SPECTRAL APPROXIMATION OF A NONLINEAR ELASTIC LIMITING STRAIN 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 is 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 equations, and is shown to converge, at a linear rate, to the unique solution of the numerical method. The theoretical results are illustrated with numerical experiments.

Keywords: Fourier spectral method, convergence, nonlinear elasticity, implicit constitutive theory

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, [13], [14]). 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 [14], 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 modeling 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. For example, Rajagopal's limiting-strain elastic models stemming from implicit constitutive theory seem to provide good description of Fung's experimental data concerning the passive response of biological tissues, which indicate that the stress/strain response of the tissue is, to a good approximation, exponential (see, for example, [6]).

As has been indicated in reference [3] and the survey article [1], limiting strain models have been thus far studied in several situations. In the case of special deformations such as shearing, compressions, torsion, etc., Rajagopal himself, and Bustamante and Rajagopal aimed to assess whether the models exhibit the expected responses (cf. [4], [16], [15]). In the case of anti-plane strain (stress) problems, considered in domains with nonconvex cross-sections (including thus the domains with V-notches or cracks), the resulting scalar problem in two space dimensions has been analyzed by methods of asymptotic analysis in [17], by performing systematic computational tests in [9], and by analytical methods of modern theory of nonlinear partial differential equations in [2]; the last result establishes the existence of weak solutions in nonconvex domains for values of the model parameter r in the range $r \in (0, 2)$, see equation (2) below, and in convex domains for the range $r \in (0, \infty)$. A detailed computational study of the complete problem in planar domains was performed in Ortiz et al. [12]. The recent paper [3] has been the first one with focus on the mathematical analysis of general boundary-value problems (which include systems of $\frac{1}{2}d(d+3)$ time-independent nonlinear partial differential equations of first order), featuring in

Date: May 28, 2018.

This work was done while the first author was visiting student from the Politecnico di Milano at the Mathematical Institute of the University of Oxford.

limiting strain models, in bounded subsets of $\mathbb{R}^d$, $d \geq 2$; the existence of a weak solution was shown, in the case of periodic boundary conditions, for all values of the model parameter r in the range $(0, \frac{2}{d})$ and the existence of a renormalized solution was established for all values of $r \in (0, \infty)$. The subsequent paper [1] surveys the physical background and the mathematical analysis of boundary-value problems associated with models with limiting small strain, and presents the first analytical result concerning the existence of weak solutions in general three-dimensional domains.

The analysis of numerical algorithms for limiting strain models is currently lacking. The present paper is a first step 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 [3], i.e., in an axiparallel parallelopipedal domain 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.

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. This essential simplification helps us to introduce not only the concept of weak solution to the problem under consideration, but also the concept of a renormalized solution. The spatially periodic setting also helps us to *construct* the solution via a specific numerical method, namely the Fourier spectral method. The proof of existence of weak and renormalized solutions to the model in this geometry presented in [3] is therefore, at the same time, a proof of the convergence of the sequence of numerical approximations to the unknown analytical solution. This simplified setting allows us to provide a fairly complete picture regarding numerical analysis for a nontrivial example of a strain-limiting nonlinear elastic model.

The problem under consideration here is 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. Our objective is to construct a Fourier spectral approximation (S_N, u_N) to (S, u), where S is the stress tensor and u is the displacement, which belong to suitable function spaces consisting of $d \times d$ 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

$$F(S) := \frac{S}{(1 + |S|^r)^{\frac{1}{r}}}, \qquad S \in \mathbb{R}^{d\times d},$$

where $r > 0$, and $|\cdot|$ denotes the Frobenius norm on $\mathbb{R}^{d\times d}$, defined by $|X|^2 := X : X = \mathrm{tr}(X^{\mathrm{T}} X)$. 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)
  $$(F(A) - F(B)) : (A - B) \geq c_a \, \frac{|A - B|^2}{(1 + |A| + |B|)^{r+1}} \qquad \forall\, A, B \in \mathbb{R}^{d\times d},$$
  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 [3].

Lemma 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. *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, (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 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 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,*

$$\big||x|^\mu - |y|^\mu\big| \leq \tfrac{1}{\mu}\big||x|-|y|\big|^\mu \leq \tfrac{1}{\mu}|x-y|^\mu \qquad \forall\, x,y \in \mathbb{R}^{d\times d}.$$

(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,*

$$\big||x|^\mu - |y|^\mu\big| \leq \mu R^{\mu-1}\big||x|-|y|\big| \leq \mu R^{\mu-1}|x-y| \qquad \forall\, x,y \in \mathcal{B}(0,R).$$

(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 [3] 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 here 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 [3], 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 the Sobolev space $[H^s(\Omega)]^{d\times d} \times [H^s(\Omega)]^{d\times d}$, with $s > \frac{d}{2}$, and 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 equations, and is shown to converge to the unique solution of the numerical method. In section 6 we report some numerical experiments in order to test the theoretical results of the paper in some 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 of 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 signified 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, 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

$$\mathrm{H}_\#(\mathrm{div};\Omega) := \{v \in [\mathrm{L}^2_\#(\Omega)]^d \,:\, \text{such that } \operatorname{div} v \in \mathrm{L}^2_\#(\Omega)\},$$

equipped with the norm

$$\|v\|_{\mathrm{H}(\mathrm{div};\Omega)} := \left(\|v\|^2_{\mathrm{L}^2(\Omega)} + \|\operatorname{div} v\|^2_{\mathrm{L}^2(\Omega)}\right)^{\frac{1}{2}}.$$

For $s > 0$, the (potentially, fractional-order,) Hilbertian Sobolev spaces of periodic functions $\mathrm{H}^s_\#(\Omega)$ and $\mathrm{H}^s_*(\Omega)$, are defined analogously, through closure of $\mathrm{C}^\infty(\overline{\Omega})$ in the norm of $\mathrm{H}^s(\Omega)$.

We shall require the following periodic version of Korn's inequality [3].

Lemma 4 (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 inequalities hold:*

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

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, $D^{\mathrm{dev}}(v) := D(v) - \frac{1}{d}(\operatorname{div} v)\mathrm{I}$ 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

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

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, with integral average over the set Ω equal to 0. We want to highlight that the previous definition of Σ_N is slightly different from the one in [3], where indeed Σ_N is a subset of $H_{\#,\mathrm{symm}}(\mathrm{div};\Omega)$; but all the results proved in [3] remain unchanged if we assume that Σ_N is a space of functions with null intergal average. This is the reason why we decided to make this little modification at the beginning of the present study.

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 [3], 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 1 below, guaranteeing the existence and uniqueness of a solution to the discrete problem (4)–(6), was established in [3]; 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 [8], Corollary 1.1, p.279).

Lemma 5. *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 1. *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 $(S^i_N, u^i_N) \in \Sigma_N \times V_N$ that solve (4)–(6) for $i = 1, 2$. Hence,

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

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,

$$\begin{aligned}
&-(\operatorname{div}(S_N^1 - S_N^2), u_N^1 - u_N^2) - (D(u_N^1 - u_N^2), S_N^1 - S_N^2) \\
&\quad = (S_N^1 - S_N^2, \nabla(u_N^1 - u_N^2)) - (D(u_N^1 - u_N^2), S_N^1 - S_N^2) \\
&\quad = (S_N^1 - S_N^2, D(u_N^1 - u_N^2)) - (D(u_N^1 - u_N^2), S_N^1 - S_N^2) = 0.
\end{aligned}$$

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 4, 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

$$\mathfrak{F}(U_N) := P_N F(U_N + \hat{S}_N), \qquad U_N \in \Sigma_{N,0},$$

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),

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

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

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

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 5 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

$$(D(u_N), T_N) = (F(S_N), T_N) \qquad \forall\, T_N \in \Sigma_N.$$

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

$$b(v_N, T_N) := -(v_N, \operatorname{div} T_N) \quad \text{and} \quad \ell(T_N) := (F(S_N), T_N)\,.$$

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 Lemma 4.1(ii) on p.40 of Girault & Raviart [7].

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 $v_N \in V_N$. Part (iii) of Lemma 4.1 on p.40 of Girault & Raviart [7] implies, again thanks to the inf-sup condition (3), 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 1. *The statement in the final paragraph of the proof above, that $\hat{S}_N \in \Sigma_N$, can be refined: in fact, $\hat{S}_N \in \Sigma_{N,0}^\perp$, where $\Sigma_{N,0}^\perp$ 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 Part (iii) of Lemma 4.1 on p.40 of [7]*, 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 Part (iii) of Lemma 4.1 on p.40 of [7].

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 4) 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. [3]), $[\mathrm{C}^\infty_*(\overline{\Omega})]^d$ 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

$$\int_\Omega T(x) : D(v_n(x))\,\mathrm{d}x \overset{n\to\infty}{\to} \int_\Omega T(x) : D(v(x))\,\mathrm{d}x \qquad \forall\, T \in [\mathrm{L}^1_\#(\Omega)]^{d\times d}.$$

We recall the following result from [3] concerning the convergence of the sequence of approximate solutions generated by (4)–(6) to a weak solution of the boundary-value problem.

Theorem 2. *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 [3] 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. Theorem 5.1 in [3]) 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

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

where $0 \leq s' \leq s$.

Theorem 3. *Suppose that $(S, D(u)) \in [\mathrm{H}^s_*(\Omega)]^{d\times d} \times [\mathrm{H}^s_*(\Omega)]^{d\times d}$, where $s > \frac{d}{2}$. Then, there exists a positive constant c_*, independent of N, and a positive integer N_* such that*

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

and

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

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:

$$(\operatorname{div} S_{N,0}, v_N) = 0 \qquad \forall\, v_N \in V_N, \tag{15}$$

$$(F(S_{N,0} + \hat{S}_N), T_N) + (u_N, \operatorname{div} T_N) = 0 \qquad \forall\, T_N \in \Sigma_N. \tag{16}$$

Hence, in particular,

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

and therefore

$$\begin{aligned}(F(S_{N,0} + \hat{S}_N) - F(\hat{S}_N), T_N) &= -(F(\hat{S}_N), T_N)\\ &= (F(S) - F(\hat{S}_N), T_N) - (F(S), T_N)\\ &= (F(S) - F(\hat{S}_N), T_N) - (D(u), T_N)\\ &= (F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) - (P_N D(u), T_N)\\ &= (F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) - (D(u), T_N)\\ &= (F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) + (u, \operatorname{div} T_N)\\ &= (F(S) - F(\hat{S}_N), T_N) - (D(u) - P_N D(u), T_N) \qquad \forall\, T_N \in \Sigma_{N,0}. \end{aligned} \tag{18}$$

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 (18) that

$$(F(S_{N,0} + \hat{S}_N) - F(\hat{S}_N), T_N) = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}. \tag{19}$$

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

$$\begin{aligned}|\ell(T_N)| &\leq \Big(c_b \|S - \hat{S}_N\|_{\mathrm{L}^2(\Omega)} + \|D(u) - P_N D(u)\|_{\mathrm{L}^2(\Omega)}\Big) \|T_N\|_{\mathrm{L}^2(\Omega)}\\ &\leq \big(c_b c_1 N^{-s} \|S\|_{\mathrm{H}^s(\Omega)} + c_1 N^{-s} \|D(u)\|_{\mathrm{H}^s(\Omega)}\big) \|T_N\|_{\mathrm{L}^2(\Omega)}\\ &\leq c_b c_1 N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right) \|T_N\|_{\mathrm{L}^2(\Omega)} \qquad \forall\, T_N \in \Sigma_N. \end{aligned} \tag{20}$$

Our objective is to prove that there exist a $c_* > 0$, independent of N, and $N_* \in \mathbb{N}$, such that for each $N \geq N_*$ there exists a unique $S_{N,0} \in \Sigma_{N,0}$ such that (19) holds and

$$\|S_{N,0}\|_{\mathrm{L}^2(\Omega)} \leq c_* N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right). \tag{21}$$

We shall use a fixed point theorem to this end. In order to define the fixed point mapping, we begin by noting that, by Lemma 3.2 in [3],

$$(F(A) - F(B)) : C = \int_0^1 G(\theta A + (1-\theta) B; A - B, C) \,\mathrm{d}\theta,$$

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

$$|G(\gamma; \alpha, \beta)| \leq \frac{2\,|\alpha|\,|\beta|}{(1 + |\gamma|^r)^{\frac{1}{r}}} \qquad \forall\, \alpha, \beta, \gamma \in \mathbb{R}^{d\times d}_{\mathrm{sym}}, \tag{22}$$

$$G(\gamma; \alpha, \alpha) \geq \frac{|\alpha|^2}{(1 + |\gamma|^r)^{1+\frac{1}{r}}} \qquad \forall\, \alpha, \gamma \in \mathbb{R}^{d\times d}_{\mathrm{sym}}. \tag{23}$$

We define the set

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

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 (19) 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 (19), we consider the following mapping: to each $\varphi \in \mathfrak{B}_{N,0}$ we assign $S_{N,\varphi} \in \Sigma_{N,0}$ such that

$$\int_\Omega \int_0^1 G(\hat{S}_N + \theta\varphi; S_{N,\varphi}, T_N)\,\mathrm{d}\theta\,\mathrm{d}x = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}. \tag{24}$$

It follows from (23) that, for $\hat{S}_N \in \Sigma_N$ and $\varphi \in \mathfrak{B}_{N,0}$ fixed, (24) has at most one solution $S_{N,\varphi} \in \Sigma_{N,0}$. Since $\Sigma_{N,0}$ is a finite-dimensional linear space and (24) 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 (23), (24) and (20),

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

Thus we deduce that

$$\|S_{N,\varphi}\|_{\mathrm{L}^2(\Omega)} \leq c_b c_1 N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right) (1 + (\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} + \|\varphi\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}}. \tag{25}$$

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_*,

$$\begin{aligned}
&c_b c_1 N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right) (1 + (\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} + \|\varphi\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}} \\
&\qquad \leq c_* N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right) \qquad \forall\, N \geq N_*.
\end{aligned} \tag{26}$$

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

$$c_b c_1 (1 + (\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} + \|\varphi\|_{\mathrm{L}^\infty(\Omega)})^r)^{1+\frac{1}{r}} \leq c_* \qquad \forall\, N \geq N_*. \tag{27}$$

We shall derive a sufficient condition for (27) to hold by replacing $\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)}$ and $\|\varphi\|_{\mathrm{L}^\infty(\Omega)}$ in (27) 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,

$$\|\hat{S}_N\|_{\mathrm{L}^\infty(\Omega)} \leq 2\|S\|_{\mathrm{L}^\infty(\Omega)} \qquad \forall\, N \geq N_{**}.$$

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 condition for (27) to hold is that

$$c_b c_1 (1 + (2\|S\|_{\mathrm{L}^\infty(\Omega)} + C_{\mathrm{inv}} N^{\frac{d}{2}} \|\varphi\|_{\mathrm{L}^2(\Omega)})^r)^{1+\frac{1}{r}} \leq c_* \qquad \forall\, N \geq N_*, \tag{28}$$

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_*, (28) becomes equivalent to the inequality

$$N^{\frac{d}{2}} \|\varphi\|_{\mathrm{L}^2(\Omega)} \leq \|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)} \qquad \forall\, N \geq N_*. \tag{29}$$

As $\varphi \in \mathfrak{B}_{N,0}$, a sufficient condition for (29) to hold is that

$$c_* N^{\frac{d}{2}-s} \leq 1 \qquad \forall\, N \geq N_*. \tag{30}$$

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, (30) holds; and, therefore, (29), (28), (27) all hold, and, since (27) is equivalent to (26), it follows that (26) also holds. Having shown the existence of c_* and N_* such that (26) holds, it follows from (25) 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

$$\int_\Omega \int_0^1 G(\hat{S}_N + \theta\varphi_i; S_{N,\varphi_i}, T_N)\,\mathrm{d}\theta\,\mathrm{d}x = \ell(T_N) \qquad \forall\, T_N \in \Sigma_{N,0}. \tag{31}$$

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 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,

$$\delta_1=\min(1,r), \quad \delta_2<1-\varepsilon, \quad \delta_3=\min(1,r-2+2\varepsilon), \quad \delta_4=\min(1,r).$$

These statements follow from Lemma 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,

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

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

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

Hence, for $(S,D(u))\in[\mathrm{H}^s_\#(\Omega)]^{d\times d}\times[\mathrm{H}^s_\#(\Omega)]^{d\times d}$ 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.,

$$\int_\Omega\int_0^1 G(\hat{S}_N+\theta S_{N,*};S_{N,*},T_N)\,\mathrm{d}\theta\,\mathrm{d}x=\ell(T_N) \qquad \forall\,T_N\in\Sigma_{N,0}. \tag{32}$$

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 (32) is equivalent to

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

Recall from (17) that, on the other hand,

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

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, 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

$$S_{N,0} := S_N - \hat{S}_N = S_N - P_N S,$$

such that (17) holds. Now, by the triangle inequality and (10), and because $S_{N,0} \in \mathfrak{B}_{N,0}$, we have that

$$\begin{aligned}
\|S - S_N\|_{\mathrm{L}^2(\Omega)} &\leq \|S - P_N S\|_{\mathrm{L}^2(\Omega)} + \|S_{N,0}\|_{\mathrm{L}^2(\Omega)} \\
&\leq c_1 N^{-s} \|S\|_{\mathrm{H}^s(\Omega)} + c_* N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right) \qquad (33)\\
&\leq (c_1 + c_*) N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right) \qquad \forall\, N \geq N_*. \qquad (34)
\end{aligned}$$

Further, by (14), (P2b) and (33), we have that, for all $N \geq N_*$,

$$\begin{aligned}
\|P_N D(u) - D(u_N)\|_{\mathrm{L}^2(\Omega)} &= \sup_{T_N \in \Sigma_N \setminus \{0\}} \frac{(P_N D(u) - D(u_N), T_N)}{\|T_N\|_{\mathrm{L}^2(\Omega)}} \\
&= \sup_{T_N \in \Sigma_N \setminus \{0\}} \frac{(D(u) - D(u_N), T_N)}{\|T_N\|_{\mathrm{L}^2(\Omega)}} \\
&= \sup_{T_N \in \Sigma_N \setminus \{0\}} \frac{(F(S) - F(S_N), T_N)}{\|T_N\|_{\mathrm{L}^2(\Omega)}} \\
&\leq \|F(S) - F(S_N)\|_{\mathrm{L}^2(\Omega)} \leq c_b \|S - S_N\|_{\mathrm{L}^2(\Omega)} \\
&\leq c_b c_1 N^{-s} \|S\|_{\mathrm{H}^s(\Omega)} + c_b c_* N^{-s} \left(\|S\|_{\mathrm{H}^s(\Omega)} + \|D(u)\|_{\mathrm{H}^s(\Omega)}\right). \qquad (35)
\end{aligned}$$

From (35), by the triangle inequality and noting that $c_b \geq 1$, it follows that, for all $N \geq N_*$,

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

That completes the proof. □

Remark 2. *We note that by Korn's inequality (cf. Lemma 4),*

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

For each $N \geq 1$, 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 explore 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 $k = 1, 2, \dots$, we define $(S_N^k, u_N^k) \in \Sigma_N \times V_N$ as the solution of the following problem

$$-(\operatorname{div} S_N^k, v_N) = (f, v_N) \qquad \forall\, v_N \in V_N, \qquad (36)$$

$$(S_N^k, T_N) - \lambda(D(u_N^k), T_N) = (S_N^{k-1}, T_N) - \lambda(F(S_N^{k-1}), T_N) \qquad \forall\, T_N \in \Sigma_N, \qquad (37)$$

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 (36), (37). To this end, let $S_{N,0}^k := S_N^k - S_N^{k-1}$, and note that

$$(\operatorname{div} S_{N,0}^k, v_N) = 0 \qquad \forall\, v_N \in V_N, \qquad (38)$$

$$(S_{N,0}^k, T_N) - \lambda(D(u_N^k), T_N) = -\lambda(F(S_N^{k-1}), T_N) \qquad \forall\, T_N \in \Sigma_N. \qquad (39)$$

Hence, $S_{N,0}^k \in \Sigma_{N,0}$, and therefore,

$$(S_{N,0}^k, T_N) = -\lambda(F(S_N^{k-1}), T_N) \qquad \forall\, T_N \in \Sigma_{N,0}.$$

Consequently, $S^k_{N,0}$ is uniquely defined as the orthogonal projection of $-\lambda F(S^{k-1}_N)$ 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 $S^k_N = S^{k-1}_N + S^k_{N,0} \in \Sigma_N$. For S^k_N thus fixed, we rewrite (37) as follows:

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

By introducing the bilinear form $b(v,T) := -(v, \operatorname{div} T)$ on $V_N \times \Sigma_N$ and the linear functional $\ell(T) := \frac{1}{\lambda}(S^k_N - S^{k-1}_N, T) + (F(S^{k-1}_N), T)$ on Σ_N, the proof of existence of a unique solution u^k_N to the problem $b(u^k_N, T_N) = \ell(T_N)$ 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^k_N satisfying $b(u^k_N, T_N) = \ell(T_N)$ for all $T_N \in \Sigma_N$ therefore follows from the fundamental theorem of the theory of mixed variational problems, stated in Lemma 4.1(ii) on p.40 of Girault & Raviart [7].

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

Theorem 4. *Let*

$$c_a := \min(1, 2^{r-\frac{1}{r}}), \quad c_\diamond := 1 + (2 + C_{\mathrm{inv}} N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}}) \|S_N\|_{\mathrm{L}^\infty(\Omega)}, \quad c_0 := \frac{c_a}{c_\diamond},$$

and let $\lambda \in \left(0, \frac{1}{2}c_0\right)$. *Then,*

$$L^2 := 1 - 2c_0\lambda + 4\lambda^2 \in (0,1),$$

and, for each $k \geq 1$,

$$\|S_N - S^k_N\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u^k_N)\|^2_{\mathrm{L}^2(\Omega)} \leq L^{2k} \|S_N\|^2_{\mathrm{L}^2(\Omega)} \qquad \forall k \geq 1,$$

Proof. We subtract (36), (37) from (13), (14), respectively; hence,

$$(\operatorname{div}(S_N - S^k_N, v_N) = 0 \qquad \forall\, v_N \in V_N, \tag{40}$$

$$\begin{aligned}(S_N - S^k_N, T_N) &= (S_N - S^{k-1}_N, T_N) - \lambda(F(S_N) - F(S^{k-1}_N), T_N)\\ &\quad + \lambda(D(u_N - u^k_N), T_N) \qquad \forall\, T_N \in \Sigma_N.\end{aligned} \tag{41}$$

Equation (40) implies that $S_N - S^k_N \in \Sigma_{N,0}$; thus, by taking $T_N = S_N - S^k_N$ in (41), we have that

$$\|S_N - S^k_N\|^2_{\mathrm{L}^2(\Omega)} = (S_N - S^{k-1}_N, S_N - S^k_N) - \lambda(F(S_N) - F(S^{k-1}_N), S_N - S^k_N). \tag{42}$$

Next, we take $T_N = S_N - S^{k-1}_N$ in (41); hence,

$$(S_N - S^k_N, S_N - S^{k-1}_N) = \|S_N - S^{k-1}_N\|^2_{\mathrm{L}^2(\Omega)} - \lambda(F(S_N) - F(S^{k-1}_N), S_N - S^{k-1}_N). \tag{43}$$

Finally, we take $T_N = P_N(F(S_N) - F(S^{k-1}_N))$ in (41); thus,

$$\begin{aligned}(S_N - S^k_N, F(S_N) - F(S^{k-1}_N)) &= (S_N - S^k_N, P_N(F(S_N) - F(S^{k-1}_N)))\\ &= (S_N - S^{k-1}_N, P_N(F(S_N) - F(S^{k-1}_N)))\\ &\quad - \lambda(F(S_N) - F(S^{k-1}_N), P_N(F(S_N) - F(S^{k-1}_N)))\\ &\quad + \lambda(D(u_N - u^k_N), P_N(F(S_N) - F(S^{k-1}_N)))\\ &= (S_N - S^{k-1}_N, F(S_N) - F(S^{k-1}_N))\\ &\quad - \lambda(F(S_N) - F(S^{k-1}_N), P_N(F(S_N) - F(S^{k-1}_N)))\\ &\quad + \lambda(D(u_N - u^k_N), F(S_N) - F(S^{k-1}_N)).\end{aligned} \tag{44}$$

Substitution of (43) and (44) into (42) yields

$$\begin{aligned}\|S_N - S^k_N\|^2_{\mathrm{L}^2(\Omega)} &= \|S_N - S^{k-1}_N\|^2_{\mathrm{L}^2(\Omega)} - \lambda(F(S_N) - F(S^{k-1}_N), S_N - S^{k-1}_N)\\ &\quad - \lambda(S_N - S^{k-1}_N, F(S_N) - F(S^{k-1}_N))\\ &\quad + \lambda^2(F(S_N) - F(S^{k-1}_N), P_N(F(S_N) - F(S^{k-1}_N)))\\ &\quad - \lambda^2(D(u_N - u^k_N), F(S_N) - F(S^{k-1}_N)).\end{aligned} \tag{45}$$

We shall transform the final term in (45) by taking $T_N = D(u_N - u_N^k)$ in (41):

$$\lambda \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} = (S_N - S_N^k, D(u_N - u_N^k)) - (S_N - S_N^{k-1}, D(u_N - u_N^k)) + \lambda(F(S_N) - F(S_N^{k-1}), D(u_N - u_N^k)). \tag{46}$$

As the first two terms on the right-hand side of (46) are both equal to 0 and $\lambda > 0$, it follows that

$$(D(u_N - u_N^k), F(S_N) - F(S_N^{k-1})) = \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)}. \tag{47}$$

Substituting (47) into (45), we arrive at the following identity:

$$\begin{aligned}
&\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad = \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}) \\
&\qquad + \lambda^2 (F(S_N) - F(S_N^{k-1}), P_N(F(S_N) - F(S_N^{k-1}))).
\end{aligned} \tag{48}$$

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

$$\begin{aligned}
&\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad = \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}) \\
&\qquad + \lambda^2 \|F(S_N) - F(S_N^{k-1})\|_{\mathrm{L}^2(\Omega)} \|P_N(F(S_N) - F(S_N^{k-1}))\|_{\mathrm{L}^2(\Omega)} \\
&\quad \leq \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}) \\
&\qquad + \lambda^2 \|F(S_N) - F(S_N^{k-1})\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad \leq (1 + 4\lambda^2) \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)} - 2\lambda(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}).
\end{aligned} \tag{49}$$

We focus our attention on the second term on the right-hand side of (49).

Thanks to Lemma 2,

$$\begin{aligned}
(F(S_N) - F(S_N^{k-1}), S_N - S_N^{k-1}) &\geq c_a \int_\Omega \frac{|S_N - S_N^{k-1}|^2}{(1 + |S_N| + |S_N^{k-1}|)^{r+1}} \,\mathrm{d}x \\
&\geq \frac{c_a}{1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^{k-1}\|_{\mathrm{L}^\infty(\Omega)}} \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)},
\end{aligned} \tag{50}$$

where $c_a = \min\left(1, 2^{r - \frac{1}{r}}\right)$. As $S_N^0 := 0$, there exists a positive constant $c_\diamond$, independent of k (but possibly dependent on N), such that

$$1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^0\|_{\mathrm{L}^\infty(\Omega)} \leq c_\diamond.$$

Suppose, for induction, that we have already shown that

$$1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^m\|_{\mathrm{L}^\infty(\Omega)} \leq c_\diamond \qquad \forall\, m \in \{0, \dots, k-1\}, \tag{51}$$

for some $k \geq 1$. It then follows from (50) and (51) that

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

with $c_0 := \frac{c_a}{c_\diamond}$. Substituting this into the right-hand side of (49) we deduce that

$$\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \leq (1 - 2c_0\lambda + 4\lambda^2) \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)}. \tag{52}$$

Let us choose

$$\lambda \in \left(0, \tfrac{1}{2} c_0\right).$$

Then,

$$L^2 := 1 - 2c_0\lambda + 4\lambda^2 \in (0, 1).$$

Consequently, (52) yields

$$\|S_N - S_N^k\|^2_{\mathrm{L}^2(\Omega)} + \lambda^2 \|D(u_N - u_N^k)\|^2_{\mathrm{L}^2(\Omega)} \leq L^2 \|S_N - S_N^{k-1}\|^2_{\mathrm{L}^2(\Omega)}, \qquad L \in (0, 1). \tag{53}$$

In order to complete the inductive step, it remains to show that (51) holds for all $m \in \{0, \dots, k\}$, $k \geq 1$. To this end, we note that (53) implies that

$$\|S_N - S_N^k\|_{\mathrm{L}^2(\Omega)} \leq L^k \|S_N - S_N^0\|_{\mathrm{L}^2(\Omega)} = L^k \|S_N\|_{\mathrm{L}^2(\Omega)}. \tag{54}$$

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

$$\begin{aligned} \|S_N^k\|_{\mathrm{L}^\infty(\Omega)} &\leq \|S_N - S_N^k\|_{\mathrm{L}^\infty(\Omega)} + \|S_N\|_{\mathrm{L}^\infty(\Omega)} \\ &\leq C_{\mathrm{inv}} N^{\frac{d}{2}} L^k \|S_N\|_{\mathrm{L}^2(\Omega)} + \|S_N\|_{\mathrm{L}^\infty(\Omega)}. \end{aligned} \tag{55}$$

Hence,

$$\begin{aligned} 1 + \|S_N\|_{\mathrm{L}^\infty(\Omega)} + \|S_N^k\|_{\mathrm{L}^\infty(\Omega)} &\leq 1 + 2\|S_N\|_{\mathrm{L}^\infty(\Omega)} + C_{\mathrm{inv}} N^{\frac{d}{2}} L^k \|S_N\|_{\mathrm{L}^2(\Omega)} \\ &\leq 1 + (2 + C_{\mathrm{inv}} L^k N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}}) \|S_N\|_{\mathrm{L}^\infty(\Omega)} \\ &\leq 1 + (2 + C_{\mathrm{inv}} N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}}) \|S_N\|_{\mathrm{L}^\infty(\Omega)}. \end{aligned} \tag{56}$$

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 $c_\diamond$, (51) holds with $k-1$ replaced by k, which then completes the inductive step. In particular, this implies that (53), and therefore also (54), holds for all $k \geq 1$.

Thus, from (53) and (54) we deduce that

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

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 3. *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* $C_{\mathrm{inv}} L^k N^{\frac{d}{2}} |\Omega|^{\frac{1}{2}} \leq 1$ *for all* $k \geq k_0$. *For example, one can take*

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

Using this refined upper bound in (56) *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 3 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 3, we can further redefine* $c_\diamond$ *as*

$$c_\diamond := 1 + \tfrac{1}{2}c_a + 6\|S\|_{\mathrm{L}^\infty(\Omega)},$$

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* (36), (37) *is independent of* N, *provided that* $(S, D(u)) \in [\mathrm{H}^s_\#(\Omega)]^{d\times d} \times [\mathrm{H}^s_\#(\Omega)]^{d\times d}$ *with* $s > \frac{d}{2}$.

6. Numerical experiments

Throughout this section we shall report some concrete examples which are useful to compare the behaviour of the numerical simulations with the theoretical results previously shown in the paper. First of all we shall consider 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, meaning that we can solve it without using an iterative method. Then we can estimate the numerical rate of convergence of the discrete solution (S_N, u_N) to the analytical one and compare it with the one stated in Theorem 3. Our second example will be split into two parts: the aim of the first one is to test the convergence of the iterative method (proved in section 5, Theorem 4) in a concrete application; in the last one we shall consider a more complicated example where we cannot evaluate the exact solution of the problem.

All the examples that we chose to report are simplified cases of the 3-dimensional problem (1)-(2). We will name these two applications 1D example and 2D example respectively, and the reason for this choice is that both the load function and the variables depend on one and two spacial coordinates, respectively. Being in the 3-dimensional case, our unknown quantities, the stress tensor S and the displacement u (belonging to suitable function spaces, cf. Theorem 2), are respectively a 3×3 symmetric matrix and a 3-component vector.

It is worth noting that, in order to perform our numerical simulations, we have to choose a specific value of the parameter r of the model. Our work in this paper is still far from a concrete engineering application (because of all our assumptions, for example the hypothesis of periodicity, which we have explained in section 2), which could require a specific value of r because of the properties of the material, etc.; our choice was to fix $r = 0.5$ in all our simulations, and the only explanation of it is that, with that value of r, the existence and uniqueness of a weak solution to our 3-dimensional problem (as it was stated in Theorem 2) was proved in [3]. Anyway we want to highlight that the following conceptual analysis is still valid for all $r \in (0, \infty)$, even if the outputs are obtained with $r = 0.5$.

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 $f = (0, 0, f_3(x))$, 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 with the 1D body lying horizontally, and the force acting vertically, the strength (but not the direction) of the force being dependent on the horizontal location x.

Thanks to the assumptions which we have made in this example and looking for a displacement vector u of the type $u(x) = (0, 0, u_3(x))$, the strong formulation (1)-(2) of the problem becomes:

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

with 2π-periodic boundary conditions. We have denoted with the symbol S_{ij} the entry of the matrix S in position (i, j).

6.1.1. *Weak formulation, discrete problem and algebraic interpretation.* In order to perform some numerical simulations we needed to derive the algebraic interpretation of the weak formulation of our problem.

To avoid notational clutter let:

$$S = S_{13}, \qquad u = u_3, \qquad f = f_3.$$

Our 1D example in its strong formulation thus is:

$$\begin{cases} -S(x) = f(x)\ , \\ \frac{1}{2}u(x) = \frac{S(x)}{(1+|\sqrt{2}\, S(x)|^r)^{\frac{1}{r}}} = F_1(S(x))\ , \end{cases} \tag{57}$$

where $x \in \Omega$ and $|\cdot|$ stands for the modulus operation.

The weak formulation of the previous problem can be stated as follows: find $(S, u) \in \Sigma \times V$ such that

$$\begin{cases} F_1(S) = \frac{1}{2}u, \\ (S, v) = (f, v) & \forall v \in V, \end{cases} \tag{58}$$

where $\Sigma := L^1_*(\Omega)$, $V := \{\omega \in L^1_\#(\Omega) : \omega' \in L^\infty_\#(\Omega), \int_\Omega \omega(x)\,\mathrm{d}x = 0\}$ (note that V is the 1-dimensional version of the space $D^1_{*,\infty}(\Omega)$ defined in section 3.3). Under the assumption $f \in W^{1,t}_*(\Omega)$ for some $t > 1$, Theorem 2 guarantees the existence of a unique solution to (58).

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

$$\begin{cases} (F_1(S_N) - \frac{1}{2}u_N, T_N) = 0 & \forall T_N \in \Sigma_N, \\ (S_N, v_N) = (f, v_N) & \forall v_N \in V_N, \end{cases} \tag{59}$$

where $\Sigma_N := \mathcal{S}_N$, $V_N := \mathcal{S}_N$ and $\mathcal{S}_N$ is the space of the 2π-periodic real-valued trigonometric polynomials of degree N with integral average over the set Ω equal to 0. Note that in this particular case the spaces Σ_N and V_N coincide but we prefer to denote them with two different symbols in order to use the same notation throughout the paper. We can characterise the space $\mathcal{S}_N$ as follows:

$$\mathcal{S}_N = \mathrm{span}\{\, \varphi_k^s(x),\ 1 \le k \le N\} \cup \mathrm{span}\{\, \varphi_k^c(x),\ 1 \le k \le N\}, \tag{60}$$

where $\varphi_k^s(x) := \sin(kx)$ and $\varphi_k^c(x) := \cos(kx)$.

Regarding the discrete problem (59), we can state the following Fourier expansions of our unknown quantities:

$$S_N(x) = \sum_{k=1}^{N} \tilde{S}_k^s \varphi_k^s(x) + \sum_{k=1}^{N} \tilde{S}_k^c \varphi_k^c(x), \qquad u_N(x) = \sum_{k=1}^{N} \tilde{u}_k^s \varphi_k^s(x) + \sum_{k=1}^{N} \tilde{u}_k^c \varphi_k^c(x), \tag{61}$$

with $x \in \Omega$.

In order to derive the algebraic interpretation of the system (59), considering the discrete Fourier expansions (61) and taking the basis functions of $\mathcal{S}_N$ as test functions, we eventually obtain the following system:

$$\begin{cases} l\,\pi\,\tilde{S}_l^c = (f, \sin(lx)) & \forall l = 1,\dots,N, \\ -m\,\pi\,\tilde{S}_m^s = (f, \cos(mx)) & \forall m = 1,\dots,N, \\ -\frac{1}{2}\,p\,\pi\,\tilde{u}_p^c = (F_1(S_N), \sin(px)) & \forall p = 1,\dots,N, \\ \frac{1}{2}\,q\,\pi\,\tilde{u}_q^s = (F_1(S_N), \cos(qx)) & \forall q = 1,\dots,N. \end{cases} \tag{62}$$

Note that the previous system is decoupled: we can first solve $(62)_1$ and $(62)_2$ to get $\tilde{S}_k^s$ and $\tilde{S}_k^c$ (thus the solution S_N) and then consider $(62)_3$ and $(62)_4$ finding u_N. Therefore, once we have fixed the body force f, the integrals we have to evaluate in order to find the discrete solution are the right-hand sides of the system (62). The structure of the terms f and $F_1(S_N)$ usually makes it difficult to compute exactly those integrals: we shall use suitable quadrature rules.

6.1.2. *Numerical simulations in a simple case with $r = 0.5$.* As it can be easily understood, it is generally challenging to know the analytical solution of our problem, but in this $1D$ example we can choose a specific f that allows us to "calculate" analytically the exact solution (the meaning of the inverted commas will be clear in a while). Thus, let us consider a simple example where we know the continuous solution: we want to compare it with the numerical one and then estimate the numerical error behaviour.

Focusing on the strong formulation (57) of our problem, we can take the absolute value of $(57)_2$ and then raise it to the r-th power, to deduce that:

$$|\sqrt{2}\,S|^r = \frac{|u|^r}{2^{\frac{r}{2}} - |u|^r}.$$

Substituting the previous expression in $(57)_2$, we obtain

$$S = \frac{1}{\sqrt{2}}\,\frac{u}{(2^{\frac{r}{2}} - |u|^r)^{\frac{1}{r}}}. \tag{63}$$

Using the last equality together with $(57)_1$, it follows that

$$f = -\frac{1}{\sqrt{2}}\left(\frac{u}{(2^{\frac{r}{2}} - |u|^r)^{\frac{1}{r}}}\right). \tag{64}$$

Now, rather than fixing an f and calculating the analytical solution, we fix the displacement u and thanks to (64) we get the right-hand side which corresponds to such a displacement. Note that, given such a u, the exact stress tensor component S is given by (63). In this way, after the previous calculation of f, we have found an example where we know the exact solution and therefore we can compare our numerical simulations with it.

Take $u = \sin(x)$ and note that f and S are always well defined for all $x \in \Omega$ (the denominators in (63) and (64) are different from zero for all $r > 0$). It follows that:

$$f = 2^{\frac{r-1}{2}}\,\sin x\,(2^{\frac{r}{2}} - |\cos x|^r)^{\frac{-r-1}{r}}.$$

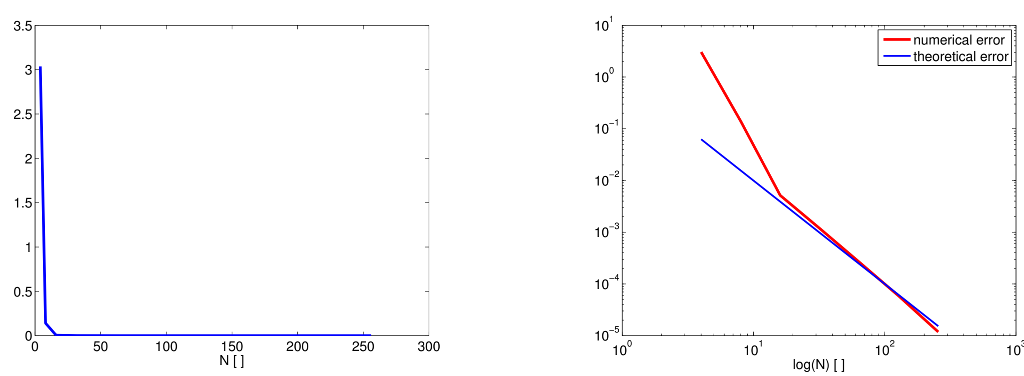
Using the previous right-hand side f, our exact solution is thus the following:

$$S = \frac{\cos x}{\sqrt{2}\,(2^{\frac{r}{2}} - |\cos x|^r)^{\frac{1}{r}}},$$
$$u = \sin x.$$

We have thus estimated the exact solution to the problem (57), which is also solution to the weak formulation (58) (and we know that this is the unique weak solution in the case of $r = 0.5$ by Theorem 2). Now we are ready to show the comparison between our analytical solution and the numerical one: the latter has been calculated by the system (62), where all the right-hand sides have been approximated with a global adaptive quadrature rule, i.e. the MATLAB [11] 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)-(12)) in terms of N: the error decreases quickly to zero as N increases. Knowing the exact solution, this example gives us the possibility to evaluate not only the error evolution in terms of N but also to estimate the numerical rate of convergence in comparison with the theoretical one, which was stated in Theorem 3. Thus, starting from $N = 4$ until $N = 256$, we reported (in red) the sum of the two L^2-norm errors of S_N and u_N with respect to the degree N in Figure 1B, where we have used the logarithmic scale on both axes. Noting the regularity of the analytical solution, i.e., $S \in H^2_*(\Omega)$ and $u \in C^\infty_*(\Omega)$, we expect that the rate of convergence is around 2 from the theoretical estimates (11)-(12). From Figure 1B it is clear that the numerical error decreases with the rate N^{-2} expected from the theory.
Furthermore, in Figure 1B, we can observe that the numerical result is a good estimation of the theoretical one only when the parameter N exceeds a given threshold: this is due to the fact that the estimations (11)-(12) proved in Theorem 3 holds for all $N \geq N_*$.

Finally, it is worth making a comment about the fact that we have compared a theoretical estimation (Theorem 3) proved for a Fourier spectral Galerkin method with a numerical one where we have used a quadrature rule to integrate the right-hand sides of the system (62): we are aware of the fact that we are not considering the quadrature error, but it is also clear from Figure 1B that this error is negligible compared with the approximation error.



(A) Sum of the L^2-norm errors in terms of N (r=0.5)

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

FIGURE 1. Numerical error behaviour

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 with the structure $f = (0, 0, f_3(x, y))$, 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" circumstance where a vertical force acts on the 2D horizontal body and the strength (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 type $u(x,y) = (0,0,u_3(x,y))$, the strong formulation (1)-(2) of our problem becomes:

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

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

6.2.1. *Weak formulation, discrete problem and algebraic interpretation.* As in the first example, we derive the algebraic system starting from the weak formulation with the aim of performing some numerical experiments.

In order to avoid notational clutter let:

$$S_1 = S_{13}, \qquad S_2 = S_{23}, \qquad u = u_3, \qquad f = f_3.$$

Our 2D model problem in strong form thus becomes:

$$\begin{cases} \frac{1}{2}(u(x,y))_x = \frac{S_1(x,y)}{(1+(2\,S_1^2(x,y)+2\,S_2^2(x,y))^{\frac{r}{2}})^{\frac{1}{r}}}, \\ \frac{1}{2}(u(x,y))_y = \frac{S_2(x,y)}{(1+(2\,S_1^2(x,y)+2\,S_2^2(x,y))^{\frac{r}{2}})^{\frac{1}{r}}}, \\ -(S_1(x,y))_x - (S_2(x,y))_y = f(x,y), \end{cases} \tag{65}$$

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

The weak formulation of the continuous problem is the following: find $(S_1, S_2, u) \in \Sigma^2 \times V$ such that

$$\begin{cases} \frac{S_1}{(1+(2\,S_1^2+2\,S_2^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}u_x = 0, \\ \frac{S_2}{(1+(2\,S_1^2+2\,S_2^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}u_y = 0, \\ ((S_1,S_2)^T, \nabla v) = (f, v) \qquad \forall v \in V, \end{cases} \tag{66}$$

where $\Sigma := L^1_*(\Omega)$, $V := \{\omega \in L^1_\#(\Omega) : \omega' \in L^\infty_\#(\Omega),\ \int_\Omega \omega(x)\,\mathrm{d}x = 0\}$. The existence and uniqueness of the solution to the previous problem is guaranteed by Theorem 2, under the assumption $f \in W^{1,t}_*(\Omega)$ for some $t > 1$.

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

$$\begin{cases} \left(\frac{S_{1,N}}{(1+(2\,S_{1,N}^2+2\,S_{2,N}^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}(u_N)_x, T_N\right) = 0 \qquad \forall T_N \in \Sigma_N, \\ \left(\frac{S_{2,N}}{(1+(2\,S_{1,N}^2+2\,S_{2,N}^2)^{\frac{r}{2}})^{\frac{1}{r}}} - \frac{1}{2}(u_N)_y, T_N\right) = 0 \qquad \forall T_N \in \Sigma_N, \\ ((S_{1,N}, S_{2,N})^T, \nabla v_N) = (f, v_N) \qquad \forall v_N \in V_N, \end{cases} \tag{67}$$

where $\Sigma_N := \mathcal{S}_N^2$, $V_N = \mathcal{S}_N^2$ and $\mathcal{S}_N^2$ is the space of the 2π-periodic real-valued trigonometric polynomials of degree N with integral average over the set Ω equal to 0 (the generalisation of the set defined in (60), now the domain Ω is not an interval but a square). As usual by the expression "2π-periodic" we mean 2π-periodic in each of the two co-ordinate directions. We can characterise the space $\mathcal{S}_N^2$ as follows:

$$\mathcal{S}_N^2 = \operatorname{span}\{\,\varphi_k^s(x,y),\ \forall\, k \in K_N\} \cup \operatorname{span}\{\,\varphi_k^c(x,y),\ \forall\, k \in K_N\},$$

where

$$
\begin{aligned}
&K_N := \{m = (m_1, m_2) \in \mathbb{R}^2 : 0 \le m_i \le N,\, \forall i = 1,2, m \ne (0,0)^T\},\\
&\varphi_k^s(x,y) := \sin(k \cdot (x,y)^T),\\
&\varphi_k^c(x,y) := \cos(k \cdot (x,y)^T),
\end{aligned}
$$

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

As regards the problem (67), the Fourier expansions of our unknown quantities are:

$$S_{1,N}(x,y) = \sum_{k\in K_N} \tilde{S}^s_{1,k}\, \varphi^s_k(x,y) + \sum_{k\in K_N} \tilde{S}^c_{1,k}\, \varphi^c_k(x,y), \tag{68}$$

$$S_{2,N}(x,y) = \sum_{k\in K_N} \tilde{S}^s_{2,k}\, \varphi^s_k(x,y) + \sum_{k\in K_N} \tilde{S}^c_{2,k}\, \varphi^c_k(x,y), \tag{69}$$

$$u_N(x,y) = \sum_{k\in K_N} \tilde{u}^s_k\, \varphi^s_k(x,y) + \sum_{k\in K_N} \tilde{u}^c_k\, \varphi^c_k(x,y), \tag{70}$$

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

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 algebraic linear system that we need to solve at each step of our iterative method. Let us consider the linearization of (67) which we have discussed in section 5, obtaining the analogue of (36)-(37) for our $2D$ example.

Given the initial guess $S^0_{1,N} \equiv 0, S^0_{2,N} \equiv 0$, find $(S^n_{1,N}, S^n_{2,N}, u^n_N) \in \Sigma^2_N \times V_N \quad \forall n = 1,2,\ldots$ (suitable stopping criterion) such that

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

For the rest of this paragraph we will consider the parameter λ to be given: we will get back to the problem of choosing λ in the paragraphs dedicated to the numerical experiments.

Substituting the discrete Fourier expansions (68), (69), (70) and considering the basis functions of the space $\mathcal{S}^2_N$ as test functions, after some calculations we have the following system:

$$
\begin{cases}
2\,\pi^2\,\tilde{S}^{s,n}_{1,p} + \lambda\, p_1\, \pi^2\, \tilde{u}^{c,n}_p = f^{s,n-1}_{1,p} & \forall p \in K_N,\\
2\,\pi^2\,\tilde{S}^{c,n}_{1,q} - \lambda\, q_1\, \pi^2\, \tilde{u}^{s,n}_q = f^{c,n-1}_{1,q} & \forall q \in K_N,\\
2\,\pi^2\,\tilde{S}^{s,n}_{2,t} + \lambda\, t_2\, \pi^2\, \tilde{u}^{c,n}_t = f^{s,n-1}_{2,t} & \forall t \in K_N,\\
2\,\pi^2\,\tilde{S}^{c,n}_{2,z} - \lambda\, z_2\, \pi^2\, \tilde{u}^{s,n}_z = f^{c,n-1}_{2,z} & \forall z \in K_N,\\
2\, l_1\, \pi^2\, \tilde{S}^{c,n}_{1,l} + 2\, l_2\, \pi^2\, \tilde{S}^{c,n}_{2,l} = f^s_l & \forall\, l \in K_N,\\
2\, m_1\, \pi^2\, \tilde{S}^{s,n}_{1,m} + 2\, m_2\, \pi^2\, \tilde{S}^{s,n}_{2,m} = f^c_m & \forall\, m \in K_N,
\end{cases}
\tag{71}
$$

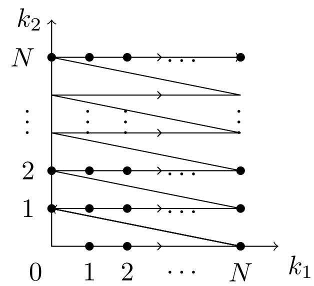


FIGURE 2. New notation where each dot represents one and only one vector $k \in K_N$ such that $k = (k_1, k_2)$.

where the right-hand sides of such a system are defined as:

$$\begin{aligned}
f_{1,p}^{s,n-1} &= \left(S_{1,N}^{n-1} - \lambda \frac{S_{1,N}^{n-1}}{(1 + (2\,(S_{1,N}^{n-1})^2 + 2\,(S_{2,N}^{n-1})^2)^{\frac{r}{2}})^{\frac{1}{r}}}, \varphi_p^s \right)_N, \\
f_{1,p}^{c,n-1} &= \left(S_{1,N}^{n-1} - \lambda \frac{S_{1,N}^{n-1}}{(1 + (2\,(S_{1,N}^{n-1})^2 + 2\,(S_{2,N}^{n-1})^2)^{\frac{r}{2}})^{\frac{1}{r}}}, \varphi_p^c \right)_N, \\
f_{2,p}^{s,n-1} &= \left(S_{2,N}^{n-1} - \lambda \frac{S_{2,N}^{n-1}}{(1 + (2\,(S_{1,N}^{n-1})^2 + 2\,(S_{2,N}^{n-1})^2)^{\frac{r}{2}})^{\frac{1}{r}}}, \varphi_p^s \right)_N, \\
f_{2,p}^{c,n-1} &= \left(S_{2,N}^{n-1} - \lambda \frac{S_{2,N}^{n-1}}{(1 + (2\,(S_{1,N}^{n-1})^2 + 2\,(S_{2,N}^{n-1})^2)^{\frac{r}{2}})^{\frac{1}{r}}}, \varphi_p^c \right)_N, \\
f_l^s &= (f, \varphi_l^s)_N, \\
f_m^c &= (-f, \varphi_m^c)_N,
\end{aligned}$$

with $p, l, m \in K_N$, the parameter n refers to the n-th step of our iterative method and the symbol $(\cdot)_N$ stands for numerical integration.

At this point we need to find a new notation to indicate the vectors $k \in K_N$ in order to deal with scalar subscripts: in this way we can then easily obtain the algebraic formulation of our n-th step of the iterative method. Ordering the vectors $k \in K_N$ in the sense specified by the oriented line in Figure 2, we can define a bijective function, which lets us jump from one notation to the other easily. The function

$$\begin{aligned}
w &: K_N \to D := \{1, \ldots, |K_N|\} \subset \mathbb{N} \\
w &: k \mapsto \hat{k}(k) := k_2\,(N+1) + k_1,
\end{aligned}$$

where we have used the notation $k = (k_1, k_2)$, allows us to know the scalar $\hat{k} \in D$ related to the vector $k \in K_N$. Note further that $|K_N| = (N+1)^2 - 1$ stands for the cardinality of the set K_N. On the other hand if we are interested in finding the vector $k \in K_N$ which corresponds to the scalar $\hat{k} \in D$, it is trivial to see the following:

$$\begin{aligned}
k_1(\hat{k}) &:= \hat{k} \;\texttt{mod}\,(N+1), \\
k_2(\hat{k}) &:= \hat{k} \;\texttt{div}\,(N+1),
\end{aligned}$$

where the operator `mod` returns the remainder of the division whereas the operator `div` gives the integer part of the quotient. To avoid notational clutter we will denote the scalar $\hat{k}$ simply with k in what follows.

Using this new notation, the system (71) becomes:

$$\begin{cases} 2\,\pi^2\,\tilde{S}^{s,n}_{1,p} + \lambda\, p_1(p)\,\pi^2\,\tilde{u}^{c,n}_p = f^{s,n-1}_{1,p} & \forall p = 1,\dots,|K_N|,\\ 2\,\pi^2\,\tilde{S}^{c,n}_{1,q} - \lambda\, q_1(q)\,\pi^2\,\tilde{u}^{s,n}_q = f^{c,n-1}_{1,q} & \forall q = 1,\dots,|K_N|,\\ 2\,\pi^2\,\tilde{S}^{s,n}_{2,t} + \lambda\, t_2(t)\,\pi^2\,\tilde{u}^{c,n}_t = f^{s,n-1}_{2,t} & \forall t = 1,\dots,|K_N|,\\ 2\,\pi^2\,\tilde{S}^{c,n}_{2,z} - \lambda\, z_2(z)\,\pi^2\,\tilde{u}^{s,n}_z = f^{c,n-1}_{2,z} & \forall z = 1,\dots,|K_N|,\\ 2\,l_1(l)\,\pi^2\,\tilde{S}^{c,n}_{1,l} + 2\,l_2(l)\,\pi^2\,\tilde{S}^{c,n}_{2,l} = f^s_l & \forall l = 1,\dots,|K_N|,\\ 2\,m_1(m)\,\pi^2\,\tilde{S}^{s,n}_{1,m} + 2\,m_2(m)\,\pi^2\,\tilde{S}^{s,n}_{2,m} = f^c_m & \forall m = 1,\dots,|K_N|. \end{cases} \tag{72}$$

Finally we define the following matrices and vectors of our data:

$$\begin{aligned}
&A := \texttt{diag}\,(2\,\pi^2) && \forall\, p = 1,\dots,|K_N| && \Longrightarrow && A \in \mathbb{R}^{|K_N|\times|K_N|},\\
&B := \texttt{diag}\,(p_1(p)\,\pi^2) && \forall\, p = 1,\dots,|K_N| && \Longrightarrow && B \in \mathbb{R}^{|K_N|\times|K_N|},\\
&C := \texttt{diag}\,(p_2(p)\,\pi^2) && \forall\, p = 1,\dots,|K_N| && \Longrightarrow && C \in \mathbb{R}^{|K_N|\times|K_N|},\\
&f^{s,n-1}_1 := \big[f^{s,n-1}_{1,p}\big] && \forall\, p = 1,\dots,|K_N| && \Longrightarrow && f^{s,n-1}_1 \in \mathbb{R}^{|K_N|},\\
&f^{c,n-1}_1 := \big[f^{c,n-1}_{1,p}\big] && \forall\, p = 1,\dots,|K_N| && \Longrightarrow && f^{c,n-1}_1 \in \mathbb{R}^{|K_N|},\\
&f^{s,n-1}_2 := \big[f^{s,n-1}_{2,p}\big] && \forall\, p = 1,\dots,|K_N| && \Longrightarrow && f^{s,n-1}_2 \in \mathbb{R}^{|K_N|},\\
&f^{c,n-1}_2 := \big[f^{c,n-1}_{2,p}\big] && \forall\, p = 1,\dots,|K_N| && \Longrightarrow && f^{c,n-1}_2 \in \mathbb{R}^{|K_N|},\\
&f^s := \big[f^s_l\big] && \forall\, l = 1,\dots,|K_N| && \Longrightarrow && f^s \in \mathbb{R}^{|K_N|},\\
&f^c := \big[f^c_m\big] && \forall\, m = 1,\dots,|K_N| && \Longrightarrow && f^c \in \mathbb{R}^{|K_N|},
\end{aligned}$$

where we denote with the term $\texttt{diag}(\cdot)$ a diagonal matrix with each element in position (p,p) given by the expression in parenthesis.

The unknown quantities vectors (the Fourier coefficients) are:

$$\begin{aligned}
&S^{s,n}_1 := \big[\tilde{S}^{s,n}_{1,k}\big] && \forall\, k = 1,\dots,|K_N| && \Longrightarrow && S^{s,n}_1 \in \mathbb{R}^{|K_N|},\\
&S^{c,n}_1 := \big[\tilde{S}^{c,n}_{1,k}\big] && \forall\, k = 1,\dots,|K_N| && \Longrightarrow && S^{c,n}_1 \in \mathbb{R}^{|K_N|},\\
&S^{s,n}_2 := \big[\tilde{S}^{s,n}_{2,k}\big] && \forall\, k = 1,\dots,|K_N| && \Longrightarrow && S^{s,n}_2 \in \mathbb{R}^{|K_N|},\\
&S^{c,n}_2 := \big[\tilde{S}^{c,n}_{2,k}\big] && \forall\, k = 1,\dots,|K_N| && \Longrightarrow && S^{c,n}_2 \in \mathbb{R}^{|K_N|},\\
&u^{s,n} := \big[\tilde{u}^{s,n}_k\big] && \forall\, k = 1,\dots,|K_N| && \Longrightarrow && u^{s,n} \in \mathbb{R}^{|K_N|},\\
&u^{c,n} := \big[\tilde{u}^{c,n}_k\big] && \forall\, k = 1,\dots,|K_N| && \Longrightarrow && u^{c,n} \in \mathbb{R}^{|K_N|}.
\end{aligned}$$

Finally our system (72) becomes:

$$\begin{bmatrix} A & 0 & 0 & 0 & 0 & \lambda\,B\\ 0 & A & 0 & 0 & -\lambda\,B & 0\\ 0 & 0 & A & 0 & 0 & \lambda\,C\\ 0 & 0 & 0 & A & -\lambda\,C & 0\\ 0 & 2\,B & 0 & 2\,C & 0 & 0\\ 2\,B & 0 & 2\,C & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} S^{s,n}_1\\ S^{c,n}_1\\ S^{s,n}_2\\ S^{c,n}_2\\ u^{s,n}\\ u^{c,n} \end{bmatrix} = \begin{bmatrix} f^{s,n-1}_1\\ f^{c,n-1}_1\\ f^{s,n-1}_2\\ f^{c,n-1}_2\\ f^s\\ f^c \end{bmatrix}, \tag{73}$$

or equivalently

$$E\,U^n = F^n,$$

where the $6|K_N| \times 6|K_N|$ matrix E and the $6|K_N|$-component vectors U^n and F^n are defined by (73). As before, the parameter n refers to the n-th step of the iterative method. Note that in (73) the symbol 0 stands for a $|K_N| \times |K_N|$ matrix with null entries.

Summing up, at each step n of our iterative method, we need to evaluate the right-hand side F^n and to solve (73) in order to find the Fourier coefficients $S^{s,n}_1$, $S^{c,n}_1$, $S^{s,n}_2$, $S^{c,n}_2$, $u^{s,n}$, $u^{c,n}$. We finally obtain the solution at the n-th step using the truncated Fourier expansions (68), (69) and (70).

So far we have not said anything as regards the stopping criterion of such an iterative method. Given a tolerance *tol*, a possible choice could be the following:

$$\frac{\|S_{1,N}^{n} - S_{1,N}^{n-1}\|_{L^2(\Omega)} + \|S_{2,N}^{n} - S_{2,N}^{n-1}\|_{L^2(\Omega)} + \|u_N^{n} - u_N^{n-1}\|_{L^2(\Omega)}}{\|S_{1,N}^{n-1}\|_{L^2(\Omega)} + \|S_{2,N}^{n-1}\|_{L^2(\Omega)} + \|u_N^{n-1}\|_{L^2(\Omega)}} \leq tol \tag{74}$$

6.2.2. *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 make a comparison between it and the numerical one.

Let us focus on the strong formulation (65) of the $2D$ case: consider the sum of $(65)_1^2$ and $(65)_2^2$, multiply both members of such an equality by 2 before taking the square root, and finally raise it at 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 plugging-in the previous expression in $(65)_1$ and $(65)_2$, we have that:

$$S_1 = \frac{u_x}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}, \tag{75}$$

$$S_2 = \frac{u_y}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}. \tag{76}$$

Substituting the previous expressions for S_1 and S_2 in $(65)_3$, we get the expression of the right-hand side f in terms of the displacement component u and its derivatives:

$$f = -\left(\frac{u_x}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}\right)_x - \left(\frac{u_y}{\sqrt{2}\,(2^{\frac{r}{2}} - (u_x^2 + u_y^2)^{\frac{r}{2}})^{\frac{1}{r}}}\right)_y. \tag{77}$$

Now, following the same path as in the corresponding section 6.1.2 regarding our $1D$ example, we fix the displacement u and then, thanks to (77), we obtain the analytical right-hand side f which corresponds to such a displacement. Note further that, given such a u, we can evaluate the exact components S_1 and S_2 of the stress tensor by (75) and (76).

Consider, for example, $u = \frac{1}{2}\sin(x+y)$ and note that f, as well as S_1 and S_2, is always well defined for all $x \in \Omega$ (in this second example we have to multiply the function $\sin(x+y)$ by $\frac{1}{2}$ to assure that the denominators in (75), (76) and (77) are different from zero for all $r > 0$). After some calculations, it follows that:

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

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

Concerning the choice of the parameter λ of the iterative model, Remark 3 (note that Theorem 3 holds in this example) and the knowledge of the exact solution gives us the possibility to fix the parameter λ in the correct "interval of convergence" $(0, \frac{1}{2}\,c_0)$. In this case we have (using $r = 0.5$):

$$\|S\|_{L^\infty} \leq 6, \qquad c_* = 37, \qquad c_0 = \frac{1}{37 * 2^{3/2}},$$

and we can finally choose $\lambda = \frac{1}{4}\,c_0 = \frac{1}{37*2^{7/2}} \simeq 0.002$: this is the value we used in our numerical simulation.

We now aim to test the accuracy of the numerical solution evaluated with the iterative method that we have shown before. Therefore we solve at each iteration the system (73), where the vectors which constitute the right-hand side of such a system have been approximated with a local adaptive quadrature rule (using the MATHEMATICA [10] command *NIntegrate*).

In the particular case which we chose to report here we used $N = 5$ and, knowing the exact solution, a slightly different stopping criterion (rather than (74)), which is the following

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

with $tol = 10^{-2}$. After 2 iterations the relative errors (in the L^2-norm) concerning S_1 (or S_2) and the displacement are:

$$\frac{\|S^2_{1,N} - S_1\|_{L^2(\Omega)}}{\|S_1\|_{L^2(\Omega)}} = 0.006, \qquad \frac{\|u^2_N - u\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}} = 0.001, \tag{78}$$

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

We tested our iterative method fixing the degree N and starting from different initial guesses: our method converges always 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 the solutions to the linearized problems converges to (S_N, u_N), the solution to the nonlinear spectral Galerkin problem (4)-(6), as $k \to \infty$.

Knowing the analytical solution, we can see that Theorem 3 holds and we expect that the solution $(S_{1,N}, S_{2,N}, u_N)$ converges to the continuous solution to the nonlinear problem (66) as $N \to \infty$. As one can see from (78), just with $N = 5$ the solution of the iterative method is really close to the analytical one, which confirms the theoretical result (we can therefore think that $N = 5 \geq N_*$, cf. section 4).

6.2.3. *Numerical simulations in the case of a concentrated load with* $r = 0.5$. As we have explained in section 1, the limiting strain models are characterised 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 like to take a Dirac delta function but we chose to regularize it because in this case we do not know the exact solution (thus we cannot check its regularity as we did in section 6.1.2) and Theorem 2 does not guarantee the minimal regularity that we need to use Theorem 3 in order to prove that our numerical solution converges to the analytical one as $N \to +\infty$. It could be part of a future work to show a regularity result for our continuous solution which could guarantee, with stronger (compared to the one assumed in Theorem 2) hypothesis regarding f, the minimal regularity to prove the convergence by Theorem 3. Using $f \in C^\infty_0(\Omega)$ we want to avoid any convergence problem.

For any $h > 0$ and $x \in (0, 2\pi)$ let us 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 used a global adaptive quadrature rule to approximate that constant). Note that φ_h is an approximation to the delta function concentrated at $x = \pi$, for small $h > 0$. Note further that all the functions defined above belong to all Sobolev spaces $W^{s,p}(\Omega)$ for any value $s \geq 0$ and for any value of $p \in [1, \infty]$.

In the $2D$ case we would like to approximate a $2D$ Dirac delta function concentrated at the point (π, π). Thus we can define

$$f_h(x, y) := \varphi_h(x)\varphi_h(y).$$

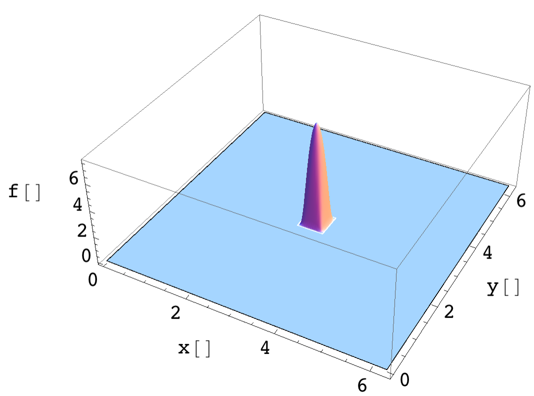
where $(x, y) \in \Omega$.

As regards the choice of the parameter λ featuring in the iterative method, we cannot repeat the same comments that we have made in section 6.2.2 because now we do not know the exact solution. Therefore the selection of the parameter λ is more critical because we do not know the exact interval $(0, \frac{1}{2}\, c_0)$ where we have to choose it. In such a case we decided to fix a small random value of λ and to run the simulation: if we obtain a plausible output our iterative method is likely working, otherwise we are not probably converging to the real solution because $\lambda > \frac{1}{2}\, c_0$. In the example which follows we made the same choice of section 6.2.2, i.e., $\lambda = \frac{1}{37 * 2^{7/2}} \simeq 0.002$ and we shall see that the method works properly.

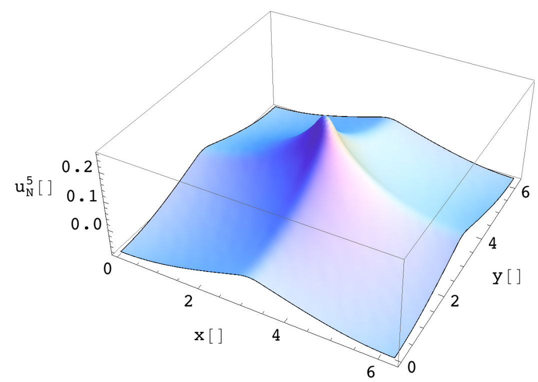
Using $f = f_h$ as defined above for a given $h > 0$, we have to solve iteratively the system (73) until the stopping criterion defined in (74) is fulfilled, using a fixed tolerance $tol = 10^{-3}$. As we did in section 6.2.2, the elements of the vectors on the right-hand side have been approximated with a local adaptive quadrature rule (the MATHEMATICA [10] command *NIntegrate*).

With the choice of the parameter $h = 0.3$, the corresponding body force f is reported in Figure 3A; fixing the degree $N = 30$, we obtain the numerical displacement shown in Figure 3B after 5 iterations. Increasing the entity of the degree N does not change the numerical solution consistently, therefore we are confident that the solution that we reported is close to the analytical one.

This example shows what we expect: the displacement (Figure 3B) has a peak in $(x, y) = (\pi, \pi)$ where the body force has a peak as well, but the magnitude of that peak is smaller than the one of 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 3. Numerical simulation with the regularized delta function as body force

7. CONCLUSIONS

This paper provides an initial step towards the rigorous mathematical analysis of numerical approximations to nonlinear elastic limiting strain models. We have constructed a spectral approximation of the model problem under consideration and have shown that the spectral method exhibits optimal order convergence. We have also proposed an iterative method for the numerical solution of the finite-dimensional system of nonlinear equations featuring in the method and have shown that the iterations converge, at a linear rate, to the unique solution of the numerical method. Our aim in future work will be to extend the results developed here to limiting strain models in general multidimensional domains, in the spirit of the PDE analysis developed in the paper [1], using a finite element method.

REFERENCES

[1] M. BULÍČEK, J. MÁLEK, K. R. RAJAGOPAL, AND E. SÜLI, *On elastic solids with limiting small strain: modelling and analysis*, EMS Surveys in Mathematical Sciences, 1 (2015). To appear.

[2] M. BULÍČEK, J. MÁLEK, K. R. RAJAGOPAL, AND J. WALTON, *Existence of solutions for the anti-plane stress for a new class of "strain-limiting" elastic bodies*. Submitted for publication, 2013.

[3] M. BULÍČEK, J. MÁLEK, AND E. SÜLI, *Analysis and approximation of a strain-limiting nonlinear model*, Mathematics and Mechanics of Solids, (2014). To appear.

[4] R. BUSTAMANTE AND K. R. RAJAGOPAL, *Solutions of some simple boundary value problems within the context of a new class of elastic materials*, International Journal of Non-Linear Mechanics, 46 (2011), pp. 376–386.

[5] C. Canuto and A. Quarteroni, *Approximation results for orthogonal polynomials in Sobolev spaces*, Math. Comp., 38 (1982), pp. 67–86.
[6] A. D. Freed and D. R. Einstein, *An implicit elastic theory for lung parenchyma*, Internat. J. Engrg. Sci., 62 (2013), pp. 31–47.
[7] V. Girault and P.-A. Raviart, *Finite Element Approximation of the Navier–Stokes Equations*, vol. 749 of Lecture Notes in Mathematics, Springer-Verlag, 1979.
[8] ———, *Finite Element Methods for Navier–Stokes Equations*, vol. 5 of Springer Ser. Comp. Math., Springer-Verlag, 1986.
[9] V. Kulvait, J. Málek, and K. Rajagopal, *Anti-plane stress state of a plate with a V-notch for a new class of elastic solids*, Int. J. Fract., (2012), pp. 1–15. doi:10.1007/s10704-012-9772-5.
[10] Mathematica, *version 9.0.1*, The Wolfram Centre, Oxfordshire OX29 8FD, 2013.
[11] MATLAB, *version 8.3 (R2014a)*, The MathWorks Inc., Natick, MA 01760, 2014.
[12] A. Ortiz, R. Bustamante, and K. R. Rajagopal, *A numerical study of a plate with a hole for a new class of elastic bodies*, Acta Mech., 223 (2012), pp. 1971–1981.
[13] K. R. Rajagopal, *On implicit constitutive theories*, Appl. Math., 48 (2003), pp. 279–319.
[14] ———, *Elasticity of elasticity*, Zeitschrift fur Angewandte Math Phys, 58 (2007), pp. 309–417.
[15] ———, *Non-linear elastic bodies exhibiting limiting small strain*, Math. Mech. Solids, 16 (2011), pp. 122–139.
[16] ———, *On a new class of models in elasticity*, J. Math. Comp. Appl., 15 (2011), pp. 506–528.
[17] K. R. Rajagopal and J. Walton, *Modeling fracture in the context of strain-limiting theory of elasticity*, Int. J. Fract., 169 (2011), pp. 39–48.

Ferrari Gestione Sportiva, Via Ascari 55/57, Maranello MO 41053, Italy
E-mail address: `nicolo.gelmetti@mail.polimi.it`

Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, United Kingdom
E-mail address: `Endre.Suli@maths.ox.ac.uk`