MATEMATIČKI VESNIK
МАТЕМАТИЧКИ ВЕСНИК
71, 1–2 (2019), 63–89
March 2019

research paper
оригинални научни рад

SPECTRAL APPROXIMATION OF A STRAIN-LIMITING NONLINEAR ELASTIC MODEL

Nicolò Gelmetti and Endre Süli

Abstract. We construct a numerical algorithm for the approximate solution of a nonlinear elastic limiting strain model based on the Fourier spectral method. The existence and uniqueness of the numerical solution are proved. Assuming that the weak solution to the boundary-value problem possesses suitable Sobolev regularity, the sequence of numerical solutions is shown to converge to the weak solution of the problem at an optimal rate. The numerical method represents a finite-dimensional system of nonlinear equations. An iterative method is proposed for the approximate solution of this system of equations and is shown to converge, at a linear rate, to the unique solution of the numerical method. The theoretical results are illustrated by numerical experiments.

1. Introduction

During the past decade there has been considerable progress in developing implicit constitutive models for the description of nonlinear responses of materials (see, for example, [10, 11]). In the field of solid mechanics one of the main achievements of implicit constitutive theory is in providing a theoretical background for nonlinear models involving the linearized strain. In particular, within the realm of implicit constitutive theory, it is possible to have models in which the linearized strain is in all circumstances a bounded function, even when the stress is very large. This subclass of implicit constitutive models, proposed by Rajagopal in [11], are referred to as *limiting strain models*, and have the potential to be useful in modelling stress concentration effects in instances when the gradient of the displacement is relatively small (e.g. in modelling brittle materials near crack tips or notches, or concentrated loads inside the body or on its boundary). Models with limiting *finite* strain are also found to be useful in describing the response of various soft tissues that exhibit the phenomenon of finite extensibility. We refer the reader to [1, 3, 4], for example, for a survey of

2010 Mathematics Subject Classification: 65N35, 74B20
Keywords and phrases: Spectral method; convergence; nonlinear elasticity; limiting strain models.

physical aspects of limiting strain models and recent results concerning the existence of solutions.

The mathematical literature on the analysis of numerical algorithms for limiting strain models is extremely limited: apart from the recent paper by Bonito et al. [2], concerned with the construction and convergence analysis of low-order mixed finite element approximations of limiting strain models on multidimensional polytopal domains subject to a homogeneous Dirichlet boundary condition, the present paper appears to be the only other work in the direction of rigorous analysis of a numerical method for a limiting strain model. The numerical algorithm considered here is posed in the context of the paper [4], i.e., in an axiparallel parallelopipedal domain $\Omega \subset \mathbb{R}^d$, subject to periodic boundary conditions, as this is the only setting involving the complete nonlinear system of equations in the model for which existence of a solution of any kind has been shown for the complete range $r \in (0, \infty)$ of the model parameter r (weak solution for $r \in (0, 2/d)$ and a renormalized solution for the complete range of $r \in (0, \infty)$).

2. Formulation of the problem and summary of the main results

As has been explained above, we shall consider a domain of a special form: namely an axiparallel parallelepiped, with spatially periodic boundary conditions in the various co-ordinate directions.

The problem under consideration here is therefore the following: suppose that $\Omega = (0, 2\pi)^d$, with $d \geq 2$, and f is a given d-component vector-function (the load-vector), which is 2π-periodic in each of the d co-ordinate directions. The case of $d = 1$ is also covered by the theory contained herein, but in the univariate case the solution to the problem can be written down in closed form, so the problem is uninteresting from the theoretical point of view. Our objective is to construct a Fourier spectral approximation (S_N, u_N) to (S, u), where $N \in \mathbb{N}$ is the degree of the d-variate trigonometric polynomial used, S is the stress tensor and u is the displacement, which belong to suitable function spaces consisting of symmetric $d \times d$ matrix functions and d-component vector functions, respectively, that are 2π-periodic in each co-ordinate direction, such that

$$-\operatorname{div} S = f \tag{1}$$

and

$$D(u) = F(S). \tag{2}$$

Here $F \in \mathrm{C}^1(\mathbb{R}^{d\times d}_{\mathrm{sym}}; \mathbb{R}^{d\times d})$ is defined by $F(S) := \frac{S}{(1+|S|^r)^{\frac{1}{r}}}$, $S \in \mathbb{R}^{d\times d}$, where $r > 0$, and $|\cdot|$ denotes the Frobenius norm on $\mathbb{R}^{d\times d}$, defined by $|X|^2 := X : X = \operatorname{tr}(X^{\mathrm{T}} X)$. It is a straightforward matter to show that the function F has the following properties:

(P1) $F(0) = 0$ and $|F(A)| \leq 1$ for all $A \in \mathbb{R}^{d\times d}$;

(P2) There exist two constants $c_a = c_a(r) > 0$ and $c_b \geq 1$ such that the following inequalities hold:

(P2a) $(F(A)-F(B)):(A-B)\geq c_a\,\dfrac{|A-B|^2}{(1+|A|+|B|)^{r+1}}\qquad\forall\,A,B\in\mathbb{R}^{d\times d},$

and $$F(A):A\geq c_a\,\frac{|A|^2}{1+|A|}\qquad\forall\,A\in\mathbb{R}^{d\times d};$$

(P2b) $$|F(A)-F(B)|\leq c_b|A-B|\qquad\forall\,A,B\in\mathbb{R}^{d\times d}.$$

The existence of such positive constants c_a and c_b appearing in **(P2a)** and **(P2b)** is an immediate consequence of the following two lemmas, whose proofs are contained in [4].

LEMMA 2.1. *For any $y\geq 0$ and $r>0$, we have that*

$$\min(1,2^{-1+\frac{1}{r}})\,(1+y)\leq(1+y^r)^{\frac{1}{r}}\leq\max(1,2^{-1+\frac{1}{r}})\,(1+y).$$

LEMMA 2.2. *Let $r>0$, and consider the mapping*

$$X\in\mathbb{R}^{d\times d}\mapsto F(X):=X(1+|X|^r)^{-\frac{1}{r}}\in\mathbb{R}^{d\times d}.$$

Then, for each $A,B\in\mathbb{R}^{d\times d}$, we have that $|F(A)-F(B)|\leq 2|A-B|$, and

$$(F(A)-F(B)):(A-B)\geq\min(1,2^{r-\frac{1}{r}})\,|A-B|^2\,(1+|A|+|B|)^{-r-1}.$$

Thanks to Lemma 2.2, **(P2b)** holds with $c_b=2$ and the first inequality in **(P2a)** holds with $c_a=\min\left(1,2^{r-\frac{1}{r}}\right)$; thanks to Lemma 2.1, the second inequality in **(P2a)** holds with $c_a=\min\left(1,2^{1-\frac{1}{r}}\right)$.

The next lemma collects some elementary but helpful results concerning the function F and related functions that will arise in our analysis.

LEMMA 2.3. *The following statements hold:*
(a) *Suppose that $\alpha>0$. The function $t\in[0,\infty)\mapsto(1+t)^{-\alpha}\in(0,1]$ is Lipschitz continuous, with Lipschitz constant α.*

(b) *Suppose that $\mu\in(0,1]$; then, the function $x\in\mathbb{R}^{d\times d}\mapsto|x|^\mu\in[0,\infty)$ is Hölder-continuous; in particular, $\big||x|^\mu-|y|^\mu\big|\leq\frac{1}{\mu}\big||x|-|y|\big|^\mu\leq\frac{1}{\mu}|x-y|^\mu\quad\forall\,x,y\in\mathbb{R}^{d\times d}$.*

(c) *Suppose that $\mu\in[1,\infty)$ and let $\mathcal{B}(0,R)$ be the closed ball in $\mathbb{R}^{d\times d}$ with radius $R>0$, centred at the origin; then, the function $x\in\mathcal{B}(0,R)\mapsto|x|^\mu\in[0,\infty)$ is Lipschitz-continuous; in particular, $\big||x|^\mu-|y|^\mu\big|\leq\mu R^{\mu-1}\big||x|-|y|\big|\leq\mu R^{\mu-1}|x-y|$ $\forall\,x,y\in\mathcal{B}(0,R)$.*

(d) *The composition of a $(0,1]$-valued Lipschitz-continuous function defined on $[0,\infty)$ and a $[0,\infty)$-valued Hölder continuous function defined on $\mathcal{B}(0,R)$, with Hölder exponent $\min(1,r)$, is a $(0,1]$-valued Hölder-continuous function defined on $\mathcal{B}(0,R)$, with exponent $\min(1,r)$.*

In particular, for any $\alpha>0$ and $r>0$, the function $x\in\mathcal{B}(0,R)\mapsto(1+|x|^r)^{-\alpha}\in(0,1]$ is Hölder continuous, with exponent $\min(1,r)$.

(e) *Suppose that $p>d^2$; then, $\mathrm{W}^{1,p}(\mathcal{B}(0,R))\hookrightarrow\mathrm{C}^{0,\alpha}(\mathcal{B}(0,R))$ with $\alpha=1-\frac{d^2}{p}$. In particular, for any $\varepsilon\in(0,1)$, the function $x\in\mathcal{B}(0,R)\mapsto\frac{x}{|x|^\varepsilon}\in\mathcal{B}(0,R^{1-\varepsilon})$ belongs to $\mathrm{W}^{1,p}(\mathcal{B}(0,R))$ for $p\in[1,\frac{d^2}{\varepsilon})$, and hence to $\mathrm{C}^{0,\delta}(\mathcal{B}(0,R))$ for $\delta\in(0,1-\varepsilon)$.*

The paper is structured as follows. In Section 3 we formulate the numerical approximation of the problem and recall from [4] various results concerning the existence and uniqueness of weak solutions for the range $r \in (0, \frac{2}{d})$ and the existence of a renormalized solution for the range $r \in (0, \infty)$. The existence proofs are based on various weak compactness arguments and are omitted as they do not directly relate to the topic of the present paper. For the sake of completeness of our discussion of the numerical method here we have however included the proof, from [4], of the existence and uniqueness of a solution to the numerical approximation of the boundary-value problem under consideration. In Section 4 we assume that the pair $(S, D(u))$ has additional regularity beyond that of a weak solution, i.e., that it belongs to a Sobolev space of high enough differentiability index so as to ensure continuity over $\overline{\Omega}$ of all components of S and $D(u)$, and we use a fixed point argument to prove that the numerical method exhibits optimal order convergence in the L^2 norm. The numerical method represents a finite-dimensional system of nonlinear equations. In Section 5 an iterative method is proposed for the approximate solution of this system of equations, and is shown to converge to the unique solution of the discretized problem. In Section 6 we report numerical experiments in order to illustrate the theoretical results of the paper through concrete examples. We conclude, in Section 7, with a summary of the main results of the paper and indications of some relevant open problems.

3. Definition of the approximation: existence and uniqueness of solutions

Consider the domain $\Omega := (0, 2\pi)^d$ in $\mathbb{R}^d$, $d \geq 2$. All function spaces consisting of real-valued 2π-periodic functions (by which we mean 2π-periodic in each of the d co-ordinate directions) will be labelled with the subscript $\#$; subspaces of these, consisting of 2π-periodic functions whose integral over Ω is equal to 0, will be labelled with the subscript $*$; in order to avoid notational clutter we shall not use the symbols $\#$ and $*$ in the various norm signs. It will be clear from the argument of the norm which of the symbols $\#$ or $*$ is intended. For example, $\mathrm{L}^p_{\#}(\Omega)$ will denote the Lebesgue space of all real-valued 2π-periodic functions v such that $|v|^p$ is integrable on Ω, equipped with the norm $\|\cdot\|_{\mathrm{L}^p(\Omega)}$. It is understood that the usual modification is made when $p = \infty$. Spaces of d-component vector functions, where each component belongs to a certain function space X, will be denoted by $[X]^d$, while spaces of $d \times d$ component matrix functions each of whose components is an element of X will be denoted by $[X]^{d\times d}$. Letting $\mathrm{C}^\infty_{\#}(\overline{\Omega})$ denote the linear space consisting of the restriction to $\overline{\Omega}$ of all real-valued 2π-periodic C^∞ functions defined on $\mathbb{R}^d$, we note that $\mathrm{C}^\infty_{\#}(\overline{\Omega})$ is dense in $\mathrm{L}^p_{\#}(\Omega)$ for all $p \in [1, \infty)$; analogously, $\mathrm{C}^\infty_*(\overline{\Omega})$ is dense in $\mathrm{L}^p_*(\Omega)$ for $1 \leq p < \infty$. The Sobolev space $\mathrm{W}^{1,p}_{\#}(\Omega)$, $1 \leq p < \infty$, will be defined as the closure of $\mathrm{C}^\infty_{\#}(\overline{\Omega})$ in the Sobolev norm $\|\cdot\|_{\mathrm{W}^{1,p}(\Omega)}$, where $\|v\|_{\mathrm{W}^{1,p}(\Omega)} := \left(\|v\|^p_{\mathrm{L}^p(\Omega)} + \|\nabla v\|^p_{\mathrm{L}^p(\Omega)}\right)^{\frac{1}{p}}$; here, $\|\nabla v\|_{\mathrm{L}^p(\Omega)} := \||\nabla v|\|_{\mathrm{L}^p(\Omega)}$, where $|\nabla v|$ denotes the Euclidean norm of ∇v. Analogously, $\mathrm{W}^{1,p}_*(\Omega)$, $1 \leq p < \infty$, will be defined as the closure of $\mathrm{C}^\infty_*(\overline{\Omega})$ in the Sobolev

norm $\|\cdot\|_{\mathrm{W}^{1,p}(\Omega)}$.

In the case of a d-component vector-valued function v defined on Ω the definition of the norm $\|v\|_{\mathrm{W}^{1,p}(\Omega)}$ is the same as above, except that $\|v\|_{\mathrm{L}^p(\Omega)} := \||v|\|_{\mathrm{L}^p(\Omega)}$, with $|\cdot|$ again signifying the Euclidean norm, while $\|\nabla v\|_{\mathrm{L}^p(\Omega)} := \||\nabla v|\|_{\mathrm{L}^p(\Omega)}$, where now $|\nabla v|$ denotes the Frobenius norm of the $d \times d$ matrix ∇v.

We further define $\mathrm{H}_{\#}(\mathrm{div};\Omega) := \{v \in [\mathrm{L}^2_{\#}(\Omega)]^d : \text{such that } \mathrm{div} v \in \mathrm{L}^2_{\#}(\Omega)\}$, equipped with the norm $\|v\|_{\mathrm{H}(\mathrm{div};\Omega)} := \left(\|v\|^2_{\mathrm{L}^2(\Omega)} + \|\mathrm{div} v\|^2_{\mathrm{L}^2(\Omega)}\right)^{\frac{1}{2}}$.

For $s > 0$, the (potentially fractional-order) Hilbertian Sobolev spaces of periodic functions $\mathrm{H}^s_{\#}(\Omega) := \mathrm{W}^{s,2}_{\#}(\Omega)$ and $\mathrm{H}^s_{*}(\Omega) := \mathrm{W}^{s,2}_{*}(\Omega)$ are defined analogously, as the closure of $\mathrm{C}^\infty_{\#}(\overline{\Omega})$ and $\mathrm{C}^\infty_{*}(\overline{\Omega})$, respectively, in the norm of $\mathrm{H}^s(\Omega) := \mathrm{W}^{s,2}(\Omega)$.

Our reason to work with function spaces whose elements integrate over Ω to 0 is that the functions S and u appearing in equations (1) and (2) can be modified by arbitrary additive constants without violating the equalities. In order to ensure uniqueness of the solution to the problem it is therefore necessary to fix these arbitrary additive constants, and we do so by demanding that the integrals of S and u over Ω are equal to 0.

We shall require the following periodic version of Korn's inequality [4].

LEMMA 3.1 (Korn's inequality in L^p). *Let $p \in (1,\infty)$, $d \geq 2$ and $\Omega := (0,2\pi)^d$. There exists a positive constant c_p such that the following inequality holds:*

$$\|\nabla v\|_{\mathrm{L}^p(\Omega)} \leq c_p \left(\|D(v)\|_{\mathrm{L}^p(\Omega)} + \|\mathrm{div}\, v\|_{\mathrm{L}^p(\Omega)}\right) \qquad \forall\, v \in [\mathrm{W}^{1,p}_{*}(\Omega)]^d,$$

and, hence, also, with a possibly different constant c_p,

$$\|\nabla v\|_{\mathrm{L}^p(\Omega)} \leq c_p \|D(v)\|_{\mathrm{L}^p(v)} \qquad \forall\, v \in [\mathrm{W}^{1,p}_{*}(\Omega)]^d.$$

Let, further, $D^{\mathrm{dev}}(v) := D(v) - \frac{1}{d}(\mathrm{div}\, v)\mathrm{I}$ denote the deviatoric part of $D(v)$, where I *is the identity matrix in $\mathbb{R}^{d\times d}$; then, there exists a positive constant c_p such that*

$$\|\nabla v\|_{\mathrm{L}^p(\Omega)} \leq c_p \|D^{\mathrm{dev}}(v)\|_{\mathrm{L}^p(\Omega)} \qquad \forall\, v \in [\mathrm{W}^{1,p}_{*}(\Omega)]^d.$$

Besides being dependent on p, the constant c_p also depends on d, but we do not explicitly indicate that. In each case, the left-hand side of the inequality can be further bounded below by $C_p\|v\|_{\mathrm{W}^{1,p}(\Omega)}$, where C_p is another positive constant dependent on p and d, but independent of v.

3.1 Construction of the numerical method

Let

$$\Sigma_N \subset \mathrm{H}_{*,\mathrm{sym}}(\mathrm{div};\Omega) := \{S \in [\mathrm{L}^2_{\#}(\Omega)]^{d\times d} : S = S^{\mathrm{T}},\ \mathrm{div}\, S \in [\mathrm{L}^2_{\#}(\Omega)]^d, \int_\Omega S(x)\,\mathrm{d}x = 0\},$$

equipped with norm $\|S\|_{\mathrm{H}(\mathrm{div};\Omega)} := \left(\|S\|^2_{\mathrm{L}^2(\Omega)} + \|\mathrm{div}\, S\|^2_{\mathrm{L}^2(\Omega)}\right)^{\frac{1}{2}}$, and

$$V_N \subset [\mathrm{W}^{1,2}_{*}(\Omega)]^d := \left\{v \in [\mathrm{W}^{1,2}_{\#}(\Omega)]^d : \int_\Omega v(x)\,\mathrm{d}x = 0\right\}$$

be a pair of finite-dimensional spaces consisting of, respectively, $\mathbb{R}^{d\times d}$-valued and $\mathbb{R}^d$-valued functions, whose components are 2π-periodic real-valued trigonometric polynomials of degree N, $N \geq 1$, in each of the d coordinate directions, whose integral

over Ω is equal to 0. We note that the above definition of Σ_N is slightly different from the one in [4], where, instead, Σ_N was taken to be a subset of $\mathrm{H}_{\#,\mathrm{sym}}(\mathrm{div};\Omega)$, without requiring that its elements have zero integral over Ω; having said this, all the results proved in [4] continue to hold if we assume, as we have done above, that elements of the spaces concerned integrate to zero over the domain Ω.

The pair of spaces (Σ_N, V_N) satisfies the following inf-sup condition: let $b(v,T) := -(v, \mathrm{div}\, T)$; then, there exists a positive constant $c_{\text{inf-sup}}$, independent of N, such that

$$\inf_{v_N \in V_N \setminus \{0\}} \sup_{T_N \in \Sigma_N \setminus \{0\}} \frac{b(v_N, T_N)}{\|v_N\|_{\mathrm{L}^2(\Omega)} \|T_N\|_{H(\mathrm{div};\Omega)}} \geq c_{\text{inf-sup}}. \tag{3}$$

For a short proof of (3) we refer to the Appendix in [4], where it is shown that $c_{\text{inf-sup}} \geq \frac{1}{3}$.

Suppose that $f \in [\mathrm{L}^1_*(\Omega)]^d$; in order to avoid trivialities, it will be assumed throughout that $f \neq 0$ (and therefore $S \neq 0$). We consider the following discrete problem: find $(S_N, u_N) \in \Sigma_N \times V_N$ such that

$$-(\mathrm{div}\, S_N, v_N) = (f, v_N) \qquad \forall\, v_N \in V_N, \tag{4}$$

$$\hat{D}_N := F(S_N), \tag{5}$$

$$(D(u_N), T_N) = (\hat{D}_N, T_N) \qquad \forall\, T_N \in \Sigma_N. \tag{6}$$

We are now ready to embark on the proof of existence and uniqueness of a solution to the discrete problem (4)–(6).

3.2 Existence and uniqueness of solutions to the numerical method

Theorem 3.3 below, guaranteeing the existence and uniqueness of a solution to the discrete problem (4)–(6), was established in [4]; for the sake of completeness of our analysis of the discretization, and for the convenience of the reader, we include its proof here. It relies on the following corollary of Brouwer's fixed point theorem (cf. Girault & Raviart [7, Corollary 1.1, p.279]).

LEMMA 3.2. *Let $\mathcal{H}$ be a finite-dimensional Hilbert space whose inner product is denoted by $(\cdot,\cdot)_{\mathcal{H}}$ and the corresponding norm by $\|\cdot\|_{\mathcal{H}}$. Let $\mathfrak{F}$ be a continuous mapping from $\mathcal{H}$ into $\mathcal{H}$ with the following property: there exists a $\mu > 0$ such that $(\mathfrak{F}(v), v)_{\mathcal{H}} > 0$ for all $v \in \mathcal{H}$ with $\|v\|_{\mathcal{H}} = \mu$. Then, there exists an element $u \in \mathcal{H}$ such that $\|u\|_{\mathcal{H}} \leq \mu$ and $\mathfrak{F}(u) = 0$.*

THEOREM 3.3. *Suppose that $f \in [\mathrm{L}^1_{\#}(\Omega)]^d$ and $N \geq 1$. Then, the discrete problem (4)–(6) has a unique solution $(S_N, u_N) \in \Sigma_N \times V_N$.*

Proof. Assuming for the moment the existence of a solution $(S_N, u_N) \in \Sigma_N \times V_N$ to (4)–(6), we shall show that the solution must be unique. Suppose otherwise, that there exist $(S_N^i, u_N^i) \in \Sigma_N \times V_N$ that solve (4)–(6) for $i = 1, 2$. Hence,

$$-(\mathrm{div}\,(S_N^1 - S_N^2), v_N) - (D(u_N^1 - u_N^2), T_N) + \big(F(S_N^1) - F(S_N^2), T_N\big) = 0$$

for all $(T_N, v_N) \in \Sigma_N \times V_N$. We take $T_N = S_N^1 - S_N^2$ and $v_N = u_N^1 - u_N^2$, and note that, after partial integration in the first term,

$$-(\mathrm{div}\,(S_N^1 - S_N^2), u_N^1 - u_N^2) - (D(u_N^1 - u_N^2), S_N^1 - S_N^2)$$

$$=(S_N^1 - S_N^2, \nabla(u_N^1 - u_N^2)) - (D(u_N^1 - u_N^2), S_N^1 - S_N^2)$$
$$=(S_N^1 - S_N^2, D(u_N^1 - u_N^2)) - (D(u_N^1 - u_N^2), S_N^1 - S_N^2) = 0,$$

where in the next to last equality we have used that S_N^1 and S_N^2 are symmetric $d \times d$ matrix functions, whereby the same is true of their difference. Consequently,

$$\big(F(S_N^1) - F(S_N^2), S_N^1 - S_N^2\big) = 0.$$

Property **(P2a)** then implies that $S_N^1 \equiv S_N^2$ on Ω, and hence $\hat{D}_N^1 \equiv \hat{D}_N^2$ on Ω, which yields that $D(u_N^1 - u_N^2) \equiv 0$ on Ω. By Korn's inequality stated in Lemma 3.1, we then have that $u_N^1 - u_N^2 \equiv 0$ on Ω, thus completing the proof of uniqueness of the solution to discrete problem (4)–(6).

Next we prove the existence of a solution to (4)–(6). First we choose any $\hat{S}_N \in \Sigma_N$ such that $-(\operatorname{div} \hat{S}_N, v_N)=(f, v_N)$ for all $v_N \in V_N$, and let $S_{N,0}:=S_N-\hat{S}_N$. The existence of such an $\hat{S}_N$ will be shown below; for the time being, we shall proceed by taking the existence of such an $\hat{S}_N$ for granted. Clearly, $-(\operatorname{div} S_{N,0}, v_N)=0$ for all $v_N \in V_N$, which then motivates us to define $\Sigma_{N,0}:=\{T_N \in \Sigma_N : -(\operatorname{div} T_N, v_N)=0 \text{ for all } v_N \in V_N\}$. As $0 \in \Sigma_{N,0}$, the set $\Sigma_{N,0}$ is nonempty. Problem (4)–(6) can be therefore restated in the following equivalent form: find $(S_{N,0}, u_N) \in \Sigma_{N,0} \times V_N$ such that

$$(D(u_N), T_N) = \Big(F(S_{N,0} + \hat{S}_N), T_N\Big) \qquad \forall\, T_N \in \Sigma_N. \tag{7}$$

Now, for $T_N \in \Sigma_{N,0}$, $(D(v_N), T_N) = (\nabla v_N, T_N) = -(v_N, \operatorname{div} T_N) = -(\operatorname{div} T_N, v_N) = 0$ for all $v_N \in V_N$. Hence, (7) indicates that we should seek $S_{N,0} \in \Sigma_{N,0}$ such that

$$\Big(F(S_{N,0} + \hat{S}_N), T_N\Big) = 0 \qquad \forall\, T_N \in \Sigma_{N,0}. \tag{8}$$

Let us consider the nonlinear operator $\mathfrak{F} : \Sigma_{N,0} \to \Sigma_{N,0}$, defined on the finite-dimensional Hilbert space $\Sigma_{N,0}$, equipped with the inner product and norm of $[\mathrm{L}^2_{\#}(\Omega)]^{d\times d}$, by $\mathfrak{F}(U_N) := P_N F(U_N+\hat{S}_N),\ U_N \in \Sigma_{N,0}$, where P_N denotes the orthogonal projector in $[\mathrm{L}^2_{\#}(\Omega)]^{d\times d}$ onto $\Sigma_{N,0}$.

Thanks to property **(P2b)**, we then have that

$$\|\mathfrak{F}(U_N^1) - \mathfrak{F}(U_N^2)\|_{\mathrm{L}^2(\Omega)} \le c_b \|U_N^1 - U_N^2\|_{\mathrm{L}^2(\Omega)} \qquad \forall\, U_N^1, U_N^2 \in \Sigma_{N,0},$$

and therefore $\mathfrak{F} : \Sigma_{N,0} \to \Sigma_{N,0}$ is (globally) Lipschitz continuous on $\Sigma_{N,0}$.

Note further that, by **(P2a)** and **(P1)**,

$$\begin{aligned}
(\mathfrak{F}(U_N), U_N) &= \Big(F(U_N+\hat{S}_N), U_N\Big) \\
&= \Big(F(U_N+\hat{S}_N), U_N + \hat{S}_N\Big) - \Big(F(U_N+\hat{S}_N), \hat{S}_N\Big) \\
&\ge c_a \int_\Omega \frac{|U_N + \hat{S}_N|^2}{1 + |U_N + \hat{S}_N|}\,\mathrm{d}x - \|\hat{S}_N\|_{\mathrm{L}^1(\Omega)} \\
&\ge \frac{1}{2} c_a \int_\Omega \frac{|U_N|^2}{1 + |U_N + \hat{S}_N|}\,\mathrm{d}x - c_a \int_\Omega \frac{|\hat{S}_N|^2}{1 + |U_N + \hat{S}_N|}\,\mathrm{d}x - \|\hat{S}_N\|_{\mathrm{L}^1(\Omega)} \\
&\ge \frac{1}{2} c_a \int_\Omega \frac{|U_N|^2}{1 + |U_N + \hat{S}_N|}\,\mathrm{d}x - c_a \|\hat{S}_N\|^2_{\mathrm{L}^2(\Omega)} - \|\hat{S}_N\|_{\mathrm{L}^1(\Omega)} \quad \forall\, U_N \in \Sigma_{N,0}.
\end{aligned}$$

As $|U_N + \hat{S}_N| \leq |U_N| + |\hat{S}_N| \leq \|U_N\|_{\mathrm{L}^\infty(\Omega)} + \|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)}$, it follows by the Nikol'skiĭ inequality $\|U_N\|_{\mathrm{L}^\infty(\Omega)} \leq C_{\mathrm{inv}} N^{\frac{d}{2}} \|U_N\|_{\mathrm{L}^2(\Omega)}$ that for any $U_N \in \Sigma_{N,0}$ such that $\|U_N\|_{\mathrm{L}^2(\Omega)} = \mu > 0$, we have that

$$(\mathfrak{F}(U_N), U_N) \geq \frac{c_a \mu^2}{2(1 + C_{\mathrm{inv}} N^{\frac{d}{2}} \mu + \|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)})} - |\Omega| \, \|\hat{S}_N\|^2_{\mathrm{L}^\infty(\Omega)} - |\Omega| \, \|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)}.$$

For $N \geq 1$ fixed (and therefore $\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)}$ also fixed), the expression on the right-hand side of the last displayed inequality is a continuous function of $\mu \in (0, \infty)$, which converges to $+\infty$ as $\mu \to +\infty$; thus, there exists a $\mu_0 = \mu_0(d, N, \|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)})$, such that $(\mathfrak{F}(U_N), U_N) > 0$ for all $U_N \in \Sigma_{N,0}$ satisfying $\|U_N\|_{\mathrm{L}^2(\Omega)} = \mu$, for $\mu > \mu_0$.

By taking $\mathcal{H} = \Sigma_{N,0}$, equipped with the inner product and norm of $[\mathrm{L}^2_{\#}(\Omega)]^{d\times d}$, we deduce from Lemma 3.2 the existence of an $S_{N,0} \in \Sigma_{N,0}$ that solves (8), and thus, recalling that $S_N = S_{N,0} + \hat{S}_N$, we have also shown the existence of an $S_N \in \Sigma_N$ such that $-(\operatorname{div} S_N, v_N) = (f, v_N)$ for all $v_N \in V_N$.

Having shown the existence of $S_N \in \Sigma_N$, we return to (7) in order to show the existence of a $u_N \in V_N$ such that $(D(u_N), T_N) = (F(S_N), T_N) \quad \forall\, T_N \in \Sigma_N$. Equivalently, we wish to show the existence of a $u_N \in V_N$ such that

$$b(u_N, T_N) = \ell(T_N) \qquad \forall\, T_N \in \Sigma_N, \tag{9}$$

where $b(v_N, T_N) := -(v_N, \operatorname{div} T_N)$ and $\ell(T_N) := (F(S_N), T_N)$. We note that $\ell(T_N) = 0$ for all $T_N \in \Sigma_{N,0}$, i.e., $\ell \in (\Sigma_{N,0})^0$ (the annihilator of $\Sigma_{N,0}$).

The existence of a unique $u_N \in V_N$ satisfying (9) then follows, thanks to the inf-sup condition (3), from the fundamental theorem of the theory of mixed variational problems stated in [6, Lemma 4.1(ii) on p.40].

At the very beginning of our proof of existence of solutions we postulated the existence of an $\hat{S}_N \in \Sigma_N$ such that $-(\operatorname{div} \hat{S}_N, v_N) = (f, v_N)$ for all $v_N \in V_N$. Again thanks to the inf-sup condition (3), [6, Lemma 4.1 (iii) on p.40] implies the existence of an $\hat{S}_N \in \Sigma_N$ such that $b(v_N, \hat{S}_N) = (f, v_N)$ for all $v_N \in V_N$; i.e., $-(\operatorname{div} \hat{S}_N, v_N) = (f, v_N)$ for all $v_N \in V_N$. Thus we have proved both the existence and the uniqueness of solutions to the discrete problem (4)–(6). □

Remark 3.4. The statement in the final paragraph of the proof above, that $\hat{S}_N \in \Sigma_N$, can be refined: in fact, $\hat{S}_N \in \Sigma^{\perp}_{N,0}$, where $\Sigma^{\perp}_{N,0}$ is the orthogonal complement of $\Sigma_{N,0}$ in Σ_N with respect to the $[\mathrm{L}^2_{\#}(\Omega)]^{d\times d}$ inner product.

The regularity hypothesis, that $f \in [\mathrm{L}^1_{\#}(\Omega)]^d$, is only used in the final paragraph of the proof. We note in particular that in order to apply [6, Lemma 4.1 (iii) on p.40], it is not necessary to demand that $f \in [\mathrm{L}^2_{\#}(\Omega)]^d$. Indeed, the Nikol'skiĭ inequality $\|v_N\|_{\mathrm{L}^\infty(\Omega)} \leq C_{\mathrm{inv}} N^{\frac{d}{2}} \|v_N\|_{\mathrm{L}^2(\Omega)}$ for any $v_N \in V_N$, implies that $|(f, v_N)| \leq C_{\mathrm{inv}} N^{\frac{d}{2}} \|f\|_{\mathrm{L}^1(\Omega)} \|v_N\|_{\mathrm{L}^2(\Omega)}$, and hence $v_N \mapsto (f, v_N)$ is a bounded linear functional on (the Hilbert space) V_N, equipped with the $[\mathrm{L}^2_{\#}(\Omega)]^d$ norm, as is required in [6, Lemma 4.1 (iii) on p.40].

3.3 Convergence of the sequence of numerical solutions

Next we will address the question of convergence of the sequence of approximate solutions generated by (4)–(6). To this end, we define the function space

$$D^{1,\infty}_*(\Omega) := \left\{ w \in [\mathrm{L}^1_\#(\Omega)]^d \,:\, D(w) \in [\mathrm{L}^\infty_\#(\Omega)]^{d\times d},\ \int_\Omega w(x)\,\mathrm{d}x = 0 \right\}.$$

Trivially, $V_N \subset D^{1,\infty}_*(\Omega)$ for each $N \geq 1$. As, by Hölder's inequality, $\|D(w)\|_{\mathrm{L}^p(\Omega)} < \infty$ for any $w \in D^{1,\infty}_*(\Omega)$ and any $p \in [1,\infty)$, Korn's inequality (cf. Lemma 3.1) implies that the seminorm $w \in D^{1,\infty}_*(\Omega) \mapsto \|D(w)\|_{\mathrm{L}^\infty(\Omega)}$ is in fact a norm on $D^{1,\infty}_*(\Omega)$. Furthermore (cf. [4]), $[\mathrm{C}^\infty_*(\overline{\Omega})]^d$ is weak-$*$ dense in $D^{1,\infty}_*(\Omega)$ against $[\mathrm{L}^1_*(\Omega)]^{d\times d}$, in the sense that for each $v \in D^{1,\infty}_*(\Omega)$ there exists a sequence $\{v_n\}_{n\geq 1} \subset [\mathrm{C}^\infty_*(\overline{\Omega})]^d$ such that $\int_\Omega T(x) : D(v_n(x))\,\mathrm{d}x \overset{n\to+\infty}{\to} \int_\Omega T(x) : D(v(x))\,\mathrm{d}x \quad \forall\, T \in [\mathrm{L}^1_*(\Omega)]^{d\times d}$.

We recall the following result from [4] concerning the convergence of the sequence of approximate solutions generated by (4)–(6) to a weak solution of the boundary-value problem.

THEOREM 3.5. *Suppose that $f \in [\mathrm{W}^{1,t}_\#(\Omega)]^d$ for some $t > 1$; then, there exists a unique pair $(S,u) \in [\mathrm{L}^1_*(\Omega)]^{d\times d} \times D^{1,\infty}_*(\Omega)$, such that*

$$(S, D(v)) = (f, v) \qquad \forall\, v \in D^{1,\infty}_*(\Omega),$$

and

$$D(u) = F(S) \quad \text{with} \begin{cases} r \in (0,1] & \text{if } d = 2, \\ r \in \left(0, \frac{2}{d}\right) & \text{if } d > 2. \end{cases}$$

Furthermore, the sequence of (uniquely defined) solution pairs $(S_N, u_N) \in \Sigma_N \times V_N$, $N \geq 1$, generated by (4)–(6), converges to (S,u) in the following sense:

(a) *The sequence $\{u_N\}_{N\geq 1}$ converges to u strongly in $[\mathrm{L}^p_*(\Omega)]^d$ and weakly in $[\mathrm{W}^{1,p}_*(\Omega)]^d$ for all $p \in [1,\infty)$;*

(b) *The sequence $\{D(u_N)\}_{N\geq 1}$ converges to $D(u)$ weakly in $[\mathrm{L}^p_*(\Omega)]^{d\times d}$ for all $p \in [1,\infty)$;*

(c) *The sequence $\{S_N\}_{N\geq 1}$ converges to S strongly in $[\mathrm{L}^s_*(\Omega)]^{d\times d}$ for all values of s in the range $[1, \frac{d(1-r)}{d-2})$ for $r \in (0, \frac{2}{d})$ when $d > 2$, and for $r \in (0,1]$ when $d = 2$;*

(d) *The sequence $\{D(u_N)\}_{N\geq 1}$ converges to $D(u)$ weakly in $[\mathrm{W}^{1,2}_*(\Omega)]^{d\times d}$, and therefore also strongly in $[\mathrm{L}^p_*(\Omega)]^{d\times d}$ for all $p \in [1, \frac{2d}{d-2})$, $d \geq 2$;*

(e) *If $r \in (0, \frac{1}{d-1})$, $d \geq 2$, then the sequence $\{S_N\}_{N\geq 1}$ converges to S weakly in $[\mathrm{W}^{1,\theta}_*(\Omega)]^{d\times d}$ for all $\theta \in [1, \frac{d(1-r)}{d-r-1})$.*

It is further shown in [4] that the boundary-value problem under consideration has a renormalized solution (S,u) for all $r > 0$, which, if $S \in [\mathrm{W}^{1,1}_*(\Omega)]^{d\times d}$ or $S \in [\mathrm{L}^{r+1}_*(\Omega)]^{d\times d}$, coincides with the unique weak solution to the problem (cf. [4, Theorem 5.1]) for any $r > 0$.

In the next section, assuming additional regularity of the solution (S,u), we derive an optimal bound in the L^2 norm on the error between $(S, D(u))$ and its numerical approximation $(S_N, D(u_N))$.

4. Error analysis of the numerical method

The proof of the next theorem will rely on the following classical approximation result (cf., for example, Theorem 1.1 in [5]): suppose that $T \in [\mathrm{H}^s_*(\Omega)]^{d\times d}$; then, there exists a positive constant $c_1 = c_1(s,d)$, independent of N, such that

$$\|T - P_N T\|_{\mathrm{H}^{s'}(\Omega)} \le c_1 N^{s'-s} \|T\|_{\mathrm{H}^s(\Omega)} \qquad \forall\, N \ge 1, \tag{10}$$

where $0 \le s' \le s$.

THEOREM 4.1. *Suppose that $(S,u) \in [\mathrm{H}^s_*(\Omega)]^{d\times d} \times [\mathrm{H}^{s+1}_*(\Omega)]^d$, where $s > \frac{d}{2}$. Then, there exists a positive constant c_*, independent of N, and a positive integer N_* such that*

$$\|S - S_N\|_{\mathrm{L}^2(\Omega)} \le (c_1 + c_*) N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \right) \ \forall\, N \ge N_*, \tag{11}$$

$$\|D(u) - D(u_N)\|_{\mathrm{L}^2(\Omega)} \le c_b (c_1 + c_*) N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \right) \ \forall\, N \ge N_*. \tag{12}$$

Proof. We begin by rewriting (4)–(6) in the following form: find $(S_N, u_N) \in \Sigma_N \times V_N$ such that

$$-(\operatorname{div} S_N, v_N) = (f, v_N) \qquad \forall\, v_N \in V_N, \tag{13}$$

$$(F(S_N), T_N) - (D(u_N), T_N) = 0 \qquad \forall\, T_N \in \Sigma_N. \tag{14}$$

Consider $\hat{S}_N := P_N S$, the orthogonal projection in $[\mathrm{L}^2_{\#}(\Omega)]^{d\times d}$ of S onto Σ_N. Clearly,

$$\begin{aligned} -(\operatorname{div} \hat{S}_N, v_N) &= -(\operatorname{div} P_N S, v_N) = (P_N S, \nabla v_N) = (P_N S, D(v_N)) \\ &= (S, D(v_N)) = (S, \nabla v_N) = -(\operatorname{div} S, v_N) = (f, v_N) \qquad \forall\, v_N \in V_N. \end{aligned}$$

Thus, by letting $S_{N,0} := S_N - \hat{S}_N$, we deduce that

$$S_{N,0} \in \Sigma_{N,0} := \{T_N \in \Sigma_N \,:\, (\operatorname{div} T_N, v_N) = 0 \quad \forall\, v_N \in V_N\}.$$

It follows that (13), (14) can be rewritten in the following form:

$$(\operatorname{div} S_{N,0}, v_N) = 0 \qquad \forall\, v_N \in V_N,$$

$$(F(S_{N,0} + \hat{S}_N), T_N) + (u_N, \operatorname{div} T_N) = 0 \qquad \forall\, T_N \in \Sigma_N.$$

Hence, in particular,

$$(F(S_{N,0} + \hat{S}_N), T_N) = 0 \qquad \forall\, T_N \in \Sigma_{N,0}, \tag{15}$$

and therefore

$$\begin{aligned} &(F(S_{N,0} + \hat{S}_N) - F(\hat{S}_N), T_N) = -(F(\hat{S}_N), T_N) \\ =&(F(S) - F(\hat{S}_N), T_N) - (F(S), T_N) = (F(S) - F(\hat{S}_N), T_N) - (D(u), T_N) \\ =&(F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) - (P_N D(u), T_N) \\ =&(F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) - (D(Q_N u), T_N) \\ =&(F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) + (Q_N u, \operatorname{div} T_N) \\ =&(F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) \qquad \forall\, T_N \in \Sigma_{N,0}, \end{aligned} \tag{16}$$

where Q_N is the orthogonal projector in $[\mathrm{L}^2_{\#}(\Omega)]^d$ onto the linear subspace V_N; here we have used that $(P_N D(u), T_N) = (D(Q_N u), T_N)$ for all $T_N \in \Sigma_{N,0}$, because

$(P_N D(v), T_N) = (D(v), T_N) = -(v, \operatorname{div} T_N) = -(Q_N v, \operatorname{div} T_N) = (D(Q_N v), T_N)$ for any $v \in [\mathrm{W}^{1,2}_*(\Omega)]^d = [\mathrm{H}^1_*(\Omega)]^d$ and any $T_N \in \Sigma_{N,0}$.

Now, for S and u fixed, consider the linear functional $\ell : \Sigma_N \to \mathbb{R}$, defined by $\ell(T_N) := (F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N), \qquad T_N \in \Sigma_N$. We then deduce from (16) that

$$(F(S_{N,0} + \hat{S}_N) - F(\hat{S}_N), T_N) = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}. \tag{17}$$

Thanks to **(P2b)** and (10), we have that

$$\begin{aligned} |\ell(T_N)| &\le \left(c_b \|S - \hat{S}_N\|_{\mathrm{L}^2(\Omega)} + \|D(u) - P_N D(u)\|_{\mathrm{L}^2(\Omega)} \right) \|T_N\|_{\mathrm{L}^2(\Omega)} \\ &\le \left(c_b c_1 N^{-s} \|S\|_{\mathrm{H}^s(\Omega)} + c_1 N^{-s} \|D(u)\|_{\mathrm{H}^s(\Omega)} \right) \|T_N\|_{\mathrm{L}^2(\Omega)} \\ &\le c_b c_1 N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \right) \|T_N\|_{\mathrm{L}^2(\Omega)} \qquad \forall\, T_N \in \Sigma_N, \end{aligned} \tag{18}$$

because $c_b \ge 1$.

Our objective is to prove that there exist $c_* > 0$, independent of N, and $N_* \in \mathbb{N}$, such that for each $N \ge N_*$ there exists a unique $S_{N,0} \in \Sigma_{N,0}$ such that (17) holds and $\|S_{N,0}\|_{\mathrm{L}^2(\Omega)} \le c_* N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \right)$. We shall use a fixed point theorem to this end. In order to define the fixed point mapping, we begin by noting that, by [4, Lemma 3.2], $(F(A) - F(B)) : C = \int_0^1 G(\theta A + (1-\theta)B; A - B, C)\, \mathrm{d}\theta$, where, for $\alpha, \beta, \gamma \in \mathbb{R}^{d \times d}$,

$$G(\gamma; \alpha, \beta) := \frac{\alpha : \beta}{(1 + |\gamma|^r)^{\frac{1}{r}}} - (\alpha : \gamma)(\beta : \gamma) \frac{|\gamma|^{r-2}}{(1 + |\gamma|^r)^{1 + \frac{1}{r}}}.$$

Note that

$$|G(\gamma; \alpha, \beta)| \le \frac{2\,|\alpha|\,|\beta|}{(1 + |\gamma|^r)^{\frac{1}{r}}} \qquad \forall\, \alpha, \beta, \gamma \in \mathbb{R}^{d \times d}_{\mathrm{sym}}, \tag{19}$$

$$G(\gamma; \alpha, \alpha) \ge \frac{|\alpha|^2}{(1 + |\gamma|^r)^{1 + \frac{1}{r}}} \qquad \forall\, \alpha, \gamma \in \mathbb{R}^{d \times d}_{\mathrm{sym}}. \tag{20}$$

We define the set

$$\mathfrak{B}_{N,0} := \left\{ T_N \in \Sigma_{N,0} \,:\, \|T_N\|_{\mathrm{L}^2(\Omega)} \le c_* N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \right) \right\}.$$

As $0 \in \mathfrak{B}_{N,0}$, the set $\mathfrak{B}_{N,0}$ is nonempty, regardless of the choice of $c_* > 0$; also, $\mathfrak{B}_{N,0}$ is a closed subset of the finite-dimensional linear space $\Sigma_{N,0}$.

Let us rewrite (17) as follows: find $S_{N,0} \in \Sigma_{N,0}$ such that

$$\int_\Omega \int_0^1 G(\theta(S_{N,0} + \hat{S}_N) + (1-\theta)\hat{S}_N; S_{N,0}, T_N)\, \mathrm{d}\theta\, \mathrm{d}x = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}.$$

Equivalently, we can write this as follows: find $S_{N,0} \in \Sigma_{N,0}$ such that

$$\int_\Omega \int_0^1 G(\hat{S}_N + \theta S_{N,0}; S_{N,0}, T_N)\, \mathrm{d}\theta\, \mathrm{d}x = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}.$$

Motivated by this equivalent restatement of (17), we consider the following mapping: to each $\varphi \in \mathfrak{B}_{N,0}$ we assign $S_{N,\varphi} \in \Sigma_{N,0}$ such that

$$\int_\Omega \int_0^1 G(\hat{S}_N + \theta\varphi; S_{N,\varphi}, T_N)\, \mathrm{d}\theta\, \mathrm{d}x = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}. \tag{21}$$

It follows from (20) that, for $\hat{S}_N \in \Sigma_N$ and $\varphi \in \mathfrak{B}_{N,0}$ fixed, (21) has at most one solution $S_{N,\varphi} \in \Sigma_{N,0}$. Since $\Sigma_{N,0}$ is a finite-dimensional linear space and (21) is a linear problem, the uniqueness of the solution implies its existence. Thus we deduce that the mapping $\varphi \in \mathfrak{B}_{N,0} \mapsto S_{N,\varphi} \in \Sigma_{N,0}$ is correctly defined. Next we will show that there exists a constant $c_* > 0$, independent of N, and $N_* \in \mathbb{N}$, such that if $\varphi \in \mathfrak{B}_{N,0}$ with $N \geq N_*$, then $S_{N,\varphi} \in \mathfrak{B}_{N,0}$, in fact.

Note that by (20), (21) and (18),

$$\begin{aligned}\frac{\|S_{N,\varphi}\|^2_{\mathrm{L}^2(\Omega)}}{(1+(\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)}+\|\varphi\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}}} &\leq \int_\Omega \int_0^1 \frac{|S_{N,\varphi}|^2}{(1+|\hat{S}_N+\theta\varphi|^r)^{1+\frac{1}{r}}}\,\mathrm{d}\theta\,\mathrm{d}x \\ &\leq \int_\Omega \int_0^1 G(\hat{S}_N+\theta\varphi; S_{N,\varphi}, S_{N,\varphi})\,\mathrm{d}\theta\,\mathrm{d}x = \ell(S_{N,\varphi}) \\ &\leq c_b c_1 N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right)\|S_{N,\varphi}\|_{\mathrm{L}^2(\Omega)}.\end{aligned}$$

Thus we deduce that

$$\begin{aligned}&\|S_{N,\varphi}\|_{\mathrm{L}^2(\Omega)} \\ &\leq c_b c_1 N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right)(1 + (\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} + \|\varphi\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}}. \qquad (22)\end{aligned}$$

In order to prove that $S_{N,\varphi} \in \mathfrak{B}_{N,0}$ for a suitable $c_* > 0$ and all $N \geq N_*$, with a certain positive integer N_*, our aim is to show that, for a suitable constant $c_* > 0$, independent of N, and a suitable positive integer N_*,

$$\begin{aligned}&c_b c_1 N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right)(1 + (\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} + \|\varphi\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}} \\ &\qquad \leq c_* N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right) \qquad \forall\, N \geq N_*. \qquad (23)\end{aligned}$$

This is equivalent to showing that, for a suitable constant $c_* > 0$, independent of N, and a suitable positive integer N_*,

$$c_b c_1 (1 + (\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} + \|\varphi\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}} \leq c_* \qquad \forall\, N \geq N_*. \qquad (24)$$

We shall derive a sufficient condition for (24) to hold by replacing $\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)}$ and $\|\varphi\|_{\mathrm{L}^\infty(\Omega)}$ in (24) by upper bounds on them.

First note that $\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} = \|P_N S\|_{\mathrm{L}^\infty(\Omega)} \leq \|S\|_{\mathrm{L}^\infty(\Omega)} + \|S - P_N S\|_{\mathrm{L}^\infty(\Omega)}$. As, by hypothesis, $s > \frac{d}{2}$, there exists an $s' \in \left(\frac{d}{2}, s\right)$. By Sobolev embedding, and using the approximation property (10) of the projector P_N, we have that

$$\|S - P_N S\|_{\mathrm{L}^\infty(\Omega)} \leq C(s', d)\|S - P_N S\|_{\mathrm{H}^{s'}(\Omega)} \leq c_1 C(s', d) N^{s'-s}\|S\|_{\mathrm{H}^s(\Omega)}.$$

As $s > s'$, there exists a positive integer N_{**} such that

$$c_1 C(s', d) N^{s'-s}\|S\|_{\mathrm{H}^s(\Omega)} \leq \|S\|_{\mathrm{L}^\infty(\Omega)} \qquad \forall\, N \geq N_{**}.$$

For example, we can take

$$N_{**} := \left\lceil \left(\frac{c_1 C(s', d)\|S\|_{\mathrm{H}^s(\Omega)}}{\|S\|_{\mathrm{L}^\infty(\Omega)}} \right)^{\frac{1}{s-s'}} \right\rceil.$$

Hence, $\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} \leq 2\|S\|_{\mathrm{L}^\infty(\Omega)} \quad \forall\, N \geq N_{**}$. Since by the Nikol'skiĭ inequality $\|T_N\|_{\mathrm{L}^\infty(\Omega)} \leq C_{\mathrm{inv}} N^{\frac{d}{2}}\|T_N\|_{\mathrm{L}^2(\Omega)}$ for any $T_N \in \Sigma_{N,0}$, it follows that a sufficient condi-

tion for (24) to hold is that

$$c_b c_1 (1 + (2\|S\|_{\mathrm{L}^\infty(\Omega)} + C_{\mathrm{inv}} N^{\frac{d}{2}} \|\varphi\|_{\mathrm{L}^2(\Omega)})^r)^{1+\frac{1}{r}} \leq c_* \qquad \forall\, N \geq N_*, \tag{25}$$

where $N_* \geq N_{**}$ is a positive integer, to be chosen below.

We define $c_* := c_b c_1 \big(1 + \big(2\|S\|_{\mathrm{L}^\infty(\Omega)} + C_{\mathrm{inv}}(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)})\big)^r\big)^{1+\frac{1}{r}}$. With this definition of c_*, (25) becomes equivalent to the inequality

$$N^{\frac{d}{2}} \|\varphi\|_{\mathrm{L}^2(\Omega)} \leq \|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \qquad \forall\, N \geq N_*. \tag{26}$$

As $\varphi \in \mathfrak{B}_{N,0}$, a sufficient condition for (26) to hold is that

$$c_* N^{\frac{d}{2}-s} \leq 1 \qquad \forall\, N \geq N_*. \tag{27}$$

Since $s > \frac{d}{2}$, there exists an $N_* \geq N_{**}$ such that this inequality holds; for example, one can take

$$N_* := \max\left(\lceil c_*^{\frac{2}{2s-d}} \rceil, N_{**} \right).$$

With c_* and N_* thus defined, (27) holds; and, therefore, (26), (25), (24) all hold, and, since (24) is equivalent to (23), it follows that (23) also holds. Having shown the existence of c_* and N_* such that (23) holds, it follows from (22) that

$$\|S_{N,\varphi}\|_{\mathrm{L}^2(\Omega)} \leq c_* N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \right) \qquad \forall\, N \geq N_*.$$

Hence, $S_{N,\varphi} \in \mathfrak{B}_{N,0}$ for all $N \geq N_*$. As the function $\varphi \mapsto S_{N,\varphi}$ maps the bounded closed ball $\mathfrak{B}_{N,0}$ contained in the finite-dimensional linear space $\Sigma_{N,0}$ into itself, Brouwer's fixed point theorem will imply the existence of a fixed point $S_{N,*} \in \mathfrak{B}_{N,0}$ for this mapping, once we have shown the continuity of this mapping.

To this end, we consider $\varphi_1, \varphi_2 \in \mathfrak{B}_{N,0}$ and the associated $S_{N,\varphi_1}, S_{N,\varphi_2} \in \mathfrak{B}_{N,0}$, $N \geq N_*$, defined, for $i = 1, 2$, by

$$\int_\Omega \int_0^1 G(\hat{S}_N + \theta\varphi_i; S_{N,\varphi_i}, T_N)\, \mathrm{d}\theta\, \mathrm{d}x = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}.$$

We thus have that

$$\begin{aligned} &\int_\Omega \int_0^1 G(\hat{S}_N + \theta\varphi_1; S_{N,\varphi_1} - S_{N,\varphi_2}, T_N)\, \mathrm{d}\theta\, \mathrm{d}x \\ &= \int_\Omega \int_0^1 G(\hat{S}_N + \theta\varphi_2; S_{N,\varphi_2}, T_N)\, \mathrm{d}\theta\, \mathrm{d}x - \int_\Omega \int_0^1 G(\hat{S}_N + \theta\varphi_1; S_{N,\varphi_2}, T_N)\, \mathrm{d}\theta\, \mathrm{d}x. \end{aligned}$$

By taking $T_N = S_{N,\varphi_1} - S_{N,\varphi_2}$ we deduce from Lemma 2.2 that

$$\begin{aligned} &\frac{\|S_{N,\varphi_1} - S_{N,\varphi_2}\|^2_{\mathrm{L}^2(\Omega)}}{(1 + (\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} + \|\varphi_1\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}}} \\ &\leq \int_\Omega \int_0^1 \Big| G(\hat{S}_N + \theta\varphi_2; S_{N,\varphi_2}, S_{N,\varphi_1} - S_{N,\varphi_2}) - G(\hat{S}_N + \theta\varphi_1; S_{N,\varphi_2}, S_{N,\varphi_1} - S_{N,\varphi_2}) \Big|\, \mathrm{d}\theta\, \mathrm{d}x. \end{aligned}$$

For $\alpha, \beta, \gamma \in \mathbb{R}^{d\times d}$, we choose $\varepsilon \in \big(\max\{0, 1 - \frac{r}{2}\}, 1\big)$ and rewrite $G(\gamma; \alpha, \beta)$ as follows:

$$G(\gamma; \alpha, \beta) := \frac{\alpha : \beta}{(1 + |\gamma|^r)^{\frac{1}{r}}} - \left(\alpha : \frac{\gamma}{|\gamma|^\varepsilon}\right)\left(\beta : \frac{\gamma}{|\gamma|^\varepsilon}\right) \frac{|\gamma|^{r-2+2\varepsilon}}{(1 + |\gamma|^r)^{1+\frac{1}{r}}}.$$

Note that with such an ε one has $r-2+2\varepsilon>0$. The functions

$$\gamma\mapsto\frac{1}{(1+|\gamma|^r)^{\frac{1}{r}}},\quad \gamma\mapsto\frac{\gamma}{|\gamma|^{\varepsilon}},\quad \gamma\mapsto|\gamma|^{r-2+2\varepsilon},\quad \gamma\mapsto\frac{1}{(1+|\gamma|^r)^{1+\frac{1}{r}}}$$

are Hölder-continuous on any bounded ball $\mathcal{B}(0,R)$ in $\mathbb{R}^{d\times d}$ of radius R; the Hölder exponents δ_i, $i=1,2,3,4$, of these four functions are, respectively, $\delta_1=\min(1,r)$, $\delta_2<1-\varepsilon$, $\delta_3=\min(1,r-2+2\varepsilon)$, $\delta_4=\min(1,r)$. These statements follow from Lemma 2.3, parts **(d)**; **(e)**; **(b)** and **(c)**; and **(d)**, respectively.

Let $\delta_0=\min(\delta_1,\delta_2,\delta_3,\delta_4)$; clearly, $\delta_0\in(0,1)$. Let $\delta\in(0,\delta_0]$. Hence,

$$\int_\Omega\int_0^1\left|G(\hat S_N+\theta\varphi_2;S_{N,\varphi_2},S_{N,\varphi_1}-S_{N,\varphi_2})-G(\hat S_N+\theta\varphi_1;S_{N,\varphi_2},S_{N,\varphi_1}-S_{N,\varphi_2})\right|\,\mathrm{d}\theta\,\mathrm{d}x$$
$$\le C(r,\varepsilon,\|S_{N,\varphi_2}\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_1\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_2\|_{\mathrm{L}^\infty(\Omega)})\int_\Omega|\varphi_1-\varphi_2|^\delta|S_{N,\varphi_1}-S_{N,\varphi_2}|\,\mathrm{d}x.$$
$$\le C(r,\varepsilon,\|S_{N,\varphi_2}\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_1\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_2\|_{\mathrm{L}^\infty(\Omega)})\,\|\varphi_1-\varphi_2\|^\delta_{L_{2\delta}(\Omega)}\,\|S_{N,\varphi_1}-S_{N,\varphi_2}\|_{\mathrm{L}^2(\Omega)}.$$

Thus we deduce that

$$\|S_{N,\varphi_1}-S_{N,\varphi_2}\|_{\mathrm{L}^2(\Omega)}$$
$$\le C(r,\varepsilon,\|\hat S_N\|_{\mathrm{L}^\infty(\Omega)},\|S_{N,\varphi_2}\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_1\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_2\|_{\mathrm{L}^\infty(\Omega)})\,\|\varphi_1-\varphi_2\|^\delta_{L_{2\delta}(\Omega)},$$

for all $\varphi_1,\varphi_2\in\mathfrak{B}_{N,0}$. As $\delta\in(0,1)$, it follows by Hölder's inequality that

$$\|S_{N,\varphi_1}-S_{N,\varphi_2}\|_{\mathrm{L}^2(\Omega)}$$
$$\le C(r,\varepsilon,\|\hat S_N\|_{\mathrm{L}^\infty(\Omega)},\|S_{N,\varphi_2}\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_1\|_{\mathrm{L}^\infty(\Omega)},\|\varphi_2\|_{\mathrm{L}^\infty(\Omega)})\,\|\varphi_1-\varphi_2\|^\delta_{L_2(\Omega)},$$

for all $\varphi_1,\varphi_2\in\mathfrak{B}_{N,0}$. We note that, for $N\ge N_*$, we have that

$$\|\hat S_N\|_{\mathrm{L}^\infty(\Omega)}\le 2\|S\|_{\mathrm{L}^\infty(\Omega)},$$
$$\|S_{N,\varphi_2}\|_{\mathrm{L}^\infty(\Omega)}\le C_{\mathrm{inv}}c_*N^{\frac{d}{2}-s}\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right),$$
$$\|\varphi_i\|_{\mathrm{L}^\infty(\Omega)}\le C_{\mathrm{inv}}c_*N^{\frac{d}{2}-s}\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right),\quad i=1,2.$$

Hence, for $(S,u)\in[\mathrm{H}^s_*(\Omega)]^{d\times d}\times[\mathrm{H}^{s+1}_*(\Omega)]^d$ fixed, with $s>\frac{d}{2}$,

$$\|S_{N,\varphi_1}-S_{N,\varphi_2}\|_{\mathrm{L}^2(\Omega)}\le C(r,\varepsilon)\,\|\varphi_1-\varphi_2\|^\delta_{L_2(\Omega)}\qquad\forall\,\varphi_1,\varphi_2\in\mathfrak{B}_{N,0},\ N\ge N_*.$$

This implies the (Hölder) continuity of the map $\varphi\in\mathfrak{B}_{N,0}\mapsto S_{N,\varphi}\in\mathfrak{B}_{N,0}$ for $N\ge N_*$. Hence, $\varphi\mapsto S_{N,\varphi}$ maps the bounded closed ball $\mathfrak{B}_{N,0}$ contained in the finite-dimensional linear space $\Sigma_{N,0}$ continuously into itself; Brouwer's fixed point theorem therefore implies the existence of a fixed point $S_{N,*}\in\mathfrak{B}_{N,0}$ for this mapping; i.e.,

$$\int_\Omega\int_0^1 G(\hat S_N+\theta S_{N,*};S_{N,*},T_N)\,\mathrm{d}\theta\,\mathrm{d}x=\ell(T_N)\qquad\forall\,T_N\in\Sigma_{N,0}.\tag{28}$$

Since the uniqueness of the fixed point is not guaranteed by Brouwer's fixed point theorem, it is not clear at this stage whether $S_{N,*}$ is equal to $S_{N,0}$. In order to show that this is the case, we proceed as follows. First note that (28) is equivalent to $(F(S_{N,*}+\hat S_N),T_N)=0\ \forall\,T_N\in\Sigma_{N,0}$. Recall from (15) that, on the other hand, $(F(S_{N,0}+\hat S_N),T_N)=0\ \forall\,T_N\in\Sigma_{N,0}$. It follows from the last two equations, and setting $T_N=(S_{N,*}+\hat S_N)-(S_{N,0}+\hat S_N)=S_{N,*}-S_{N,0}\in\Sigma_{N,0}$, that $(F(S_{N,*}+\hat S_N)-$

$F(S_{N,0}+\hat{S}_N),(S_{N,*}+\hat{S}_N)-(S_{N,0}+\hat{S}_N))=0$. By Lemma 2.2, with $A=S_{N,*}+\hat{S}_N$, $B=S_{N,0}+\hat{S}_N$, this then implies that

$$\int_\Omega \min\left(1,2^{r-\frac{1}{r}}\right)\frac{|S_{N,*}-S_{N,0}|^2}{1+|S_{N,*}+\hat{S}_N|+|S_{N,0}+\hat{S}|)^{r+1}}\,\mathrm{d}x\le 0.$$

Hence, $|S_{N,*}-S_{N,0}|^2=0$ a.e. on Ω, whereby $S_{N,*}=S_{N,0}$ a.e. on Ω. Since both $S_{N,*}$ and $S_{N,0}$ are trigonometric polynomials, it follows that $S_{N,*}(x)=S_{N,0}(x)$ for all $x\in\Omega$.

Thus we have finally shown that there exists a unique $S_{N,0}\in\mathfrak{B}_{N,0}$, with $S_{N,0}:=S_N-\hat{S}_N=S_N-P_NS$, such that (15) holds. Now, by the triangle inequality and (10), and because $S_{N,0}\in\mathfrak{B}_{N,0}$, we have that

$$\begin{aligned}\|S-S_N\|_{\mathrm{L}^2(\Omega)}&\le\|S-P_NS\|_{\mathrm{L}^2(\Omega)}+\|S_{N,0}\|_{\mathrm{L}^2(\Omega)}\\&\le c_1N^{-s}\|S\|_{\mathrm{H}^s(\Omega)}+c_*N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right)\qquad(29)\\&\le(c_1+c_*)N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right)\qquad\forall\,N\ge N_*.\end{aligned}$$

Further, by (14), **(P2b)** and (29), and noting that $P_ND(u)-D(u_N)\in\Sigma_N$, we have that, for all $N\ge N_*$,

$$\begin{aligned}\|P_ND(u)-D(u_N)\|_{\mathrm{L}^2(\Omega)}&=\sup_{T_N\in\Sigma_N\setminus\{0\}}\frac{(P_ND(u)-D(u_N),T_N)}{\|T_N\|_{\mathrm{L}^2(\Omega)}}\\=\sup_{T_N\in\Sigma_N\setminus\{0\}}\frac{(D(u)-D(u_N),T_N)}{\|T_N\|_{\mathrm{L}^2(\Omega)}}&=\sup_{T_N\in\Sigma_N\setminus\{0\}}\frac{(F(S)-F(S_N),T_N)}{\|T_N\|_{\mathrm{L}^2(\Omega)}}\\\le\|F(S)-F(S_N)\|_{\mathrm{L}^2(\Omega)}&\le c_b\|S-S_N\|_{\mathrm{L}^2(\Omega)}\\\le c_bc_1N^{-s}\|S\|_{\mathrm{H}^s(\Omega)}+c_bc_*N^{-s}&\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right).\qquad(30)\end{aligned}$$

From (30), by the triangle inequality and noting that $c_b\ge1$, it follows that, for all $N\ge N_*$,

$$\begin{aligned}\|D(u)-D(u_N)\|_{\mathrm{L}^2(\Omega)}&\le\|D(u)-P_ND(u)\|_{\mathrm{L}^2(\Omega)}+\|P_ND(u)-D(u_N)\|_{\mathrm{L}^2(\Omega)}\\&\le c_b(c_1+c_*)N^{-s}\|S\|_{\mathrm{H}^s(\Omega)}+(c_1+c_bc_*)N^{-s}\|D(u)\|_{\mathrm{H}^s(\Omega)}\\&\le c_b(c_1+c_*)N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right).\end{aligned}$$

That completes the proof. □

Remark 4.2. We note that by Korn's inequality (cf. Lemma 3.1),

$$\|u-u_N\|_{\mathrm{H}^1(\Omega)}\le\mathrm{const}\cdot N^{-s}\left(\|S\|_{\mathrm{H}^s(\Omega)}+\|D(u)\|_{\mathrm{H}^s(\Omega)}\right).$$

For each $N\ge1$, the numerical method (4)–(6) is a finite-dimensional system of nonlinear equations. In the next section we propose an iterative method for the solution of the discrete problem (4)–(6) and we establish its convergence, with N kept fixed.

5. Iterative solution of the finite-dimensional nonlinear system

We consider the following iterative method for the solution of (4)–(6): let $S_N^0:=0$; for $k=1,2,\ldots$, we define $(S_N^k,u_N^k)\in\Sigma_N\times V_N$ as the solution of the following problem

$$-(\operatorname{div} S_N^k, v_N) = (f, v_N) \qquad \forall\, v_N \in V_N, \tag{31}$$

$$(S_N^k, T_N) - \lambda(D(u_N^k), T_N) = (S_N^{k-1}, T_N) - \lambda(F(S_N^{k-1}), T_N) \qquad \forall\, T_N \in \Sigma_N, \tag{32}$$

where $\lambda > 0$ is a parameter, to be fixed below.

We begin by showing that this iteration is correctly defined, in the sense that, for each $k \in \mathbb{N}$, there exists a unique pair $(S_N^k, u_N^k) \in \Sigma_N \times V_N$ satisfying (31), (32). To this end, let $S_{N,0}^k := S_N^k - S_N^{k-1}$, and note that

$$(\operatorname{div} S_{N,0}^k, v_N) = 0 \qquad \forall\, v_N \in V_N, \tag{33}$$

$$(S_{N,0}^k, T_N) - \lambda(D(u_N^k), T_N) = -\lambda(F(S_N^{k-1}), T_N) \qquad \forall\, T_N \in \Sigma_N. \tag{34}$$

Hence, $S_{N,0}^k \in \Sigma_{N,0}$, and therefore, $(S_{N,0}^k, T_N) = -\lambda(F(S_N^{k-1}), T_N) \quad \forall\, T_N \in \Sigma_{N,0}$. Consequently, $S_{N,0}^k$ is uniquely defined as the orthogonal projection of $-\lambda F(S_N^{k-1})$ onto the finite-dimensional linear subspace $\Sigma_{N,0}$ of Σ_N, with respect to the inner product of $[\mathrm{L}^2_{\#}(\Omega)]^{d\times d}$, which then uniquely defines $S_N^k = S_N^{k-1} + S_{N,0}^k \in \Sigma_N$. For S_N^k thus fixed, we rewrite (32) as follows:

$$-(u_N^k, \operatorname{div} T_N) = \frac{1}{\lambda}(S_N^k - S_N^{k-1}, T_N) + (F(S_N^{k-1}), T_N) \qquad \forall\, T_N \in \Sigma_N.$$

By introducing the bilinear form $b(v, T) := -(v, \operatorname{div} T)$ on $V_N \times \Sigma_N$ and the linear functional $\ell(T) := \frac{1}{\lambda}(S_N^k - S_N^{k-1}, T) + (F(S_N^{k-1}), T)$ on Σ_N, the proof of existence of a unique solution u_N^k to the problem $b(u_N^k, T_N) = \ell(T_N)$ for all $T_N \in \Sigma_N$ proceeds analogously as in the case of problem (9): the bilinear form $b(\cdot,\cdot)$ satisfies the inf-sup condition (3), and the linear functional $\ell \in (\Sigma_{N,0})^0$ (the annihilator of $\Sigma_{N,0}$). The existence of a unique solution u_N^k satisfying $b(u_N^k, T_N) = \ell(T_N)$ for all $T_N \in \Sigma_N$ therefore follows from the fundamental theorem of the theory of mixed variational problems, stated in [6, Lemma 4.1(ii) on p.40].

Next, we will show that, for each fixed $N \geq 1$, $(S_N^k, u_N^k) \to (S_N, u_N)$ as $k \to +\infty$.

THEOREM 5.1. *Let* $c_a := \min(1, 2^{r-\frac{1}{r}})$, $c_\diamond := 1 + (2 + C_{\mathrm{inv}} N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}}) \|S_N\|_{\mathrm{L}^\infty(\Omega)}$, $c_0 := \frac{c_a}{c_\diamond}$, *and let* $\lambda \in \big(0, \frac{1}{2}c_0\big)$. *Then,* $L^2 := 1 - 2c_0\lambda + 4\lambda^2 \in (0,1)$, *and, for each* $k \geq 1$, $\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \leq L^{2k} \|S_N\|^2_{\mathrm{L}^2(\Omega)} \quad \forall\, k \geq 1$.

Proof. We subtract (31), (32) from (13), (14), respectively; hence,

$$(\operatorname{div}(S_N - S_N^k, v_N) = 0 \quad \forall\, v_N \in V_N, \tag{35}$$

$$(S_N - S_N^k, T_N) = (S_N - S_N^{k-1}, T_N) - \lambda(F(S_N) - F(S_N^{k-1}), T_N) + \lambda(D(u_N - u_N^k), T_N) \quad \forall\, T_N \in \Sigma_N. \tag{36}$$

Equation (35) implies that $S_N - S_N^k \in \Sigma_{N,0}$; thus, by taking $T_N = S_N - S_N^k$ in (36), we have that

$$\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} = (S_N - S_N^{k-1}, S_N - S_N^k) - \lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^k). \tag{37}$$

Next, we take $T_N = S_N - S_N^{k-1} \in \Sigma_{N,0}$ in (36); hence,

$$(S_N - S_N^k, S_N - S_N^{k-1}) = \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - \lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}). \tag{38}$$

Finally, we take $T_N = P_N(F(S_N) - F(S_N^{k-1}))$ in (36); thus,

$$\begin{aligned}
&(S_N - S_N^k, F(S_N) - F(S_N^{k-1})) = (S_N - S_N^k, P_N(F(S_N) - F(S_N^{k-1})))\\
&= (S_N - S_N^{k-1}, P_N(F(S_N) - F(S_N^{k-1}))) - \lambda(F(S_N) - F(S_N^{k-1}), P_N(F(S_N) - F(S_N^{k-1})))\\
&\quad + \lambda(D(u_N - u_N^k), P_N(F(S_N) - F(S_N^{k-1})))\\
&= (S_N - S_N^{k-1}, F(S_N) - F(S_N^{k-1})) - \lambda(F(S_N) - F(S_N^{k-1}), P_N(F(S_N) - F(S_N^{k-1})))\\
&\quad + \lambda(D(u_N - u_N^k), F(S_N) - F(S_N^{k-1})). \qquad (39)
\end{aligned}$$

Substitution of (38) and (39) into (37) yields

$$\begin{aligned}
\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} &= \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - \lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1})\\
&\quad - \lambda(S_N - S_N^{k-1}, F(S_N) - F(S_N^{k-1}))\\
&\quad + \lambda^2(F(S_N) - F(S_N^{k-1}), P_N(F(S_N) - F(S_N^{k-1})))\\
&\quad - \lambda^2(D(u_N - u_N^k), F(S_N) - F(S_N^{k-1})). \qquad (40)
\end{aligned}$$

We shall transform the final term in (40) by taking $T_N = D(u_N - u_N^k)$ in (36):

$$\begin{aligned}
\lambda\|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} &= (S_N - S_N^k, D(u_N - u_N^k)) - (S_N - S_N^{k-1}, D(u_N - u_N^k))\\
&\quad + \lambda(F(S_N) - F(S_N^{k-1}), D(u_N - u_N^k)). \qquad (41)
\end{aligned}$$

As the first two terms on the right-hand side of (41) are both equal to 0 and $\lambda > 0$, it follows that

$$(D(u_N - u_N^k), F(S_N) - F(S_N^{k-1})) = \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)}. \qquad (42)$$

Substituting (42) into (40), we arrive at the following identity:

$$\begin{aligned}
&\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2\|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)}\\
&= \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1})\\
&\quad + \lambda^2(F(S_N) - F(S_N^{k-1}), P_N(F(S_N) - F(S_N^{k-1}))). \qquad (43)
\end{aligned}$$

As $|F(A) - F(B)| \le 2|A - B|$ (cf. Lemma 2.2), it follows that

$$\begin{aligned}
&\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2\|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)}\\
&= \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1})\\
&\quad + \lambda^2\|F(S_N) - F(S_N^{k-1})\|_{\mathrm{L}^2(\Omega)}\|P_N(F(S_N) - F(S_N^{k-1}))\|_{\mathrm{L}^2(\Omega)}\\
&\le \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1})\\
&\quad + \lambda^2\|F(S_N) - F(S_N^{k-1})\|^2_{\mathrm{L}^2(\Omega)}\\
&\le (1 + 4\lambda^2)\|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}). \qquad (44)
\end{aligned}$$

We focus our attention on the second term on the right-hand side of (44).

Thanks to Lemma 2.2,

$$\begin{aligned}
&(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}) \ge c_a \int_\Omega \frac{|S_N - S_N^{k-1}|^2}{(1 + |S_N| + |S_N^{k-1}|)^{r+1}}\,\mathrm{d}x\\
&\ge \frac{c_a}{1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^{k-1}\|_{\mathrm{L}^\infty(\Omega)}}\,\|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)}, \qquad (45)
\end{aligned}$$

where $c_a = \min\left(1, 2^{r-\frac{1}{r}}\right)$. As $S_N^0 := 0$, there exists a positive constant $c_\diamond$, to be fixed below, independent of k (but possibly dependent on N), such that $1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^0\|_{\mathrm{L}^\infty(\Omega)} \le c_\diamond$. Suppose, for induction, that we have already shown that

$$1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^m\|_{\mathrm{L}^\infty(\Omega)} \le c_\diamond \qquad \forall\, m \in \{0, \ldots, k-1\}, \tag{46}$$

for some $k \ge 1$. It then follows from (45) and (46) that

$$(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}) \ge c_0 \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)},$$

with $c_0 := \frac{c_a}{c_\diamond}$. Substituting this into the right-hand side of (44) we deduce that

$$\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \le (1 - 2c_0\lambda + 4\lambda^2)\|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)}. \tag{47}$$

Let us choose $\lambda \in \left(0, \frac{1}{2}c_0\right)$. Then, $L^2 := 1 - 2c_0\lambda + 4\lambda^2 \in (0,1)$. Consequently, (47) yields

$$\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \le L^2 \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)}, \quad L \in (0,1). \tag{48}$$

In order to complete the inductive step, it remains to show that (46) holds for all $m \in \{0, \ldots, k\}$, $k \ge 1$. To this end, we note that (48) implies that

$$\|S_N - S_N^k\|_{\mathrm{L}^2(\Omega)} \le L^k \|S_N - S_N^0\|_{\mathrm{L}^2(\Omega)} = L^k \|S_N\|_{\mathrm{L}^2(\Omega)}. \tag{49}$$

Thus, by the Nikol'skiĭ inequality $\|T_N\|_{\mathrm{L}^\infty(\Omega)} \le C_{\mathrm{inv}} N^{\frac{d}{2}} \|T_N\|_{\mathrm{L}^2(\Omega)}$, $T_N \in \Sigma_N$, we have that

$$\|S_N^k\|_{\mathrm{L}^\infty(\Omega)} \le \|S_N - S_N^k\|_{\mathrm{L}^\infty(\Omega)} + \|S_N\|_{\mathrm{L}^\infty(\Omega)} \le C_{\mathrm{inv}} N^{\frac{d}{2}} L^k \|S_N\|_{\mathrm{L}^2(\Omega)} + \|S_N\|_{\mathrm{L}^\infty(\Omega)}. \tag{50}$$

Hence,

$$\begin{aligned} 1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^k\|_{\mathrm{L}^\infty(\Omega)} &\le 1 + 2\|S_N\|_{\mathrm{L}^\infty(\Omega)} + C_{\mathrm{inv}} N^{\frac{d}{2}} L^k \|S_N\|_{\mathrm{L}^2(\Omega)} \\ &\le 1 + (2 + C_{\mathrm{inv}} L^k N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}}) \|S_N\|_{\mathrm{L}^\infty(\Omega)} \\ &\le 1 + (2 + C_{\mathrm{inv}} N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}}) \|S_N\|_{\mathrm{L}^\infty(\Omega)}. \end{aligned} \tag{51}$$

Thus we define $c_\diamond := 1 + (2 + C_{\mathrm{inv}} N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}}) \|S_N\|_{\mathrm{L}^\infty(\Omega)}$ to deduce that, with this definition of $c_\diamond$, (46) holds with $k-1$ replaced by k, which then completes the inductive step. In particular, this implies that (48), and therefore also (49), holds for all $k \ge 1$.

Thus, from (48) and (49) we deduce that

$$\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \le L^{2k} \|S_N\|^2_{\mathrm{L}^2(\Omega)} \qquad \forall\, k \ge 1,$$

where $L \in (0,1)$, and hence $(S_N^k, D(u_N^k)) \to (S_N, D(u_N^k))$, as $k \to +\infty$; thus, by Korn's inequality, also $(S_N^k, u_N^k) \to (S_N, u_N^k)$ as $k \to +\infty$. □

Remark 5.2. Some remarks are in order at this point. As a function of λ, $L^2 = 1 - 2c_0\lambda + 4\lambda^2$ is minimized for $\lambda = \frac{1}{4}c_0$, yielding $L^2 = 1 - \frac{c_0^2}{4}$ (assuming that $c_0 \in (0,2)$, which can always be achieved by choosing $c_\diamond > \frac{1}{2}c_a$).

Our next remark concerns the choice of $c_\diamond$. As $C_{\mathrm{inv}} L^k N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}} \to 0$ when $k \to +\infty$, there exists a positive integer $k_0 = k_0(N)$ such that $C_{\mathrm{inv}} L^k N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}} \le 1$ for all $k \ge k_0$. For example, one can take

$$k_0 := \left\lceil \frac{\log C_{\mathrm{inv}} |\Omega|^{\frac{1}{2}} + \frac{d}{2} \log N}{\log \frac{1}{L}} \right\rceil + 1.$$

Using this refined upper bound in (51) allows us to redefine $c_\diamond$ as $c_\diamond := 1+3\|S_N\|_{\mathrm{L}^\infty(\Omega)}$. In fact, since we know from the proof of Theorem 4.1 that $\|S_N\|_{\mathrm{L}^\infty(\Omega)} \leq 2\|S\|_{\mathrm{L}^\infty(\Omega)}$ for all $N \geq N_*$, with N_* as defined in the proof of Theorem 4.1, we can further redefine $c_\diamond$ as $c_\diamond := 1 + \frac{1}{2}c_a + 6\|S\|_{\mathrm{L}^\infty(\Omega)}$, thus rendering $c_0 := \frac{c_a}{c_\diamond} \in (0,2)$ independent of N, and thereby $\lambda = \frac{1}{4}c_0$ and $L^2 = 1 - \frac{c_0^2}{4}$ become independent of N. In other words, once $N \geq N_*$ and $k \geq k_0(N) \sim \frac{d}{2}\log N$, the asymptotic rate of convergence of the iterative method (31), (32) is independent of N, provided that $(S,u) \in [\mathrm{H}^s_*(\Omega)]^{d\times d} \times [\mathrm{H}^{s+1}_*(\Omega)]^d$ with $s > \frac{d}{2}$.

6. Numerical experiments

In this section we shall report the results of numerical simulations whose aim is to assess the properties of the proposed numerical method and compare these with the theoretical results derived in the paper. We begin by considering a simple case, where we can analytically find the solution to the problem (1), (2) (and therefore check its regularity) and the discrete problem (4)–(6) is linear, so we can directly solve it without using an iterative method. We can then estimate the rate of convergence of the discrete solution (S_N, u_N) to the analytical solution (S,u) and compare the observed rate with the one asserted in Theorem 4.1. Our second example will be split into two parts: the aim of the first part is to test the convergence of the iterative method (proved in Section 5, Theorem 5.1) in a concrete example; in the second part we shall consider a more complicated example, involving a concentrated load, where we cannot compute the exact solution of the problem. The examples are simplified cases of the three-dimensional problem (1), (2). We will name them, respectively, 1D example and 2D example, the reason for this choice of terminology being that both the load function and the variables depend on one and two spatial coordinates, respectively.

Our work in this paper is still far from concrete engineering applications; we have therefore fixed the parameter r to be 0.5 in all of our simulations, because $r = 0.5$ is within the range for which the existence and uniqueness of a weak solution to our three-dimensional problem were asserted in Theorem 3.5. We wish to emphasize though that, conceptually, the discussion that follows is valid for all $r \in (0,\infty)$.

6.1 1D example

Suppose that $\Omega = (0, 2\pi)$, $r > 0$ is a fixed parameter of the model, and f is a 3-component vector-function (the load-vector) with the structure $f = (0, 0, f_3(x))^{\mathrm{T}}$, where $x \in \Omega$. Assume further that each of the components of S and u is a function of x only. This corresponds to the "physical" situation of a one-dimensional body lying horizontally, and the force acting on it vertically, the magnitude (but not the direction) of the force being dependent on the horizontal location x. Thanks to the assumptions we have made in this example and looking for a displacement vector u of

the form $u(x) = (0, 0, u_3(x))^{\mathrm{T}}$, the strong formulation (1), (2) of the problem becomes

$$\begin{cases} -(S_{13})_x = f_3 & \text{in } \Omega, \\ \frac{1}{2}(u_3)_x = \frac{S_{13}}{(1+|\sqrt{2}\,S_{13}|^r)^{\frac{1}{r}}} & \text{in } \Omega, \end{cases}$$

subject to a 2π-periodic boundary condition. We have denoted by the symbol S_{ij} the entry of the matrix S at position (i, j).

Weak formulation and discrete problem

In order to perform numerical simulations for this model problem we need to derive the algebraic interpretation of the Fourier spectral approximation of our problem. To avoid notational clutter let $S = S_{13}$, $u = u_3$, and $f = f_3$. Our 1D example in its strong formulation, for $r > 0$, is therefore

$$\begin{cases} -S'(x) = f(x) & \text{in } \Omega, \\ \frac{1}{2}u'(x) = \frac{S(x)}{(1+|\sqrt{2}\,S(x)|^r)^{\frac{1}{r}}} =: F_1(S(x)) & \text{in } \Omega, \end{cases} \tag{52}$$

where $x \in \Omega$ and $|\cdot|$ stands for the modulus operation on $\mathbb{R}$. The weak formulation of the problem is then as follows: find $(S, u) \in \Sigma \times V$ such that

$$\begin{cases} F_1(S) = \frac{1}{2}u', \\ (S, v') = (f, v) \qquad \forall\, v \in V, \end{cases} \tag{53}$$

where $\Sigma := \mathrm{L}^1_*(\Omega)$, $V := \{\omega \in \mathrm{L}^1_{\#}(\Omega) : \omega_x \in \mathrm{L}^\infty_{\#}(\Omega),\ \int_\Omega \omega(x)\,\mathrm{d}x = 0\}$ (note that V is the one-dimensional version of the space $D^1_{*,\infty}(\Omega)$ defined in Section 3.3). Under the assumption $f \in \mathrm{W}^{1,t}_{\#}(\Omega)$ for some $t > 1$, Theorem 3.5 guarantees the existence of a unique solution to (53).

The spectral Galerkin method for the discrete problem is: find $(S_N, u_N) \in \Sigma_N \times V_N$ such that

$$\begin{cases} (F_1(S_N) - \frac{1}{2}u'_N, T_N) = 0 & \forall\, T_N \in \Sigma_N, \\ (S_N, v'_N) = (f, v_N) & \forall\, v_N \in V_N, \end{cases} \tag{54}$$

where $\Sigma_N := \mathcal{S}_N$, $V_N := \mathcal{S}_N$ and $\mathcal{S}_N$ is the space of all univariate 2π-periodic real-valued trigonometric polynomials of degree $\leq N$ whose integral over Ω is equal to 0. Note that in this particular case the spaces Σ_N and V_N coincide, but we prefer to denote them with different symbols for the sake of consistency with our earlier notation.

The terms involving f and $F_1(S_N)$ are generally difficult to compute exactly: we shall therefore use suitable quadrature rules.

Numerical simulations in a simple case, with $r = 0.5$

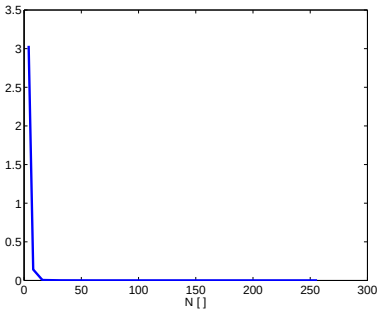
It is generally impossible to compute the analytical solution of our problem, but in this $1D$ example we can choose a specific f that allows us to find the analytical solution in closed form. Having done so, we shall compare it with the numerical solution and assess the behaviour of the approximation error in the limit of $N \to +\infty$. Taking $f(x) := 2^{\frac{r-1}{2}}\,(\sin x)\,(2^{\frac{r}{2}} - |\cos x|^r)^{-\frac{r+1}{r}}$ as the right-hand side of the equation, the exact solution (S, u) is $S(x) = \frac{\cos x}{\sqrt{2}\,(2^{\frac{r}{2}} - |\cos x|^r)^{\frac{1}{r}}}$, $u(x) = \sin x$.

We have thus found the exact solution to the problem (52), which is also a solution to the weak formulation (53) (and we know that this is the unique weak solution in the case of $r = 0.5$ by Theorem 3.5). Now we are ready to show the comparison between our analytical solution and the numerical solution: all integrals have been approximated by a global adaptive quadrature rule, using the MATLAB [8] command *integral*.

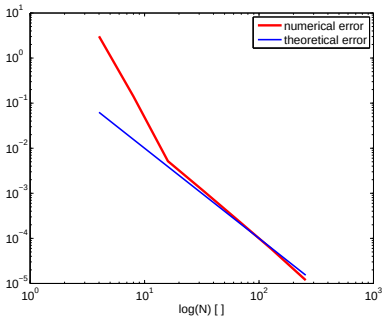
In Figure 1A we have reported the behaviour of the sum of the L^2-norm errors (the sum of the left-hand sides of (11) and (12)) against N: the error decreases rapidly to zero as N increases.

Starting from $N = 4$ up to $N = 256$, we have shown (in red) the sum of the L^2-norm errors of S_N and u_N against the degree N in Figure 1B, where we have used a logarithmic scale on both axes. Noting the regularity of the analytical solution, i.e., that $S \in \mathrm{H}^2_*(\Omega)$ and $u \in \mathrm{C}^\infty_*(\overline{\Omega})$, we expect from the error bounds (11)–(12) that the rate of convergence is (approximately) 2. From Figure 1B it is clear that the approximation error decreases as N^{-2}, in agreement with the theory. Furthermore, we observe in Figure 1B that the numerical solution is an accurate approximation of the analytical solution when the discretization parameter N exceeds a given threshold, in this case, roughly 15: this is consistent with the fact that the asymptotic error bounds (11), (12) from Theorem 4.1 hold for all $N \geq N_*$, with N_* sufficiently large.

It is clear from Figure 1B though that the quadrature error is negligible compared with the approximation error.



(A) Sum of the L^2-norm errors plotted against N ($r = 0.5$)



(B) Sum of the L^2-norm errors plotted against N using a logarithmic scale on both axes

Figure 1: Behaviour of the approximation error: 1D example

6.2 2D example

Assume that $\Omega = (0, 2\pi)^2$, $r > 0$ is a fixed parameter featuring in the model, and f is a 3-component vector-function of the form $f = (0, 0, f_3(x, y))^{\mathrm{T}}$, where $(x, y) \in \Omega$. Furthermore we suppose that each component of S and u is a function of x and y only. This example corresponds to the "physical" situation where a vertical force acts on a two-dimensional body lying in the horizontal plane, and the magnitude (but not the direction) of the force is dependent on the horizontal coordinates x and y.

Under the assumptions we have made and considering a displacement vector u of the form $u(x,y) = (0,0,u_3(x,y))^{\mathrm{T}}$, the strong formulation (1), (2) of our problem becomes

$$\begin{cases} -(S_{13}(x,y))_x - (S_{23}(x,y))_y = f_3(x,y), \\ \frac{1}{2}(u_3(x,y))_x = \frac{S_{13}(x,y)}{(1+(2\,S_{13}^2(x,y)+2\,S_{23}^2(x,y))^{\frac{r}{2}})^{\frac{1}{r}}}, \\ \frac{1}{2}(u_3(x,y))_y = \frac{S_{23}(x,y)}{(1+(2\,S_{13}^2(x,y)+2\,S_{23}^2(x,y))^{\frac{r}{2}})^{\frac{1}{r}}}, \end{cases}$$

subject to 2π-periodic boundary conditions. As before, the symbol S_{ij} stands for the entry of the matrix S at position (i,j).

Weak formulation and discrete problem

As in the first example, we derive the algebraic system starting from the weak formulation with the aim to perform numerical experiments. In order to avoid notational clutter let $S_1 = S_{13}$, $S_2 = S_{23}$, $u = u_3$, $f = f_3$. Our 2D model problem in strong form thus becomes

$$\begin{cases} \frac{1}{2}(u(x,y))_x = \frac{S_1(x,y)}{(1+(2\,S_1^2(x,y)+2\,S_2^2(x,y))^{\frac{r}{2}})^{\frac{1}{r}}}, \\ \frac{1}{2}(u(x,y))_y = \frac{S_2(x,y)}{(1+(2\,S_1^2(x,y)+2\,S_2^2(x,y))^{\frac{r}{2}})^{\frac{1}{r}}}, \\ -(S_1(x,y))_x - (S_2(x,y))_y = f(x,y), \end{cases} \tag{55}$$

where $(x,y) \in \Omega$.

The weak formulation of the continuous problem is the following: find $(S_1, S_2, u) \in \mathrm{L}^1_*(\Omega)^2 \times V$ such that

$$\begin{cases} \frac{S_1}{(1+(2\,S_1^2+2\,S_2^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}u_x = 0, \\ \frac{S_2}{(1+(2\,S_1^2+2\,S_2^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}u_y = 0, \\ ((S_1,S_2)^{\mathrm{T}}, \nabla v) = (f,v) \qquad \forall\, v \in V, \end{cases} \tag{56}$$

where $V := \{\omega \in \mathrm{L}^1_{\#}(\Omega) : \nabla\omega \in \mathrm{L}^\infty_{\#}(\Omega)^2,\ \int_\Omega \omega(x)\,\mathrm{d}x = 0\}$. The existence and uniqueness of the solution to the previous problem is guaranteed by Theorem 3.5, under the assumption $f \in W^{1,t}_{\#}(\Omega)$ for some $t > 1$.

The spectral Galerkin method for the discrete problem is: find $(S_{1,N}, S_{2,N}, u_N) \in \mathcal{S}_N^2 \times \mathcal{S}_N$ such that

$$\begin{cases} \left(\frac{S_{1,N}}{(1+(2\,S_{1,N}^2+2\,S_{2,N}^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}(u_N)_x, T_N \right) = 0 \quad \forall\, T_N \in \mathcal{S}_N, \\ \left(\frac{S_{2,N}}{(1+(2\,S_{1,N}^2+2\,S_{2,N}^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}(u_N)_y, T_N \right) = 0 \quad \forall\, T_N \in \mathcal{S}_N, \\ ((S_{1,N}, S_{2,N})^{\mathrm{T}}, \nabla v_N) = (f, v_N) \qquad\qquad \forall\, v_N \in \mathcal{S}_N, \end{cases} \tag{57}$$

where $\mathcal{S}_N$ is the space of bivariate 2π-periodic real-valued trigonometric polynomials of degree $\leq N$ whose integral over Ω is equal to 0.

In Section 5 we have studied the iterative method in the general case: we now reinterpret the results shown previously using the hypotheses of the $2D$ example: the aim is to find the system of linear algebraic equation that we need to solve in each

step of our iterative method. Let us consider the linearization of (57), which we have discussed in Section 5; we do so by applying the iterative method (31), (32) to our $2D$ example.

Given an initial guess $S^0_{1,N} \equiv 0, S^0_{2,N} \equiv 0$, find $(S^n_{1,N}, S^n_{2,N}, u^n_N) \in \mathcal{S}^2_N \times \mathcal{S}_N$ for all $n = 1, 2, \dots$ (subject to a suitable stopping criterion) such that

$$\begin{cases} (S^n_{1,N}, T_N) - \lambda\,(\frac{1}{2}\,(u^n_N)_x, T_N) = (S^{n-1}_{1,N}, T_N) - \lambda\left(\frac{S^{n-1}_{1,N}}{(1+(2\,(S^{n-1}_{1,N})^2+2\,(S^{n-1}_{2,N})^2)^{\frac{r}{2}})^{\frac{1}{r}}}, T_N\right) \\ \qquad \forall\, T_N \in \mathcal{S}_N, \\ (S^n_{2,N}, T_N) - \lambda\,(\frac{1}{2}\,(u^n_N)_y, T_N) = (S^{n-1}_{2,N}, T_N) - \lambda\left(\frac{S^{n-1}_{2,N}}{(1+(2\,(S^{n-1}_{1,N})^2+2\,(S^{n-1}_{2,N})^2)^{\frac{r}{2}})^{\frac{1}{r}}}, T_N\right) \\ \qquad \forall\, T_N \in \mathcal{S}_N, \\ ((S^n_{1,N}, S^n_{2,N})^{\mathrm{T}}, \nabla v_N) = (f, v_N) \qquad \forall\, v_N \in \mathcal{S}_N. \end{cases}$$

For the rest of this paragraph we will consider the parameter λ to be given: we will return below to the question of choosing λ.

So far we have not specified the stopping criterion for the iterative method. Given a tolerance `TOL`, a possible choice could be the following:

$$\frac{\|S^n_{1,N} - S^{n-1}_{1,N}\|_{\mathrm{L}^2(\Omega)} + \|S^n_{2,N} - S^{n-1}_{2,N}\|_{\mathrm{L}^2(\Omega)} + \|u^n_N - u^{n-1}_N\|_{\mathrm{L}^2(\Omega)}}{\|S^{n-1}_{1,N}\|_{\mathrm{L}^2(\Omega)} + \|S^{n-1}_{2,N}\|_{\mathrm{L}^2(\Omega)} + \|u^{n-1}_N\|_{\mathrm{L}^2(\Omega)}} \leq \texttt{TOL}. \tag{58}$$

Numerical simulations in a simple case, with $r = 0.5$

We start to test our iterative method in an easy case where we know the analytical solution and therefore we can compare it with the numerical solution. Let us focus on the strong formulation (55) of the two-dimensional case: consider the sum of $(55)^2_1$ and $(55)^2_2$, multiply this by 2 before taking the square root, and finally raise the resulting expression to the r-th power. We obtain that

$$(2\,(S_1^2 + S_2^2))^{\frac{r}{2}} = \frac{(u_x^2 + u_y^2)^{\frac{r}{2}}}{2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}}}.$$

Hence, by inserting this expression into $(55)_1$ and $(55)_2$, we have that

$$S_1 = \frac{u_x}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}, \qquad S_2 = \frac{u_y}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}. \tag{59}$$

Substituting these into $(55)_3$, we arrive at the following expression for the right-hand side f in terms of the displacement component u and its derivatives:

$$f = -\left(\frac{u_x}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}\right)_x - \left(\frac{u_y}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}\right)_y. \tag{60}$$

Next, we fix the displacement u and then, using (60), we obtain the right-hand side f corresponding to the chosen displacement u. Note further that, given such a u, we can compute the components S_1 and S_2 of the stress tensor using (59).

Consider, for example, $u(x, y) = \frac{1}{2}\sin(x + y)$ and note that f, as well as S_1 and S_2, are always well-defined for all $x \in \overline{\Omega}$ (in contrast with the 1D example, here we have multiplied the function $\sin(x + y)$ by $\frac{1}{2}$ to ensure that the denominators in (59)

and (60) are different from zero for all $r > 0$). After some calculations, it follows that

$$f(x,y) = 2^{\frac{r-1}{2}}\,\sin(x+y)\left(2^{\frac{r}{2}} - \left(\frac{1}{2}\,\cos^2(x+y)\right)^{\frac{r}{2}}\right)^{-\frac{r+1}{r}},$$

$$S_1(x,y) = S_2(x,y) = \frac{1}{2\sqrt{2}}\,\frac{\cos(x+y)}{(2^{\frac{r}{2}} - (\frac{1}{2}\,\cos^2(x+y))^{\frac{r}{2}})^{\frac{1}{r}}}.$$

Concerning the choice of the parameter λ in the iterative method, Remark 5.2 (note that Theorem 4.1 applies for this example) and the knowledge of the exact solution enable us to fix the parameter λ in the correct "interval of convergence", $(0, \frac{1}{2}\,c_0)$. In this case we have (using $r = 0.5$) that $\|S\|_{\mathrm{L}^\infty(\Omega)} \le 6$, $c_* = 37$, $c_0 = \frac{1}{37*2^{3/2}}$, and we can therefore choose $\lambda = \frac{1}{4}\,c_0 = \frac{1}{37*2^{7/2}} \simeq 0.002$ in our numerical simulations.

We now aim to test the accuracy of the numerical solution computed by using the proposed iterative method. The inner products have been approximated by a locally adaptive quadrature rule (using the MATHEMATICA [9] command *NIntegrate*).

In the particular case discussed here we took $N = 5$ and, since for this model problem the analytical solution is known, a slightly different stopping criterion than (58) was used, which was the following:

$$\frac{\|S^n_{1,N} - S_1\|_{\mathrm{L}^2(\Omega)} + \|S^n_{2,N} - S_2\|_{\mathrm{L}^2(\Omega)} + \|u^n_N - u\|_{\mathrm{L}^2(\Omega)}}{\|S_1\|_{\mathrm{L}^2(\Omega)} + \|S_2\|_{\mathrm{L}^2(\Omega)} + \|u\|_{\mathrm{L}^2(\Omega)}} \le \mathtt{TOL},$$

with $\mathtt{TOL} = 10^{-2}$. After 2 iterations the relative errors (in the L^2-norm) for S_1 (as well as S_2) and the displacement were

$$\frac{\|S^2_{1,N} - S_1\|_{\mathrm{L}^2(\Omega)}}{\|S_1\|_{\mathrm{L}^2(\Omega)}} = 0.006, \qquad \frac{\|u^2_N - u\|_{\mathrm{L}^2(\Omega)}}{\|u\|_{\mathrm{L}^2(\Omega)}} = 0.001, \tag{61}$$

which confirm that the numerical solution is close to the analytical solution.

We have also tested our iterative method by fixing the degree N and starting from different initial guesses: the method, in all cases, converged to the same solution $(S_{1,N}, S_{2,N}, u_N)$. This concurs with our convergence result proved in Section 5: for all $N \in \mathbb{N}$ given, the sequence (S^k_N, u^k_N) of iterates converges to (S_N, u_N), defined as the unique solution of (4)–(6), as $k \to \infty$.

In the knowledge of the analytical solution we can verify that the assumptions of Theorem 4.1 hold and we therefore expect that the sequence of solutions $(S_{1,N}, S_{2,N}, u_N)$ converges to the analytical solution of the nonlinear problem (56) as $N \to \infty$. As one can already see from (61), with just $N = 5$ and two steps of the iterative algorithm the resulting numerical solution is already close to the analytical solution, in agreement with the theoretical results.

Numerical simulations in the case of a concentrated load, with $r = 0.5$

As we have explained in the Introduction, limiting strain models are typified by the fact that the linearized strain is a bounded function even if the stress is very large. In our numerical simulations we would like to simulate the effect of a large stress concentration on the displacement: for this reason it seems reasonable to consider a regularized Dirac delta function as right-hand side f. Ideally we would have liked to

take f to be a Dirac delta function, but we chose to regularize it because in this case we do not know the exact solution (and thus we cannot check its regularity as we did in Section 6.1) and Theorem 3.5 does not guarantee the minimal regularity that we need to be able to apply Theorem 4.1 to deduce that the sequence of numerical solutions converges to the analytical solution as $N \to \infty$. The consideration of measures as source terms in the problem will be the subject of future research. By using $f \in C_0^\infty(\Omega)$ instead of a delta function we will avoid any limitations that might arise from the regularity of the data.

For any $h > 0$ and $x \in (0, 2\pi)$ we consider

$$\varphi_h(x) := \begin{cases} \frac{c}{h} \exp\left\{ \frac{1}{|\frac{x-\pi}{h}|^2 - 1} \right\}, & 0 < |\frac{x-\pi}{h}| < 1, \\ 0, & |\frac{x-\pi}{h}| > 1, \end{cases}$$

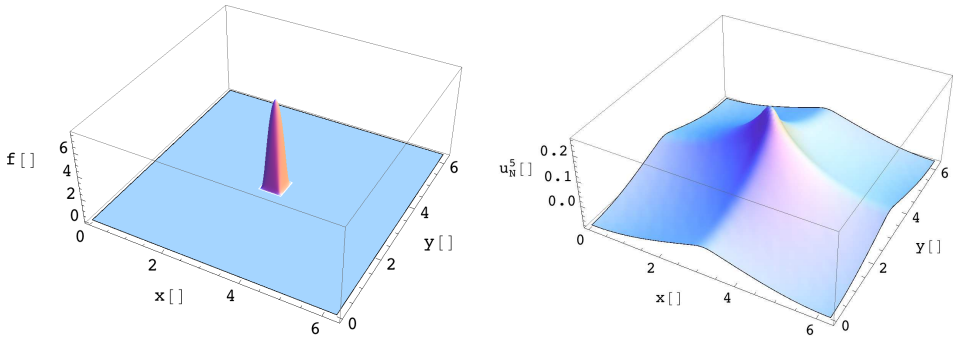
where $c = \left(\int_{-1}^{1} \exp\left\{ \frac{1}{|x|^2-1} \right\} \mathrm{d}x \right)^{-1}$ (we will use a global adaptive quadrature rule to approximate the constant c). Note that φ_h is an approximation to the delta function concentrated at $x = \pi$, for small values of $h > 0$. Note further that φ_h belongs to $W^{s,p}_{\#}(\Omega)$ for all $s \geq 0$, all $p \in [1, \infty]$, and all $h \in (0, \pi)$. In the two-dimensional case we shall approximate a bivariate Dirac delta function concentrated at the point (π, π) by $f_h(x, y) := \varphi_h(x)\varphi_h(y)$, where $(x, y) \in \Omega = (0, 2\pi)^2$.

As regards the choice of the parameter λ featuring in the iterative method, we cannot repeat the argument that we used in Section 6.2 because, this time, the exact solution is not available. Therefore the selection of the parameter λ is more critical since we do not know the precise interval $(0, \frac{1}{2} c_0)$ where we have to choose λ to guarantee convergence of the iterative method. A possible strategy is to fix a small value of λ and run a simulation with that value of λ; if a plausible output is obtained, then the iterative method is likely to have converged; otherwise λ should be reduced. In the example that follows we made the same choice as in Section 6.2, i.e., $\lambda = \frac{1}{37 * 2^{7/2}} \simeq 0.002$, and the method was seen to have worked 'properly'.

Using $f = f_h$ as defined above for a given $h > 0$, we have solved iteratively the discrete problem until the stopping criterion defined in (58) was fulfilled, using the fixed tolerance $\texttt{TOL} = 10^{-3}$. As in Section 6.2, the inner products were approximated using a local adaptive quadrature rule (the MATHEMATICA [9] command *NIntegrate*).

With the choice of the parameter $h = 0.3$, the corresponding body force f is reported in Figure 2A; fixing the degree $N = 30$, we obtained the numerical displacement shown in Figure 2B after 5 iterations. Increasing the degree N did not visibly change the numerical solution, indicating that the numerical solution that we have computed is likely to be close to the unique analytical solution (corresponding to $h = 0.3$).

This example confirms what we have expected: the displacement (Figure 2B) has a peak at $(x, y) = (\pi, \pi)$, where the body force has a peak as well, but the magnitude of that peak is significantly smaller than the one in f, and this is due to the nonlinearity of the model.



(A) Body force f with $h = 0.3$

(B) Displacement component u_N^5 with $h = 0.3$ and $N = 30$

Figure 2: Numerical simulation with the regularized delta function as body force

7. Conclusions

This paper provides a first step towards the rigorous mathematical analysis of spectral approximations of a nonlinear elastic limiting strain model. The spectral method we have constructed was shown to exhibit optimal order convergence. We have also proposed an iterative method for the numerical solution of the finite-dimensional system of nonlinear equations resulting from the Fourier spectral discretization of the problem, and have proved that it converges at a linear rate to the unique solution of the discretized problem. The recent paper [2] focuses on the analysis of low-order finite element approximations of a general limiting strain model on a bounded open polytopal domain in $\mathbb{R}^d$, $d \in \{2, 3\}$, subject to a homogeneous Dirichlet boundary condition, in the spirit of the PDE analysis developed in the paper [3]. When the Neumann part of the boundary is nonempty, the structure of the solution is potentially much more complicated. It was shown in [1] that, in general, the solution in that case belongs to the space of Radon measures, but if the problem is equipped with a so-called asymptotic radial structure, then the solution can in fact be understood as a standard weak solution, with one proviso: the attainment of the boundary value is penalized by a measure supported on the Neumann part of the boundary. The numerical analysis of a mixed Dirichlet–Neumann boundary-value problem for limiting strain models therefore possesses nontrivial new challenges, which will be considered in future publications.

Acknowledgement. The research reported in this paper was conducted while the first author was visiting student from the Politecnico di Milano at the Mathematical Institute, University of Oxford.

References

[1] L. Beck, M. Bulíček, J. Málek, E. Süli, *On the existence of integrable solutions to nonlinear elliptic systems and variational problems with linear growth*, Arch. Ration. Mech. An., **225(2)** (2017), 717–769.

[2] A. Bonito, V. Girault, E. Süli, *Finite element approximation of a strain-limiting elastic model*, IMA J. Numer. Anal., (2018), Published electronically. https://doi.org/10.1093/imanum/dry065.

[3] M. Bulíček, J. Málek, K. R. Rajagopal, E. Süli, *On elastic solids with limiting small strain: modelling and analysis*, EMS Surveys in Mathematical Sciences, **1(2)** (2014), 283–332.

[4] M. Bulíček, J. Málek, E. Süli, *Analysis and approximation of a strain-limiting nonlinear model*, Math. Mech. Solids, **20** (2015), 92–118.

[5] C. Canuto, A. Quarteroni, *Approximation results for orthogonal polynomials in Sobolev spaces*, Math. Comp., **38** (1982), 67–86.

[6] V. Girault, P. A. Raviart, *Finite Element Approximation of the Navier–Stokes Equations*, Lect. Notes Math., **749**, Springer-Verlag, 1979.

[7] V. Girault, P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer Ser. Comp. Math., **5**, Springer-Verlag, 1986.

[8] MATLAB, version 8.3 (R2014a), The MathWorks Inc. Natick, MA 01760, 2014.

[9] Mathematica, version 9.0.1, The Wolfram Centre, Oxfordshire OX29 8FD, 2013.

[10] K. R. Rajagopal, *On implicit constitutive theories*, Appl. Math., **48** (2003), 279–319.

[11] K. R. Rajagopal, *Elasticity of elasticity*, Zeitschrift fur Angewandte Math. Phys., **58** (2007), 309–417.

(received 04.07.2018; in revised form 18.12.2018; available online 23.12.2018)

Bain & Company Italy Inc., Via Crocefisso, 10, 20122 Milano MI, Italy

E-mail: nicolo.gelmetti@bain.com

Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, United Kingdom

E-mail: Endre.Suli@maths.ox.ac.uk