there holds*

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \delta^{-1}\widehat{C}\{\mathrm{dist}(\boldsymbol{\sigma},\mathrm{H}_h^{\boldsymbol{\sigma}})+\mathrm{dist}(\mathbf{u},\widetilde{\mathrm{H}}_h^{\mathbf{u}})+\mathrm{dist}(\boldsymbol{\Phi},\mathrm{H}_h^{\boldsymbol{\Phi}})\}.$$

Exactly as remarked at the end of Sec. 2.3, we also stress here that the optimal value of δ is $1/2$, whence we obtain the assumption $\underline{\widetilde{C}}\alpha L_{\mathcal{D}} \leq 1/2$ and the Céa estimate

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq 2\widehat{C}\{\mathrm{dist}(\boldsymbol{\sigma},\mathrm{H}_h^{\boldsymbol{\sigma}})+\mathrm{dist}(\mathbf{u},\widetilde{\mathrm{H}}_h^{\mathbf{u}})+\mathrm{dist}(\boldsymbol{\Phi},\mathrm{H}_h^{\boldsymbol{\Phi}})\}. \quad (3.60)$$

We end this section with the rates of convergence of our mixed finite element solution $(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)$, for which we first recall the approximation properties of the finite element subspaces involved (see Ref. 12).

$(\mathbf{AP}_h^{\boldsymbol{\sigma}})$ there exists $C>0$, independent of h, such that for each $\boldsymbol{\tau}\in\mathbb{H}^1(\Omega)\cap\mathbb{H}_0(\mathbf{div};\Omega)$ with $\mathbf{div}(\boldsymbol{\tau})\in\mathbf{H}^1(\Omega)$ there holds

$$\mathrm{dist}(\boldsymbol{\tau},\mathrm{H}_h^{\boldsymbol{\sigma}}) \leq Ch\{\|\boldsymbol{\tau}\|_{1,\Omega}+\|\mathbf{div}(\boldsymbol{\tau})\|_{1,\Omega}\}.$$

$(\mathbf{AP}_h^{\mathbf{u}})$ there exists $C>0$, independent of h, such that for each $\mathbf{v}\in\mathbf{H}^1(\Omega)$ there holds

$$\mathrm{dist}(\mathbf{v},\mathrm{H}_h^{\mathbf{u}}) \leq Ch\|\mathbf{v}\|_{1,\Omega}.$$

$(\mathbf{AP}_h^{\boldsymbol{\Phi}})$ there exists $C>0$, independent of h, such that for each $\boldsymbol{\Psi}\in\mathbb{H}^1(\Omega)\cap\mathbb{L}^2_{\mathtt{skew}}(\Omega)$ there holds

$$\mathrm{dist}(\boldsymbol{\Psi},\mathrm{H}^{\boldsymbol{\Phi}}) \leq Ch\|\boldsymbol{\Psi}\|_{1,\Omega}.$$

Note here that, while $(\mathbf{AP}_h^{\mathbf{u}})$ provides the approximation property of $\mathrm{H}_h^{\mathbf{u}}$, the fact that this space is contained in $\widetilde{\mathrm{H}}_h^{\mathbf{u}}$ implies that $\mathrm{dist}(\mathbf{v},\widetilde{\mathrm{H}}_h^{\mathbf{u}}) \leq \mathrm{dist}(\mathbf{v},\mathrm{H}_h^{\mathbf{u}})$, and hence $(\mathbf{AP}_h^{\mathbf{u}})$ also serves to estimate the distance to $\widetilde{\mathrm{H}}_h^{\mathbf{u}}$. According to the above discussion, the main result of this section is stated as follows.

Theorem 3.5. *Assume* (A2), (A3) *and that* $\underline{\widetilde{C}}\alpha L_{\mathcal{D}} \leq 1/2$. *In addition, suppose that the solution* $(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}}) := ((\boldsymbol{\sigma},\boldsymbol{\rho}),(\mathbf{u},\boldsymbol{\Phi},\boldsymbol{\lambda}))\in\mathbf{H}\times\mathbf{Q}$ *of* (3.12) *verifies* $\boldsymbol{\sigma}\in\mathbb{H}^1(\Omega)$, $\mathbf{div}(\boldsymbol{\sigma})\in\mathbf{H}^1(\Omega)$, $\mathbf{u}\in\mathbf{H}^1(\Omega)$, *and* $\boldsymbol{\Phi}\in\mathbb{H}^1(\Omega)$. *Then, there exists a positive constant* C, *independent of* h, *such that*

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq Ch\{\|\boldsymbol{\sigma}\|_{1,\Omega}+\|\mathbf{div}(\boldsymbol{\sigma})\|_{1,\Omega}+\|\mathbf{u}\|_{1,\Omega}+\|\boldsymbol{\Phi}\|_{1,\Omega}\}. \quad (3.61)$$

Proof. It is a simple consequence of the Céa estimate (3.60), the additional regularity assumptions on the solution, and the approximation properties $(\mathbf{AP}_h^{\boldsymbol{\sigma}})$, $(\mathbf{AP}_h^{\mathbf{u}})$, and $(\mathbf{AP}_h^{\boldsymbol{\Phi}})$. □

4. Implementation of the Methods

We now refer to the practical implementation of (2.8). The extension to (3.12) proceeds similarly. More precisely, in what follows, we employ a fictional time variable in a gradient flow fashion to implement the solution of problem (2.8), thus rendering problem (2.1) convex for a sufficiently small time step. This means that, given a time step Δt, $k \in \mathrm{N}$, and a previous iteration u_k, we modify the extended problem (2.6) to obtain

$$\min_{(v,\eta)\in H} \max_{\xi\in Q} \left\{ \alpha\mathcal{D}(v) + \frac{1}{2}a(v,v) + \langle v-\eta,\xi\rangle + \frac{\beta}{2}\|\eta\|^2 + \frac{1}{2\Delta t}\|v-u_k\|_{\mathcal{V}}^2 \right\}, \tag{4.1}$$

where we recall from Sec. 2.2 that $H = \mathcal{V}\times Q$. Then, the first-order conditions of this problem are given by the following: Find $((u,\rho),\lambda)\in H\times Q$ such that

$$\begin{aligned} \langle u,v\rangle + \Delta t a(u,v) + \beta\Delta t\langle\lambda,\eta\rangle + \Delta t\langle v-\eta,\rho\rangle & \\ = \alpha\Delta t F_{u_k}(v) + \langle u_k,v\rangle \quad \forall\,(v,\eta)\in H, & \\ \langle u-\lambda,\xi\rangle = 0 \quad \forall\,\xi\in Q, & \end{aligned} \tag{4.2}$$

where the nonlinear term is treated explicitly, and which is well-posed in virtue of Theorem 2.2. The resulting solution of (4.2) is then redenoted $((u_{k+1},\rho_{k+1}),\lambda_{k+1})$. Our main modification to the classical time-dependent scheme used to implement registration problems is that the extended variables prevent the orthogonality to the kernel of the adjoint operator. Now, we establish a relationship between subsequent iterations to find a bound on the time step for stability.

Lemma 4.1. *Given an initial iteration $((u_0,\rho_0),\lambda_0)\in H\times Q$ and $n\in\mathrm{N}$, we let $((u_n,\rho_n),\lambda_n)$ and $((u_{n+1},\rho_{n+1}),\lambda_{n+1})$ be the solutions of (4.2) with $k=n-1$ and $k=n$, respectively. In addition, let $\widetilde{c}_a$ be the ellipticity constant of the bilinear form a (cf. (A1)), and define $\kappa_1(\Delta t) := (\frac{1}{\Delta t}+2\widetilde{c}_a-\alpha)$ and $\kappa_2(\Delta t) := (\frac{1}{\Delta t}+\alpha L_{\mathcal{D}})$. Then, there holds*

$$\kappa_1(\Delta t)\|u_{n+1}-u_n\|_{\mathcal{V}}^2 \le \kappa_2(\Delta t)\|u_n-u_{n-1}\|_{\mathcal{V}}^2. \tag{4.3}$$

Proof. Subtracting the corresponding equations of the problems (4.2) yielding $((u_n,\rho_n),\lambda_n)$ and $((u_{n+1},\rho_{n+1}),\lambda_{n+1})$, we obtain

$$\begin{aligned} \frac{1}{\Delta t}\langle u_{n+1}-u_n,v\rangle + a(u_{n+1}-u_n,v) + \beta\langle\lambda_{n+1}-\lambda_n,\eta\rangle + \langle v-\eta,\rho_{n+1}-\rho_n\rangle & \\ = \alpha(F_{u_n}-F_{u_{n-1}})(v) + \frac{1}{\Delta t}\langle u_n-u_{n-1},v\rangle \quad \forall\,(v,\eta)\in H, & \end{aligned} \tag{4.4}$$

and

$$\langle u_{n+1}-u_n-\lambda_{n+1}+\lambda_n,\xi\rangle = 0 \quad \forall\,\xi\in Q. \tag{4.5}$$

from which, testing (4.4) and (4.5) against $(v, \eta) = (u_{n+1} - u_n, \rho_{n+1} - \rho_n)$ and $\xi = \lambda_{n+1} - \lambda_n$, respectively, we deduce that

$$\frac{1}{\Delta t}\|u_{n+1} - u_n\|_{\mathcal{V}}^2 + a(u_{n+1} - u_n, u_{n+1} - u_n) + \beta\|\lambda_{n+1} - \lambda_n\|_{\mathcal{V}}^2$$

$$= \alpha(F_{u_n} - F_{u_{n-1}})(u_{n+1} - u_n) + \frac{1}{\Delta t}\langle u_n - u_{n-1}, u_{n+1} - u_n\rangle.$$

Next, using the ellipticity of a (cf. (A1)), the Lipschitz continuity of $\nabla\mathcal{D}$ (cf. (A2)), and Young's inequality, we arrive at

$$\left(\frac{1}{2\Delta t} + \widetilde{c}_a\right)\|u_{n+1} - u_n\|_{\mathcal{V}}^2 \leq \frac{\alpha}{2}\|u_{n+1} - u_n\|^2 + \left(\frac{L_{\mathcal{D}}\alpha}{2} + \frac{1}{2\Delta t}\right)\|u_n - u_{n-1}\|^2,$$

which leads to the desired result after a minor algebraic rearrangement. □

We stress here that the estimate (4.3) (cf. Lemma 4.1) becomes useless if $\kappa_1(\Delta t) \leq 0$. According to it, we now provide a way to bound how small Δt should be in order to guarantee that $\kappa_1(\Delta t) > 0$.

Lemma 4.2. *Problem* (4.1) *is unconditionally stable in time, that is stable for any fixed time step* Δt*, if* $\alpha < 2\widetilde{c}_a$*. It is otherwise stable if* $\Delta t < \frac{1}{\alpha - 2\widetilde{c}_a}$.

Proof. We first observe that if $\alpha < 2\widetilde{c}_a$, then, independently of Δt, $\kappa_1(\Delta t)$ remains always strictly positive, bounded below precisely by $2\widetilde{c}_a - \alpha$. Otherwise, the strict positivity of $\kappa_1(\Delta t)$ is guaranteed only by imposing $\frac{1}{\Delta t} > \alpha - 2\widetilde{c}_a$. □

Unfortunately, the previous scheme does not guarantee convergence for arbitrary α. Indeed, it is clear from (4.3) that in order to obtain $\|u_{n+1} - u_n\| \leq \delta\|u_n - u_{n-1}\|$, with $\delta \in]0, 1[$, it suffices to require that $\kappa_2(\Delta t) < \kappa_1(\Delta t)$, which yields the condition $\alpha < \frac{2c_a}{L_{\mathcal{D}}+1}$. Alternatively, if we consider variable time steps, we can prove the following result.

Lemma 4.3. *Let* $\{\Delta t^k\}_{k\in\mathrm{N}}$ *be an arbitrary sequence of time steps, and given an initial iteration* $((u_0, \rho_0), \lambda_0) \in H \times Q$ *and* $n \in \mathrm{N}$*, we let* $((u_n, \rho_n), \lambda_n)$ *and* $((u_{n+1}, \rho_{n+1}), \lambda_{n+1})$ *be the solutions of* (4.2) *with* $(k, \Delta t) = (n - 1, \Delta t^n)$ *and* $(k, \Delta t) = (n, \Delta t^{n+1})$*, respectively. Then, there holds*

$$\kappa_1(\Delta t^{n+1})\|u_{n+1} - u_n\|_{\mathcal{V}}^2 \leq \kappa_2(\Delta t^n)\|u_n - u_{n-1}\|_{\mathcal{V}}^2. \tag{4.6}$$

Consequently, under the assumption

$$\frac{1}{\Delta t^{n+1}} > \frac{1}{\Delta t^n} + \alpha(L_{\mathcal{D}} + 1) - 2\widetilde{c}_a, \tag{4.7}$$

the absolute step-wise error is strictly decreasing.

Proof. The derivation of the relationship between $\|u_{n+1} - u_n\|_{\mathcal{V}}^2$ and $\|u_n - u_{n-1}\|_{\mathcal{V}}^2$ is analogous to the one in the proof of Lemma 4.1, except for a minor modification.

In fact, as time steps are different, the time derivatives gives rise to new terms which cancel out, that is

$$\left\langle \frac{u_{n+1}}{\Delta t^{n+1}} - \frac{u_n}{\Delta t^n}, u_{n+1} - u_n \right\rangle = \frac{1}{\Delta t^{n+1}} \|u_{n+1} - u_n\|^2 + \left(\frac{1}{\Delta t^{n+1}} - \frac{1}{\Delta t^n} \right) \times \langle u_n, u_{n+1} - u_n \rangle,$$

and

$$\left\langle \frac{u_n}{\Delta t^{n+1}} - \frac{u_{n-1}}{\Delta t^n}, u_{n+1} - u_n \right\rangle = \frac{1}{\Delta t^n} \langle u_n - u_{n-1}, u_{n+1} - u_n \rangle + \left(\frac{1}{\Delta t^{n+1}} - \frac{1}{\Delta t^n} \right) \times \langle u_n, u_{n+1} - u_n \rangle.$$

Finally, the condition relating the subsequent time steps Δt^{n+1} and Δt^n is obtained by imposing $\kappa_2(\Delta t^n) < \kappa_1(\Delta t^{n+1})$. □

The above formulation and its associated analysis apply straightforwardly to the mixed case, the only difference being that, while the $\mathbf{H}^1$ inner product is employed in the regularizing terms for the primal case, the $\mathbf{L}^2$ one is utilized for the mixed approach.

5. Numerical Examples

In this section, we present several numerical examples to show the effectiveness of the proposed formulations. All tests were implemented with the FEniCS library.[1] For this, we will use in the primal case the same regularizer used in the mixed formulation, that is the bilinear form defined by (3.2), which arises from the Hooke law for elastic materials. Thus, as already announced at the beginning of Sec. 2.4, the abstract unknown u utilized in Secs. 2.1–2.3, and 4, is rewritten here as $\mathbf{u}$ to denote the respective displacement vector. We consider the problem with null traction boundary conditions so that it kernel is given by the space of rigid motions Q (cf. (3.3)), and consider the similarity functional given by the squared error, i.e.

$$\mathcal{D}(\mathbf{u}) = \int_\Omega (T(\boldsymbol{x} + \mathbf{u}(\boldsymbol{x})) - R(\boldsymbol{x}))^2,$$

where the maps $R, T : \Omega \to [0, 1]$ denote the reference and target images, respectively, and are such that the gradient $\nabla\mathcal{D}$ fulfills condition (A2). In what follows, we consider the domain $\Omega = (0, 1)^2$, and all examples, except for the convergence one, use the classic time regularization scheme described in Sec. 4. Also, only in the real-case study we use the time-adaptivity strategy presented in Sec. 4. For the other examples, we used $\Delta t \propto \alpha^{-1}$ justified by Lemma 4.2, which does not account for the ellipticity constant of the problem but gives satisfactory results nonetheless. The Young modulus E and Poisson ratio ν are related to the Lamé parameters through $\lambda_s = \frac{E\nu}{(1+\nu)(1-2\nu)}$ and $\mu_s = \frac{E}{2(1+\nu)}$.

5.1. *Convergence*

We consider the reference and target images

$$R(\boldsymbol{x}) = \exp(-20\|\boldsymbol{x} - 0.3(1,1)\|^2)$$

and

$$T(\boldsymbol{x}) = \exp(-20\|\boldsymbol{x} - 0.7(1,1)\|),$$

respectively, where $\boldsymbol{x} = (x_1, x_2)^{\mathtt{t}}$ with parameters $\mu_s = \lambda_s = \beta = 1$ and $\alpha = 0.1$. We define also the individual errors

$$\mathtt{e}_0(\mathbf{u}) := \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}, \quad \mathtt{e}_1(\mathbf{u}) := \|\mathbf{u} - \mathbf{u}_h\|_{1,\Omega}, \quad \mathtt{e}_0(\boldsymbol{\sigma}) := \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega},$$

$$\mathtt{e}(\boldsymbol{\sigma}) := \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}, \quad \mathtt{e}(\mathbf{u}) := \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}, \quad \text{and} \quad \mathtt{e}(\boldsymbol{\Phi}) := \|\boldsymbol{\Phi} - \boldsymbol{\Phi}_h\|_{0,\Omega},$$

and the respective experimental rates of convergence

$$\mathtt{r}_0(\mathbf{u}) := \frac{\log(\mathtt{e}_0(\mathbf{u})/\mathtt{e}_0'(\mathbf{u}))}{\log(h/h')}, \quad \mathtt{r}_1(\mathbf{u}) := \frac{\log(\mathtt{e}_1(\mathbf{u})/\mathtt{e}_1'(\mathbf{u}))}{\log(h/h')},$$

$$\mathtt{r}_0(\boldsymbol{\sigma}) := \frac{\log(\mathtt{e}_0(\boldsymbol{\sigma})/\mathtt{e}_0'(\boldsymbol{\sigma}))}{\log(h/h')}, \quad \mathtt{r}(\boldsymbol{\sigma}) := \frac{\log(\mathtt{e}(\boldsymbol{\sigma})/\mathtt{e}'(\boldsymbol{\sigma}))}{\log(h/h')},$$

$$\mathtt{r}(\mathbf{u}) := \frac{\log(\mathtt{e}(\mathbf{u})/\mathtt{e}'(\mathbf{u}))}{\log(h/h')}, \quad \mathtt{r}(\boldsymbol{\Phi}) := \frac{\log(\mathtt{e}(\boldsymbol{\Phi})/\mathtt{e}'(\boldsymbol{\Phi}))}{\log(h/h')},$$

where $\mathtt{e}$ y $\mathtt{e}'$, with and without subindex, denote in each case the errors of two consecutive triangulations with meshsizes given by h and h'.

We report the convergence results for the primal (2.8) and mixed (3.12) formulations in Tables 1 and 2, respectively, with respect to a solution of higher resolution, where the mixed scheme is set with the BDM elements described in (3.36). We stress that this problem was solved with a low α, thus results from the point of view of registration are not satisfactory, but they help us to verify convergence, as it is theoretically established for small α without the time stabilization terms (see Sec. 4). In particular, the $O(h)$ and $O(h^2)$ rates of convergence for $\|\mathbf{u} - \mathbf{u}_h\|_{1,\Omega}$ and $\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}$, respectively, which are predicted by (2.29) (cf. Theorem 2.4) and (2.35) (cf. Theorem 2.5), are confirmed by the sixth and fourth columns of Table 1. Nevertheless, the convergence of the extended mixed scheme shown in Table 2 seems a bit slow for $\mathtt{e}_0(\boldsymbol{\sigma})$ and slightly oscillating for $\mathtt{e}(\mathbf{u})$, which could be originated by an insufficient number of degrees of freedom employed.

Table 1. Errors and convergence rates for the primal extended scheme with $\alpha = 0.1$.

N_{dofs}	$h_{\max}$	$\mathtt{e}_0(\mathbf{u})$	$\mathtt{r}_0(\mathbf{u})$	$\mathtt{e}_1(\mathbf{u})$	$\mathtt{r}_1(\mathbf{u})$
56	3.536e-01	1.756e-03	—	1.959e-02	—
168	1.768e-01	5.669e-04	1.631	1.210e-02	0.695
584	8.839e-02	1.636e-04	1.793	6.253e-03	0.952
2184	4.419e-02	4.291e-05	1.931	3.147e-03	0.990
8456	2.210e-02	1.082e-05	1.988	1.575e-03	0.998
33288	1.105e-02	2.649e-06	2.030	7.878e-04	0.999

Table 2. Errors and convergence rates for the mixed extended scheme with $\alpha = 0.1$.

N_{dofs}	$h_{\max}$	$\mathtt{e}_0(\boldsymbol{\sigma})$	$\mathtt{r}_0(\boldsymbol{\sigma})$	$\mathtt{e}(\boldsymbol{\sigma})$	$\mathtt{r}(\boldsymbol{\sigma})$
95	7.071e-01	9.059e-03	—	8.327e-02	—
327	3.536e-01	3.304e-03	1.455	5.103e-02	0.706
1223	1.768e-01	1.285e-03	1.363	3.217e-02	0.666
4743	8.839e-02	3.924e-04	1.711	1.632e-02	0.979
18695	4.419e-02	1.262e-04	1.637	8.122e-03	1.006
N_{dofs}	$h_{\max}$	$\mathtt{e}(\mathbf{u})$	$\mathtt{r}(\mathbf{u})$	$\mathtt{e}(\boldsymbol{\Phi})$	$\mathtt{r}(\boldsymbol{\Phi})$
95	7.071e-01	5.783e+00	—	6.327e+00	—
327	3.536e-01	2.194e-04	1.469	5.642e-04	1.345
1223	1.768e-01	1.126e-04	0.960	2.416e-04	1.224
4743	8.839e-02	5.731e-05	0.975	1.128e-04	1.098
18695	4.419e-02	3.129e-05	0.873	5.609e-05	1.008

5.2. *To extend or not to extend*

In this test, we compare the results of the Neumann solver with and without extending the formulation, i.e. without the added degrees of freedom in Q and their corresponding terms to (2.8), which we call the standard formulation. The translation images are defined as in the convergence test, whereas the rotation images are given by

$$R(\boldsymbol{x}) = \varphi(\boldsymbol{S}\boldsymbol{x}) \quad \text{and} \quad T(\boldsymbol{x}) = \varphi(\boldsymbol{S}\boldsymbol{R}\boldsymbol{x}),$$

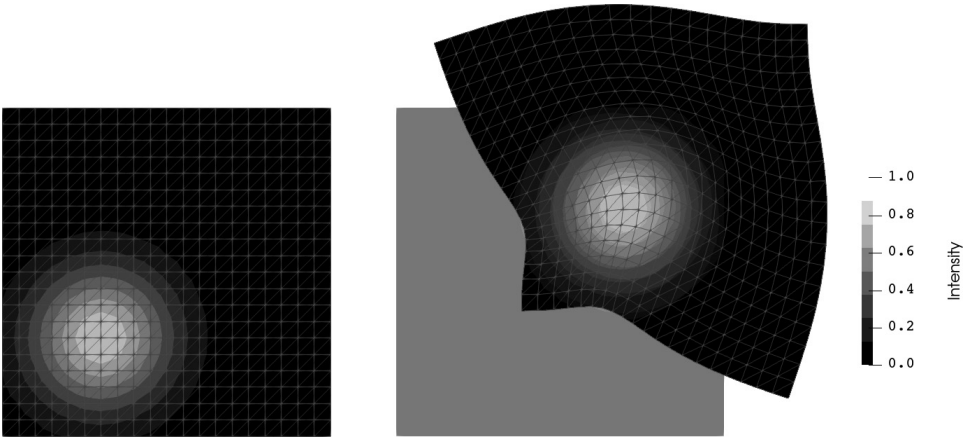
where

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix}, \quad \boldsymbol{R} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$

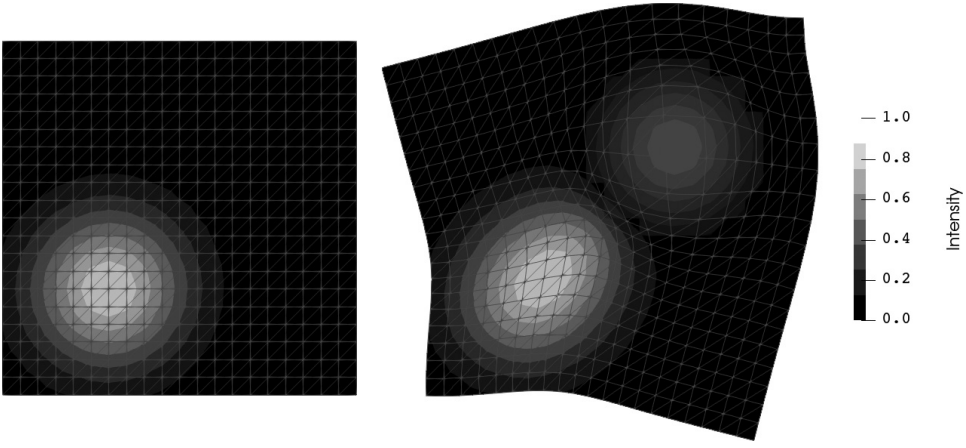
and the function $\varphi(\boldsymbol{x}) = \exp(-C|x|^2)$. The parameters used are given by $E = 10^3$, $\nu = 0.3$, $\alpha = 10^4$, $\Delta t = 0.1/\alpha$, $\beta = 1$, $C = 20$, $a = 0.4$, and the convergence criterion is given by a threshold on the similarity, so that the simulation stops when $\mathcal{D}(\mathbf{u}) \leq 0.01\mathcal{D}(\mathbf{0})$. In Figs. 1 and 2, which display the reference image and the warped reference image with the target image in the background, we notice that both translations and rotations cannot be captured up to the required tolerance without extending the formulation. In this regard, we stress that choosing a smaller Δt does not yield convergence in the non-extended scenario. This locking-like phenomenon is seen due to the choice of the convergence criterion, and indeed using another one such as the solution increments would yield convergence to a solution, albeit unsatisfactory.

5.3. *Translations in the quasi-incompressible case*

In this test, we register the translation images for the primal (2.8) and mixed (3.12) formulations, both with $E = 15$, $\nu = 0.4999$, $\alpha = 100$, $\Delta t = 0.1/\alpha$, $\beta = 1$, and time regularization terms were included and a tolerance of 10^{-8} for the absolute ℓ^∞ error between two subsequent steps was used. The results are reported in Fig. 3, where



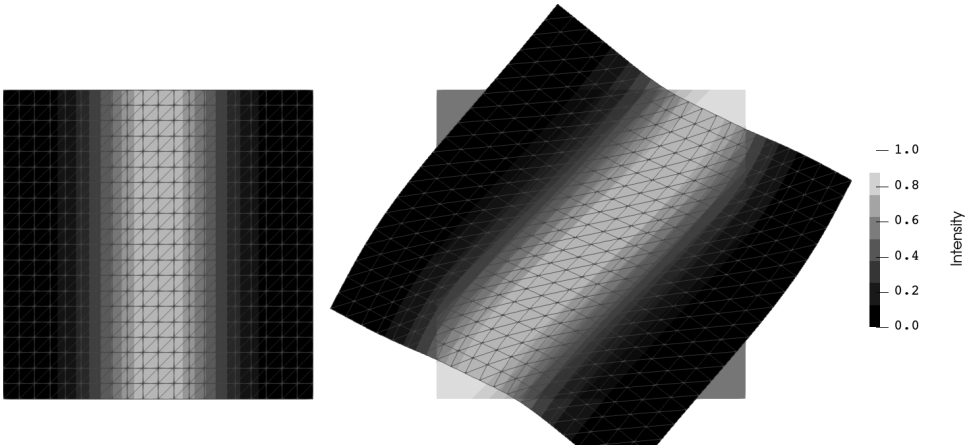
(a) Extended case



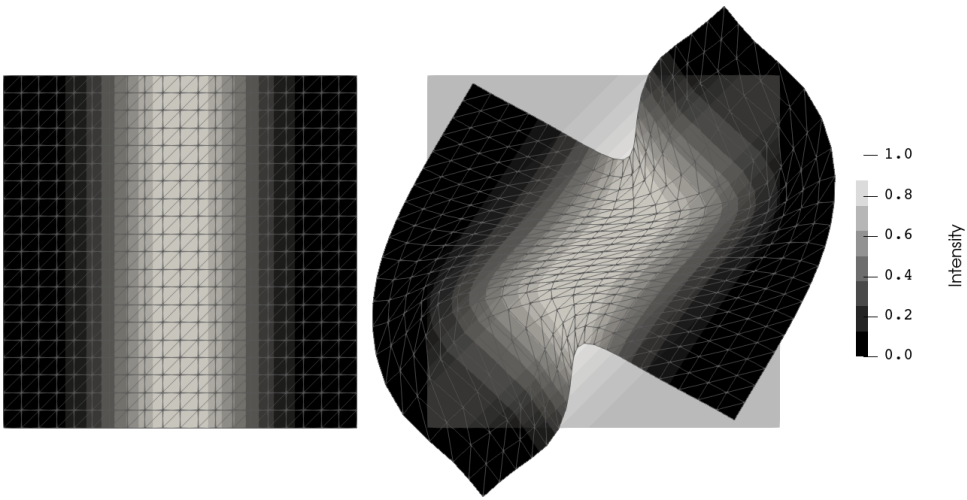
(b) Standard case

Fig. 1. Comparison warped reference images in translation example. We present the reference image $R(\boldsymbol{x})$ in the first column and the deformed reference image $R \circ (\boldsymbol{I} + \boldsymbol{u}_h)^{-1}(\boldsymbol{x})$ with the target image $T(\vec{x})$ in the background in the second column.

the rigid motion components obtained were $\lambda = (0.386, 0.396, 0.022)$ for the primal case and $\lambda = (0.402, 0.381, -0.056)$ for the mixed one. As λ is a rigid motion, the first two components are translations in x and y, whereas the third one represents a rotation. The solution in this case presents no rotation and has by construction a translation of 0.4 in each axis, which is coherent with the results obtained. We highlight that the primal formulation took 213 iterations to achieve convergence, whereas the mixed one took 102. This difference is mainly due to the locking effects generated by $\nu \approx 0.5$ in the primal formulation, which are fully overcome by the mixed one.



(a) Extended case

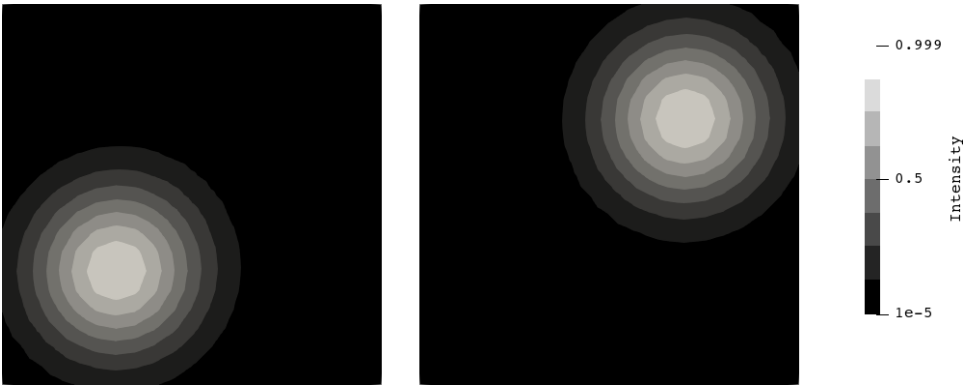


(b) Standard case

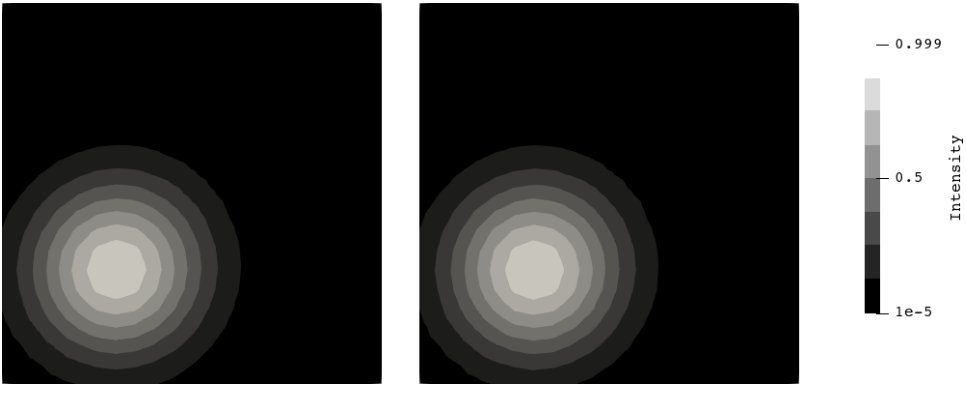
Fig. 2. Comparison warped reference images in rotation example. We present the reference image $R(\boldsymbol{x})$ in the first column and the deformed reference image $R \circ (\boldsymbol{I} + \boldsymbol{u}_h)^{-1}(\boldsymbol{x})$ with the target image $T(\vec{x})$ in the background in the second column.

Table 3. Extended vs. standard in terms of iterations and execution time on a personal computer.

	Formulation	Iterations	time $[s]$
Translation	Extended	64	3.516
	Standard	1000	—
Rotation	Extended	51	3.454
	Standard	1000	—



(a) R and T, reference and target images.



(b) Warped target images $T \circ (\boldsymbol{I} + \boldsymbol{u}_h)$ for primal and mixed formulation.

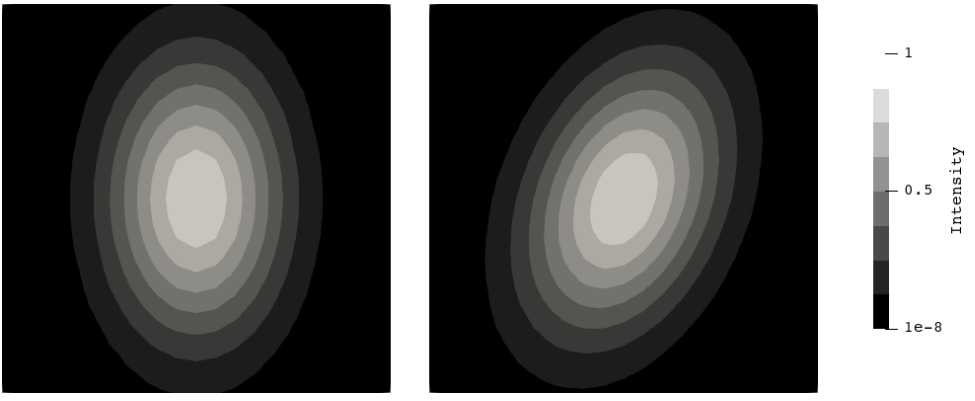
Fig. 3. Solutions of the primal and mixed formulations of the translation test.

5.4. *Rotations in the quasi-incompressible case*

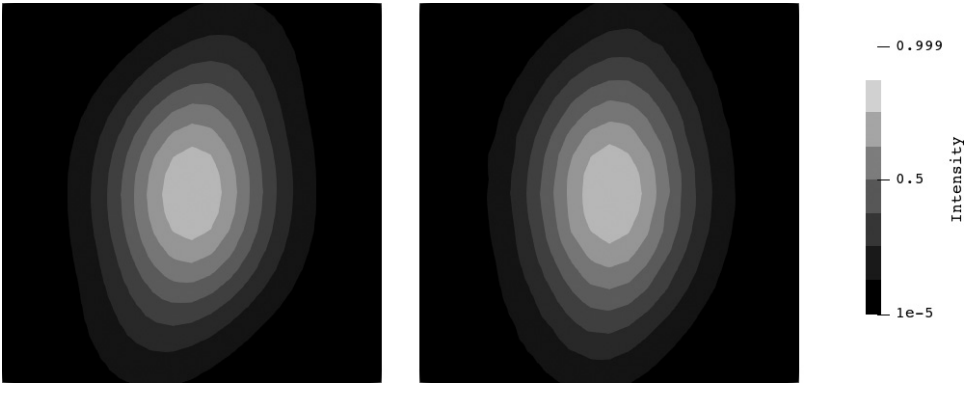
This test was performed for the same settings of the translation example but with the rotation images using $C = 20$ and $a = 0.4$. Results are reported in Fig. 4, and the rigid motions obtained in this case are

$$\lambda = (-2.84310^{-4}, 3.12010^{-5}, -7.08410^{-2}) \quad \text{and}$$
$$\lambda = (6.79810^{-5}, 6.04210^{-4}, -1.47610^{-3}),$$

for the primal and mixed cases, respectively. We remark that we did not allow for more than 1000 iterations in time, which was achieved by the primal case still without reaching the required tolerance. The mixed one instead converged after 74 iterations, which is again explained by the superiority of the mixed formulation in the quasi-incompressible case.



(a) R and T, reference and target images.



(b) R and T, reference and target images.

Fig. 4. Solutions of the primal and mixed formulations of the rotation test.

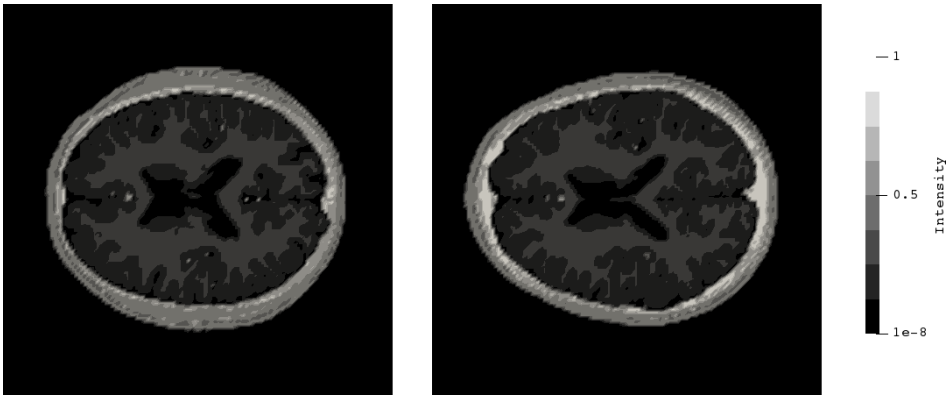
5.5. *Application to the image registration of the human brain*

The real application is performed on brain images obtained in Ref. 16. We use this case as well to test the condition on the time step given by

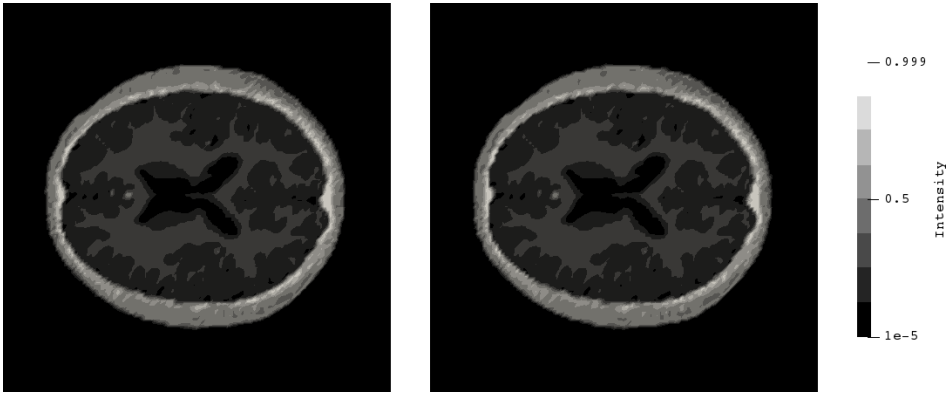
$$\Delta t^{n+1} < \frac{\Delta t^n}{1 + \Delta t^n(\alpha(L_{\mathcal{D}} + 1) - c_a)}. \tag{5.1}$$

Two important observations are in place for condition (5.1). One is that it guarantees the convergence of $\|u_{n+1} - u_n\|$, and not of $\|u_{n+1} - u_n\|/\Delta t_n$, which means that possibly the error performed by means of incorporating the time terms might not disappear. The second one is that it does not stall the simulation within a certain time. To see this, assume $\Delta t^0 = \tau = (\alpha(L_{\mathcal{D}} + 1) - c_a)^{-1}$. This choice gives $\Delta t^n < \tau/(n+1)$, and thus we cannot insure that $\sum_n \Delta t^n < \infty$.

In turn, for the simulations we use the elastic constants $E = 15$ and $\nu = 0.3$. For the others constants we consider $\alpha = 10^4, \beta = 1, \Delta t^0 = 0.01/\alpha$ and a tolerance of

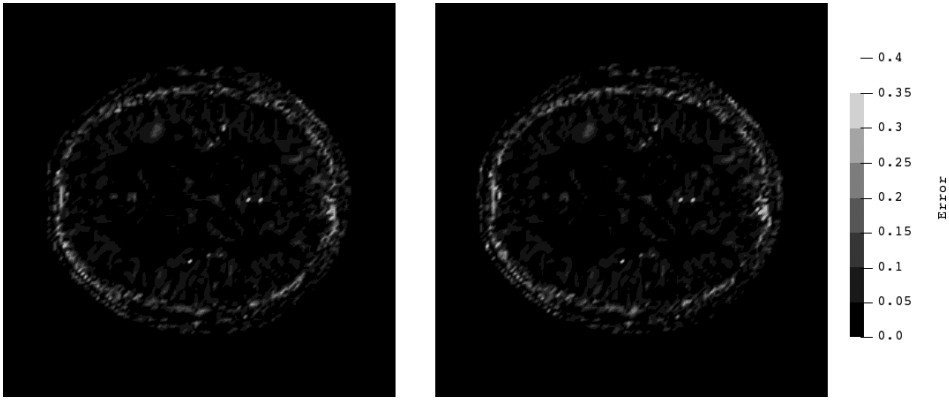


(a) Reference and template images.



(b) Primal and mixed formulation solutions.

Fig. 5. Results of registration for brain images scenario with $\alpha = 10^4$, $\beta = 1$.



(a) $|T \circ (\boldsymbol{I} + \vec{u}) - R|$ for primal and mixed formulations.

Fig. 6. Results of registration for brain images scenario with $\alpha = 10^4$, $\beta = 1$.

10^{-6} for a domain with 128×128 elements. We report the outcome in Figs. 5 and 6, that indicate sufficiently accurate results after convergence. To avoid an excessive reduction of the time step, we used (5.1) every 10 iterations.

6. Discussion

In this work, we present a way to formulate problems with Neumann boundary conditions in a mathematically consistent way so as not to lose information from the images but still keeping all the degrees of freedom from the original problem in both primal and mixed formulations, the latter being particularly important in the quasi-incompressible case. This method presents clear advantages for capturing rigid motions, i.e. translations and rotations. This gives rise to modeling considerations, such as whether it is important or not to consider a regularizer with rigid motions on its kernel. We could for instance device a model which only presents translations in its kernel, such as a gradient-regularized formulation, and use the last iterations as input for an elastic registration problem, so as to obtain relevant stress indicators. This consideration itself opens the possibility of setting iterated models for different objectives, such as a gradient regularizer initially for translations, and then an elastic regularizer for rigid motions and stress estimation. It is also important to mention that the strategy adopted in this work is purely monolithic, meaning that operator splitting techniques remain available for this kind of multivariable formulation. The bound on a possibly variable time step is the first one devised on image registration to the knowledge of the authors, and although it gives a rule for tuning parameters, it is far from sharp, and no considerations have been made so far regarding the convergence of the velocity to 0. Any scheme not satisfying this conditions risks of presenting non-convergent orbits. The issue of convergence in time remains largely an open question for image registration, and together with the construction of an efficient solver for the extended problem it presents some of the main questions to be answered in our future work.

Acknowledgments

This work received financial support from the Chilean National Agency for Research and Development (ANID) through Grant FONDECYT Regular #1180832, the Project Centro de Modelamiento Matemático (AFB170001) of the PIA program: Concurso Apoyo a Centros Cientificos y Tecnológicos de Excelencia con Financiamiento Basal, and the Becas-CONICYT Programme for foreign students; by Centro de Investigación en Ingeniería Matemática (CI^2MA), Universidad de Concepción; by the HPC-Europa3 Transnational Access programme; by the Monash Mathematics Research Fund S05802-3951284; and by the Ministry of Science and Higher Education of the Russian Federation within the framework of state support for the creation and development of World-Class Research Centers "Digital biodesign and personalised healthcare" No. 075-15-2020-926.

References

1. M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes and G. N. Wells, The FEniCS project version 1.5, *Arch. Numer. Softw.* **3** (2015) 9–23.
2. R. E. Amelon, K. Cao, K. Ding, G. E. Christensen, J. M. Reinhardt and M. L. Raghavan, Three-dimensional characterization of regional lung deformation, *J. Biomech.* **44** (2011) 2489–2495.
3. D. N. Arnold, F. Brezzi and J. Douglas, PEERS: A new mixed finite element method for plane elasticity, *Japan J. Appl. Math.* **1** (1984) 347–367.
4. D. N. Arnold, R. Falk and R. Winther, Mixed finite element methods for linear elasticity with weakly imposed symmetry, *Math. Comp.* **76** (2007) 1699–1723.
5. B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein and J. C. Gee, A reproducible evaluation of ants similarity metric performance in brain image registration, *Neuroimage* **54** (2011) 2033–2044.
6. G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag and A. V. Dalca, An unsupervised learning model for deformable medical image registration, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 9252–9260.
7. N. Barnafi, G. N. Gatica and D. E. Hurtado, Primal and mixed finite element methods for deformable image registration problems, *SIAM J. Imaging Sci.* **11** (2018) 2529–2567.
8. N. Barnafi, G. N. Gatica, D. E. Hurtado, W. Miranda and R. Ruiz-Baier, *A posteriori* error estimates for primal and mixed finite element approximations of the deformable image registration problem, preprint 2018-50, Centro de Investigación en Ingeniería Matemática (CI^2MA), Universidad de Concepción, Chile (2018), `http://www.ci2ma.udec.cl`.
9. D. Boffi, F. Brezzi and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics, Vol. 44 (Springer, 2013).
10. S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Method*, Texts in Applied Mathematics, Vol. 15, 3rd edn. (Springer, 2008).
11. F. Brezzi, J. Douglas Jr. and L. D. Marini, Two families of mixed finite elements for second order elliptic problems, *Numer. Math.* **47** (1985) 217–235.
12. F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Series in Computational Mathematics, Vol. 15 (Springer-Verlag, 1991).
13. S. Choi, E. A. Hoffman, S. E. Wenzel, M. H. Tawhai, Y. Yin, M. Castro and C. L. Lin, Registration-based assessment of regional lung function via volumetric CT images of normal subjects vs. severe asthmatics, *J. Appl. Physiol.* **115** (2013) 730–742.
14. G. E. Christensen, J. H. Song, W. Lu, I. El Naqa and D. A. Low, Tracking lung tissue motion and expansion/compression with inverse consistent image registration and spirometry, *Med. Phys.* **34** (2007) 2155–2163.
15. P. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications* (SIAM, 2013).
16. D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes and A. C. Evans, Design and construction of a realistic digital brain phantom, *IEEE Trans. Med. Imaging* **17** (1998) 463–468.
17. M. Ebner, M. Modat, S. Ferraris, S. Ourselin and T. Vercauteren, Forward–backward splitting in deformable image registration: A demons approach, in *2018 IEEE 15th Int. Symp. Biomedical Imaging* (*ISBI 2018*) (Washington, 2018), pp. 1065–1069.
18. G. N. Gatica, Analysis of a new augmented mixed finite element method for linear elasticity allowing $\mathbb{RT}_0 - \mathbb{P}_1 - \mathbb{P}_0$ approximations, *Math. Model. Numer. Anal.* **40** (2006) 1–28.

19. G. N. Gatica, *A Simple Introduction to the Mixed Finite Element Method. Theory and Applications*, SpringerBriefs in Mathematics (Springer, 2014).
20. G. N. Gatica, A. Márquez and S. Meddahi, A new dual-mixed finite element method for the plane linear elasticity problem with pure traction boundary conditions, *Comput. Methods Appl. Mech. Engrg.* **197** (2008) 1115–1130.
21. G. N. Gatica and W. L. Wendland, Coupling of mixed finite elements and boundary elements for a hyperelastic interface problem, *SIAM J. Numer. Anal.* **34** (1997) 2335–2356.
22. S. Haker, L. Zhu, A. Tannenbaum and S. Angenent, Optimal mass transport for registration and warping, *Int. J. Comput. Vision* **60** (2004) 225–240.
23. B. K. Horn and B. G. Schunck, Determining optical flow, *Artificial Intelligence* **17** (1981) 185–203.
24. D. E. Hurtado, N. Villarroel, C. Andrade, J. Retamal, G. Bugedo and A. R. Bruhn, Spatial patterns and frequency distributions of regional deformation in the healthy human lung, *Biomech. Model. Mechanobiol.* **16** (2017) 1413–1423.
25. D. E. Hurtado, N. Villarroel, J. Retamal, G. Bugedo and A. Bruhn, Improving the accuracy of registration-based biomechanical analysis: A finite element approach to lung regional strain quantification, *IEEE Trans. Med. Imag.* **35** (2016) 580–588.
26. E. Lee and M. Gunzburger, An optimal control formulation of an image registration problem, *J. Math. Imag. Vis.* **36** (2010) 69–80.
27. M. Lonsing and R. Verfürth, On the stability of BDMS *and* PEERS elements, *Numer. Math.* **99** (2004) 131–140.
28. J. Modersitzki, *Numerical Methods for Image Registration*, Numerical Mathematics and Scientific Computation (Oxford Science Publications, 2004).
29. C. Pöschl, J. Modersitzki and O. Scherzer, A variational setting for volume constrained image registration, *Inverse Probl. Imag.* **4** (2010) 505–522.
30. J. Retamal, D. E. Hurtado, N. Villarroel, A. Bruhn, G. Bugedo, M. B. P. Amato, E. L. V. Costa, G. Hedenstierna, A. Larsson and J. B. Borges Does regional lung strain correlate with regional inflammation in acute respiratory distress syndrome during nonprotective ventilation? An experimental porcine study, *Crit. Care Med.* **46** (2018) 591–599.
31. R. T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM J. Control Opt.* **14** (1976) 877–898.
32. A. Sotiras, C. Davatzikos and N. Paragios, Deformable medical image registration: A survey, *IEEE Trans. Med. Imag.* **32** (2013) 1153–1190.
33. G. Unal and G. Slabaugh, Coupled PDEs for non-rigid registration and segmentation, *IEEE Computer Society Conf. Computer Vision and Pattern Recognition* (*CVPR'05*), Vol. 1 (IEEE, 2005), pp. 168–175.
34. C. R. Vogel, *Computational Methods for Inverse Problems. With a Foreword by H. T. Banks*, Frontiers in Applied Mathematics, Vol. 23 (SIAM, 2002).
35. J. Wlazlo, R. Fessler, R. Pinnau, N. Siedow and O. Tse, Elastic image registration with exact mass preservation, preprint (2016), arXiv:1609.04043.