FIRST-ORDER SYSTEM LEAST SQUARES (FOSLS) FOR GEOMETRICALLY NONLINEAR ELASTICITY

T. A. MANTEUFFEL*, S. F. MCCORMICK*, J. G. SCHMIDT†, AND C. R. WESTPHAL‡

Abstract.
We present a first-order system least-squares (FOSLS) method to approximate the solution to the equations of geometrically nonlinear elasticity in two dimensions. With assumptions of regularity on the problem, we show H^1 equivalence of the norm induced by the FOSLS functional in the case of pure displacement boundary conditions as well as local convergence of Newton's method in a nested iteration setting. Theoretical results hold for deformations satisfying a small-strain assumption, a set we show to be largely coincident with the set of deformations allowed by the model. Numerical results confirm optimal multigrid performance and finite element approximation rates of the discrete functional with a total work bounded by about 25 fine-grid relaxation sweeps.

1. Introduction. The primary goal in the study of elasticity is to model the deformation of an elastic body under applied forces, including both internal body forces, such as gravity, and applied surface tractions. For simplicity, we consider forces whose associated density per unit volume is independent of the deformation. Under these applied forces, the elastic body is said to occupy the deformed configuration, and, in the absence of forces, the reference configuration. With this in mind, we may think of the central problem as one of finding the mapping from the reference configuration to the deformed configuration. We refer to this mapping function as the deformation and to the Jacobian of the map as the deformation gradient. Two tensor-valued physical quantities are also of interest: strain and stress. The strain tensor, a completely geometrical quantity, is purely a measure of deviation from the reference configuration, while the stress tensor is directly related to the internal force density across the deformed configuration. While the deformation itself is usually the primary unknown in the study of elasticity, the resulting stress and strain are often of interest as well. In this case, the solution methodology we describe in this paper has a distinct advantage over more traditional approaches.

The partial differential equations that are commonly used to govern the deformation are composed of two main components: the equilibrium equation and a constitutive equation. The equilibrium equation and associated boundary conditions relate a balance of forces in the deformed configuration. But, since the deformed configuration is unknown, the equation is mapped back to the reference configuration. The necessity of this mapping introduces a source of nonlinearity into the equations of elasticity.

The constitutive equation, or material law as it is sometimes called, relates the stress to the strain, taking the material properties into account. In general, a material law may be designed for a specific material in a specific range of deformations, as is often the case in applications. There can be as many material laws as materials, but we focus here on a general two-parameter linear relationship between the stress and strain. When this approximation is valid for homogenous, isotropic materials, we call them St. Venant-Kirchhoff materials. To understand the general behavior of the

*Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, CO 80309-0526. tmanteuf@colorado.edu, stevem@colorado.edu

†C&C Research Laboratories, NEC Europe Ltd., Rathausallee 10, 53757 Sankt Augustin, Germany. schmidt@ccrl-nece.de

‡Department of Mathematics and Computer Science, Wabash College, PO Box 352, Crawfordsville, IN 47933. westphac@wabash.edu

elasticity system, such materials are considered exclusively.

The model we have described here is both three-dimensional and nonlinear. In this paper, we consider the plane strain model of two-dimensional elasticity, which retains the same character as the full three-dimensional problem both physically and mathematically. It is common to linearize this problem about the reference configuration. However, inherent in the linearization of this naturally nonlinear model is the additional assumption that the displacement is small. There are many applications in which this is a valid assumption and the resulting solution remains sufficiently accurate. For example, a structure whose displacement is magnitudes of order smaller than the structure itself may be accurately modeled by this linear approximation. However, when the small displacement assumption is unreasonable, the partial differential equations of linear elasticity should be used with caution. For this reason, we choose to study a more realistic problem.

In [5, 6, 8, 16], the first-order system least-squares (FOSLS) method is applied to the equations of linear elasticity using the displacement gradient as a new variable. A suitable least-squares functional is minimized over finite element subspaces of H^1. This method allows for the displacement gradient and displacement to be approximated in a two-stage algorithm, with full H^1 control on all varaibles when the solution is sufficiently smooth. More recent methods, developed in [4, 9, 10], use the stress and displacement as primary unknowns for linear elasticity. The stress, which for linear elasticity is naturally in $H(div)$, is approximated in an $H(div)$ conforming space, thereby avoiding the need to consider effects of boundary singularities. Results from these studies show that a least-squares formulation can be effective for elasticity problems.

This leads us to consider a least-squares method for the geometrically nonlinear model of elasticity that relaxes the small displacement assumption while retaining a linear material law, thus widening the scope of problems that can be effectively treated by least-squares methods. In this model, a linear stress-strain relationship is assumed, but the full nonlinear strain-displacement relationship is preserved. Such a formulation is accurate for the so-called "large displacement, small strain" cases. While not necessarily the best model to use for a given material or for configurations with large strain, this is a common model for elastic materials, and certainly more accurate than linear elasticity. See [11] for further background in elasticity theory.

Our general approach is to linearize the equations of elasticity about a current approximation by Newton's method, to reformulate the resulting linear problem as a well-posed least-squares minimization problem, and to let its minimizer become the new approximation. The reference configuration (i.e., zero displacement) is always taken to be the initial approximation. Thus, the first Newton step reduces to the equations of linear elasticity and subsequent steps are corrections thereof. Since the constitutive equation involves products of the unknowns, we focus on using the displacement gradient as the new dependent variable. The stress and strain tensors are then just simple combinations of the this new dependent variable and can be computed in a post-processing stage with no loss of accuracy. Each Newton step is cast as an appropriate first-order system, and the associated least-squares functional is minimized over an appropriate finite element subspace of $H^1(\Omega)$.

Our approach also employs a two-stage solution process. The first stage solves for the displacement gradients, while the second stage recovers the actual displacement vector. This decoupling of the unknowns in stages is desirable for several reasons. First, when the primary interest is in the stress or strain, the second stage does

not need to performed. Second, if the problem requires several Newton steps, the deformations can be retrieved after the first stage converges. Third, this approach obviates the need to determine relative weights for the stages if they are incorporated into a single functional. Finally, decoupling the variables is somewhat more efficient than solving for them simultaneously.

We define the term H^1 ellipticity to mean H^1 equivalence with the norm induced by the homogenous FOSLS functional. The focus of much of this paper is on the formulation and efficiency of the first-stage algorithm by establishing H^1 ellipticity of the FOSLS functional for a general linearization step. The second stage is essentially a coupled Poisson problem that is ideally suited for FOSLS and already discussed in some detail in [5, 8].

2. Notation. Throughout this paper, we refer to our Newton-FOSLS algorithm as *linearized elasticity* (linearized about a current approximation) and to the first Newton step as *linear elasticity* (linearized about the reference configuration). This is not strictly standard convention, but one we find convenient in what follows.

Vector $\mathbf{u}$ and matrix $\mathbf{U}$ are represented componentwise by

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}.$$

The gradient of scalar p and vector $\mathbf{u}$ are given by

$$\nabla p = \begin{pmatrix} \partial_x p \\ \partial_y p \end{pmatrix}, \quad \text{and} \quad \nabla \mathbf{u} = \begin{pmatrix} \partial_x u_1 & \partial_y u_1 \\ \partial_x u_2 & \partial_y u_2 \end{pmatrix}.$$

Define the respective divergence, curl, and trace operators by

$$\nabla \cdot \mathbf{u} = \partial_x u_1 + \partial_y u_2, \quad \nabla \cdot \mathbf{U} = \begin{pmatrix} \partial_x U_{11} + \partial_y U_{12} \\ \partial_x U_{21} + \partial_y U_{22} \end{pmatrix},$$
$$\nabla \times \mathbf{U} = \begin{pmatrix} \partial_x U_{12} - \partial_y U_{11} \\ \partial_x U_{22} - \partial_y U_{21} \end{pmatrix}, \quad \text{and} \quad tr(\mathbf{U}) = U_{11} + U_{22}.$$

Also, denoting the formal adjoint of the curl operator by $\nabla^\perp$, we define

$$\nabla^\perp p = \begin{pmatrix} \partial_y p \\ -\partial_x p \end{pmatrix}, \quad \text{and} \quad \nabla^\perp \mathbf{u} = \begin{pmatrix} \partial_y u_1 & -\partial_x u_1 \\ \partial_y u_2 & -\partial_x u_2 \end{pmatrix}$$

We extend the respective outward unit normal and counter-clockwise unit tangential operators, $\mathbf{n}\cdot$ and $\boldsymbol{\tau}\cdot$, componentwise to block column vectors and matrices in the natural way:

$$\mathbf{n} \cdot \mathbf{U} = \begin{pmatrix} n_x U_{11} + n_y U_{12} \\ n_x U_{21} + n_y U_{22} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\tau} \cdot \mathbf{U} = \begin{pmatrix} \tau_x U_{11} + \tau_y U_{12} \\ \tau_x U_{21} + \tau_y U_{22} \end{pmatrix}.$$

We also note that $n_x = \tau_y$ and $n_y = -\tau_x$, and that $\mathbf{n} \cdot \nabla = -\boldsymbol{\tau} \cdot \nabla^\perp$.

We use standard notation for Sobolev spaces $H^k(\Omega)^d$, corresponding inner product $(\cdot,\cdot)_{k,\Omega}$, and norm $\|\cdot\|_{k,\Omega}$, for $k \geq 0$. We drop subscript Ω and superscript d when the domain and dimension are clear by context. For noninteger k, $H^k(\Omega)$ is the interpolation space between $H^{\lfloor k \rfloor}(\Omega)$ and $H^{\lceil k \rceil}(\Omega)$ as in [17]. The case of $k = 0$ corresponds to the Lebesgue measurable space, $L^2(\Omega)$, in which case we generally denote

the norm and inner product by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. Define the subspaces of $L^2(\Omega)$ induced by the divergence and the curl of vector $\mathbf{u}$ by

$$H(div) = \{\mathbf{u} \in L^2(\Omega) : \|\nabla \cdot \mathbf{u}\| < \infty\},$$
$$H(curl) = \{\mathbf{u} \in L^2(\Omega) : \|\nabla \times \mathbf{u}\| < \infty\},$$

with norms

$$\|\mathbf{u}\|^2_{H(div)} = \|\mathbf{u}\|^2 + \|\nabla \cdot \mathbf{u}\|^2,$$
$$\|\mathbf{u}\|^2_{H(curl)} = \|\mathbf{u}\|^2 + \|\nabla \times \mathbf{u}\|^2.$$

Denote by $C^k(\Omega)$ the space of k times continuously differentiable functions on Ω, an open set in $\mathbb{R}^2$. The boundary of Ω, denoted by $\partial\Omega$, is of class C^k if it satisfies the conditions of a Lipschitz boundary (see [20]) and is the union of the graphs of a finite number of C^k functions. We say that $\partial\Omega$ is a $C^{k,l}$ boundary when it is Lipschitz and is the graph of the union of a finite number of Hölder continuous $C^{k,l}$ functions.

We also make use of the following general inequalities:

$$|a|^2 + |b|^2 \leq |a+b|^2 \leq 2(|a|^2 + |b|^2). \tag{2.1}$$

3. The Nonlinear Problem. Let Ω be a bounded open connected subset of $\mathbb{R}^2$ with boundary $\partial\Omega$, which is partitioned into displacement, Γ_D, and traction, Γ_T, segments ($\bar{\Gamma}_D \cup \bar{\Gamma}_T = \partial\Omega$ and $\Gamma_D \cap \Gamma_T = \emptyset$). For simplicity, we assume that the displacements vanish on Γ_D, as is often the case in practice. The geometrically nonlinear elasticity equations may be written as

$$\begin{cases} \nabla \cdot [(\mathbf{I} + \nabla\mathbf{u})\Sigma] = \mathbf{f}, & \text{in} \quad \Omega, \\ \mathbf{n} \cdot [(\mathbf{I} + \nabla\mathbf{u})\Sigma] = \mathbf{g}, & \text{on} \quad \Gamma_T, \\ \mathbf{u} = \mathbf{0}, & \text{on} \quad \Gamma_D, \end{cases} \tag{3.1}$$

where the material law,

$$\Sigma = \Sigma(\mathbf{E}) = \lambda tr(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E}, \tag{3.2}$$

is the second Piola-Kirchhoff stress tensor and

$$\mathbf{E} = \mathbf{E}(\nabla\mathbf{u}) = \frac{\mathbf{1}}{\mathbf{2}}(\nabla\mathbf{u} + \nabla\mathbf{u}^{\mathbf{t}} + \nabla\mathbf{u}^{\mathbf{t}}\nabla\mathbf{u}) \tag{3.3}$$

is the Green-St. Venant strain tensor. This problem is often referred to as one for a St. Venant-Kirchhoff material. Again, this describes materials in configurations in which the "large displacement, small strain" assumption is valid.

We may also separate the linear and nonlinear parts of the first equation in (3.1), and write it as

$$\mu\Delta\mathbf{u} + (\lambda+\mu)\nabla\nabla\cdot\mathbf{u} + \nabla\cdot\mathbf{P_3}(\nabla\mathbf{u}) = \mathbf{f}, \tag{3.4}$$

where $\Delta\mathbf{u} = \nabla\cdot\nabla\mathbf{u}$ is the vector Laplacian of $\mathbf{u}$ and $\mathbf{P_3}(\nabla\mathbf{u})$ is the following matrix of degree 3 polynomials of the components of $\nabla\mathbf{u}$:

$$\mathbf{P_3}(\mathbf{X}) = \frac{1}{2}\lambda\left(tr(\mathbf{X}^t\mathbf{X})\mathbf{I} + tr(\mathbf{X} + \mathbf{X}^t + \mathbf{X}^t\mathbf{X})\mathbf{X}\right) + \mu\left(\mathbf{X}^2 + \mathbf{X}^t\mathbf{X} + \mathbf{X}\mathbf{X}^t + \mathbf{X}\mathbf{X}^t\mathbf{X}\right).$$

The linear part of the left-side operator in (3.4) is simply the linear elasticity equations and the nonlinear part can be thought of as a perturbation that begins to dominate as $\nabla \mathbf{u}$ becomes large compared to $\mathbf{u}$.

The unknown, $\mathbf{u}$, is the usual displacement vector. We assume that the Lamé constants, λ and μ, are bounded by satisfying $0 < \mu_0 < \mu < \mu_1$ and $0 < \lambda_0 < \lambda < \lambda_1$, for appropriate positive bounds. Physically, this corresponds to an assumption of compressibility of the material. The more difficult problem of incompressible materials is considered for linear elasticity in [5, 6, 8, 16]. A complete study of the geometrically nonlinear elasticity problem in a least-squares context in the incompressible limit remains an open problem. Without loss of generality, we scale the problem so that $\mu = 1$ and let λ determine the level of compressibility. See Section 11 for examples of Lamé constants for different materials.

The case where $\Gamma_T = \emptyset$ corresponds to a pure displacement problem, $\Gamma_D = \emptyset$ a pure traction problem, and otherwise a mixed boundary condition problem.

4. Existence and Uniqueness of Solutions. In this section, we establish existence and uniqueness results that confirm well-posedness of System (3.1). We restrict ourselves here to the pure displacement problem on domains with sufficiently smooth data and boundaries (see Remark 4.3 at the end of this section).

Let ∂ represent either first partial derivative, ∂_x or ∂_y, and suppose $\delta > 0$ and $k \geq 0$. The following lemma addresses smoothness of products of functions in $H^{1+\delta}(\Omega)$ and $H^{1+k}(\Omega)$.

Lemma 4.1 *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^2$. Then there exists a constant, C, depending only on Ω, such that, for $u \in H^{1+\delta}(\Omega)$ and $v \in H^{1+k}(\Omega)$, the product uv satisfies*

$$\|uv\|_{1+k} \leq C\|u\|_{1+\delta}\|v\|_{1+k},$$
$$\|\partial(uv)\|_k \leq C\|u\|_{1+\delta}\|v\|_{1+k}.$$

Proof. This is a consequence of the Sobolev imbedding theorem and a proof can be seen in Chapter 1 of [14]. ∎

The following theorem establishes criteria for existence and uniqueness of solutions to Problem (3.1).

Theorem 4.2 *Let Ω be a domain in $\mathbb{R}^2$ with boundary of class C^{2+m} for some $m > 0$. Then there exists a neighborhood, $\mathcal{Q}_0^m$, of the origin in $H^m(\Omega)$ and a neighborhood, $\mathcal{U}_0^{1+m}$, of the origin in $\mathcal{U}^{1+m} = \{\nabla \mathbf{v} : \mathbf{v} \in H^{2+m}(\Omega), \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\} \subset H^{1+m}(\Omega)$ such that for each $\mathbf{f} \in \mathcal{Q}_0^m$, the boundary value problem*

$$\mathcal{L}(\nabla \mathbf{u}) := \nabla \cdot [(\mathbf{I} + \nabla \mathbf{u})\Sigma(\mathbf{E}(\nabla \mathbf{u}))] = \mathbf{f} \tag{4.1}$$

has exactly one solution, $\nabla \mathbf{u}^*$, in $\mathcal{U}_0^{1+m}$.

Proof. We observe that nonlinear operator $\mathcal{L}$ maps $\nabla \mathbf{u} \in H^{1+m}(\Omega)$ into $H^m(\Omega)$ by applying Lemma 4.1, and that $\mathcal{L}$ is differentiable between these spaces (in fact, all derivatives of order ≥ 4 are zero).

Since $\mathcal{L}(\mathbf{0}) = \mathbf{0}$, we can then apply the implicit function theorem in a neighborhood of the origin in $\mathcal{U}^{1+m} \times H^m(\Omega)$. Thus, we now only need to check that the derivative of $\mathcal{L}$ at the origin, $\mathcal{L}'(\mathbf{0})$, is bijective between $\mathcal{U}^{1+m}$ and $H^m(\Omega)$ and has continuous inverse.

But $\mathcal{L}'(\mathbf{0})$ is exactly the operator of linear elasticity. It is known that if $\partial\Omega$ is a C^{2+m} boundary and $\mathbf{f} \in H^m(\Omega)$, then there is a unique weak solution to the linear pure displacement problem, $\mathbf{u} \in H^{2+m}(\Omega)$ (see [11]). This immediatly implies $\nabla\mathbf{u} \in H^{1+m}(\Omega)$. Thus, we have shown that continuous operator $\mathcal{L}'(\mathbf{0})$ is bijective. Now since $\mathcal{L}'(\mathbf{0})$ is a continuous, bijective, linear map between two Banach spaces, by the closed graph theorem, it must have a continuous inverse.

By the implicit function theorem there is, therefore, a neighborhood , $\mathcal{Q}_0^m$, of the origin in $H^m(\Omega)$ and a neighborhood, $\mathcal{U}_0^{1+m}$, of the origin in $\mathcal{U}^{1+m}$ such that there is a unique solution, $\nabla\mathbf{u}^* \in \mathcal{U}_0^{1+m}$, for any function $\mathbf{f} \in \mathcal{Q}_0^m$. ■

Thus, the pure displacement problem with sufficiently smooth data and domain is well-posed, and the solution, $\nabla\mathbf{u}$, remains small in the H^{1+m} norm, with no direct restriction on $\mathbf{u}$ itself. This is consistent with the small strains assumption in the geometrically nonlinear elastic model.

Remark 4.3 *We are ultimately interested in nonhomogenous problems on polygonal domains, which are known to have solutions less smooth than described above. At corner points and/or points of changing boundary condition type on the boundary, a locally-weighted norm can be used to remove the effect of the nonsmooth solution. In [19] a weighted-norm least-squares method is developed for problems with boundary singularities. For simplicity, we choose here to focus on the formulation and analysis of the linearized problem since, even for problems lacking full global regularity, we may expect the regularity predicted in Theorem 4.2 away from abruptly changing material interfaces in the interior of* Ω *for sufficiently smooth data* $\mathbf{f}$*.*

5. Least-Squares Formulation. We want to replace the nonlinear elasticity problem with a series of linear problems, which we then reformulate as a first-order system. Introducing the deformation, $\boldsymbol{\phi} = \mathbf{x} + \mathbf{u}$, the deformation gradient, $\boldsymbol{\Phi} = \nabla\boldsymbol{\phi}$, and the displacement gradient, $\mathbf{U} = \nabla\mathbf{u}$, Problem (3.1) becomes one of finding the zero of

$$\mathcal{F}(\mathbf{U}) = \nabla \cdot [\tfrac{1}{2}\lambda tr(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})(\mathbf{I} + \mathbf{U}) + (\mathbf{I} + \mathbf{U})(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})] - \mathbf{f}, \quad (5.1)$$

subject to the constraint

$$\nabla \times \mathbf{U} = \mathbf{0}, \quad (5.2)$$

for $\mathbf{U}$ satisfying appropriate boundary conditions (recall also that we assume $\mu = 1$).

The Fréchet derivative of $\mathcal{F}(\mathbf{U})$ in the direction of $\mathbf{V}$ is

$$\begin{aligned} \mathcal{F}'(\mathbf{U})[\mathbf{V}] = \\ \nabla \cdot [&\lambda tr(\mathbf{U} + \tfrac{1}{2}\mathbf{U}^t\mathbf{U})\mathbf{V} + \lambda tr(\mathbf{V} + \mathbf{V}^t\mathbf{U})(\mathbf{I} + \mathbf{U}) \\ &+ (\mathbf{I} + \mathbf{U})(\mathbf{V} + \mathbf{V}^t + \mathbf{U}^t\mathbf{V} + \mathbf{V}^t\mathbf{U}) + \mathbf{V}(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})]. \end{aligned} \quad (5.3)$$

Thus, Newton's method for approximating the solution of (5.1) is given by iteratively solving the linear problem

$$\begin{cases} \mathcal{F}'(\mathbf{U}_n)[\mathbf{U}_{n+1}] = \mathcal{F}'(\mathbf{U}_n)[\mathbf{U}_n] - \mathcal{F}(\mathbf{U}_n) \\ \qquad \nabla \times \mathbf{U}_{n+1} = \mathbf{0} \end{cases} \quad (5.4)$$

for $\mathbf{U}_{n+1}$, with initial approximation $\mathbf{U}_0 = \mathbf{0}$.

It is convenient to view 2×2 matrices as 4×1 vectors so that general linear operators on such quantities can be written as 4×4 matrices. Thus, define operator $\mathcal{K} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{4\times 1}$ by

$$\mathcal{K}\begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} = (U_{11}\,, U_{12}\,, U_{21}\,, U_{22})^t\,,$$

for any 2×2 matrix, $(\mathbf{U})_{ij} = U_{ij}$. If A is a 4×4 matrix, then the quantity $A\mathbf{U}$ should be interpreted as $A\mathbf{U} = \mathcal{K}^{-1}\left(A\mathcal{K}\mathbf{U}\right)$.

With the relation $\mathbf{\Phi} = \mathbf{I} + \mathbf{U}$, we define the following linear operators:

$$M_1(\mathbf{\Phi}) = \begin{pmatrix} \Phi_{11}^2 & \Phi_{11}\Phi_{12} & \Phi_{11}\Phi_{21} & \Phi_{11}\Phi_{22} \\ \Phi_{12}\Phi_{11} & \Phi_{12}^2 & \Phi_{12}\Phi_{21} & \Phi_{12}\Phi_{22} \\ \Phi_{21}\Phi_{11} & \Phi_{21}\Phi_{12} & \Phi_{21}^2 & \Phi_{21}\Phi_{22} \\ \Phi_{22}\Phi_{11} & \Phi_{22}\Phi_{12} & \Phi_{22}\Phi_{21} & \Phi_{22}^2 \end{pmatrix},$$

$$M_2(\mathbf{\Phi}) = (\Phi_{11}^2 + \Phi_{12}^2 + \Phi_{21}^2 + \Phi_{22}^2 - 2)I,$$

$$M_3(\mathbf{\Phi}) =$$

$$\begin{pmatrix} 3\Phi_{11}^2 + \Phi_{12}^2 + \Phi_{21}^2 & 2\Phi_{11}\Phi_{12} + \Phi_{21}\Phi_{22} & 2\Phi_{11}\Phi_{21} + \Phi_{12}\Phi_{22} & \Phi_{12}\Phi_{21} \\ 2\Phi_{11}\Phi_{12} + \Phi_{21}\Phi_{22} & \Phi_{11}^2 + 3\Phi_{12}^2 + \Phi_{22}^2 & \Phi_{11}\Phi_{22} & \Phi_{11}\Phi_{21} + 2\Phi_{12}\Phi_{22} \\ 2\Phi_{11}\Phi_{21} + \Phi_{12}\Phi_{22} & \Phi_{11}\Phi_{22} & \Phi_{11}^2 + 3\Phi_{21}^2 + \Phi_{22}^2 & \Phi_{11}\Phi_{12} + 2\Phi_{21}\Phi_{22} \\ \Phi_{12}\Phi_{21} & \Phi_{11}\Phi_{21} + 2\Phi_{12}\Phi_{22} & \Phi_{11}\Phi_{12} + 2\Phi_{21}\Phi_{22} & \Phi_{12}^2 + \Phi_{21}^2 + 3\Phi_{22}^2 \end{pmatrix}.$$

Using the relation $\mathbf{\Phi} = \mathbf{I} + \mathbf{U}$ as a change of variables, define the system matrix, A, as a function of $\mathbf{U}$ by

$$A(\mathbf{U}) = \lambda M_1(\mathbf{I} + \mathbf{U}) + \tfrac{1}{2}\lambda M_2(\mathbf{I} + \mathbf{U}) + M_3(\mathbf{I} + \mathbf{U}) - \mathbf{I}.$$

In this way, we may denote the linear operator in (5.4) as:

$$\mathcal{F}'(\mathbf{U})[\mathbf{V}] = \nabla \cdot A(\mathbf{U})\mathbf{V}.$$

Denoting $A_n = A(\mathbf{U}_n)$ and $\mathcal{F}_n = \mathcal{F}(\mathbf{U}_n)$, the Newton step for the $(n+1)^{st}$ iterate $\mathbf{U}$ (dropping the subscript) may now be written as

$$\begin{cases} \nabla \cdot A_n\mathbf{U} = \nabla \cdot A_n\mathbf{U}_n - \mathcal{F}_n \\ \quad \nabla \times \mathbf{U} = \mathbf{0}. \end{cases} \tag{5.5}$$

We may apply an analogous linearization technique to the traction boundary conditions by defining

$$\mathcal{T}(\mathbf{U}) = \mathbf{n} \cdot [\tfrac{1}{2}\lambda tr(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})(\mathbf{I} + \mathbf{U}) + (\mathbf{I} + \mathbf{U})(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})] - \mathbf{g}$$

and letting $\mathcal{T}_n = \mathcal{T}(\mathbf{U}_n)$. The corresponding Newton step for the traction boundaries then becomes

$$\mathbf{n} \cdot A_n\mathbf{U} = \mathbf{n} \cdot A_n\mathbf{U}_n - \mathcal{T}_n, \quad \text{on} \quad \Gamma_T. \tag{5.6}$$

Since $\mathbf{u} = \mathbf{0}$ on the displacement boundaries, we may enforce the derivative of $\mathbf{u}$ along those boundaries to be zero:

$$\boldsymbol{\tau} \cdot \mathbf{U} = \mathbf{0}, \quad \text{on} \quad \Gamma_D. \tag{5.7}$$

Thus, we may completely decouple the unknowns in $\mathbf{u}$ from the unknowns in $\mathbf{U}$. We concentrate here on the first-stage solution of $\mathbf{U}$, that is, solving the problem for $\mathbf{U}$ and later recovering $\mathbf{u}$, if necessary.

We take the initial approximation for Newton's method to be the reference configuration, $\mathbf{U}_0 = \mathbf{0}$; the system matrix for the first Newton step is

$$A_0 = \begin{pmatrix} \lambda+2 & 0 & 0 & \lambda \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ \lambda & 0 & 0 & \lambda+2 \end{pmatrix};$$

and we can write $\nabla \cdot (A_0\mathbf{U}_0) - \mathcal{F}_0 = \mathbf{f}$ and $\mathbf{n} \cdot (A_0\mathbf{U}_0) - \mathcal{T}_0 = \mathbf{g}$. Thus, we may write the first step of Newton's method as

$$\begin{cases} \nabla \cdot (A_0\mathbf{U}) = \mathbf{f}, & \text{in} \quad \Omega, \\ \nabla \times \mathbf{U} = \mathbf{0}, & \text{in} \quad \Omega, \\ \boldsymbol{\tau} \cdot \mathbf{U} = \mathbf{0}, & \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot (A_0\mathbf{U}) = \mathbf{g}, & \text{on} \quad \Gamma_T. \end{cases} \tag{5.8}$$

This is the form of the linear elasticity equations studied in [6, 8].

System (5.5) depends explicitly on the current approximation to the solution. Specifically, matrix A_n deviates from A_0 as $\boldsymbol{\Phi}$ deviates from the identity (or, as $\mathbf{U}$ deviates from $\mathbf{0}$). Much is known about the first Newton step because it is exactly the linear elasticity case. For example, assuming sufficient smoothness of the solution, a least-squares functional associated with System (5.8) can be shown to be H^1 elliptic with the aid of Korn's inequality. In fact, this ellipticity property is even retained for a modification of System (5.8) in the incompressible limit in [8]. Existence, uniqueness, and optimal finite element approximation bounds immediately follow (see [6, 8]). For the linearized problem, however, the literature reflects relatively little theory in $W^{k,2}$ Sobolev spaces, and a thorough study of these equations in a least-squares context has, to our knowledge, not been explored. Thus, we are led to develop new theory that establishes well-posedness of, and a fast solution technique for, the linearized equations consisting of (5.5), (5.6), and (5.7).

6. Problem Modification. One goal of the least-squares methodology is to develop a functional that is H^1 elliptic whenever possible. It is well known that such systems admit uniform and optimal H^1 approximations when using standard finite elements for the discretization and standard multigrid solvers for the resulting linear system (see [7]). For System (5.5), this poses a challenge because the system matrix, A_n, is generally pointwise indefinite. In this section, we introduce a modification to (5.5) that overcomes this difficulty, and we make a reasonable physical assumption that guarantees positive-definiteness of the modified system matrix. To this end, consider modifying A_n by adding to it a matrix of the form

$$B(c) = \begin{pmatrix} 0 & 0 & 0 & c \\ 0 & 0 & -c & 0 \\ 0 & -c & 0 & 0 \\ c & 0 & 0 & 0 \end{pmatrix},$$

where c is any fixed constant. It is easy to see that $\nabla \cdot B(c)\nabla \mathbf{p} = \mathbf{0}$ for any function $\mathbf{p}$, so the solution to (5.5) is unaffected by replacing A_n with $A_n + B(c)$. (We note, however, that this modification cannot be applied to the traction boundary conditions given in (5.6).) In [6, 8], this idea is applied with $c = \mu = 1$ in conjunction with a rotation of the unknowns so that the equations of linear elasticity in the incompressible limit mirror the Stokes equations. We apply the same idea here, not to transform the equations to a more well known form, but rather to shift the spectrum to be positive. Indeed, in the linear case, the spectrum of A_0, which is $\{0, 2, 2, 2\lambda+2\}$, can be shifted by $B(1)$ so that the spectrum of $A_0 + B(1)$ becomes $\{1, 1, 1, 2\lambda + 3\}$. Numerical experiments on the spectrum of A_n indicate that a choice of $c = 1$ is also most effective for shifting the spectrum to be positive for general deformations. We now study this question analytically.

Matrix $\tilde{A}_n = A_n + B(1)$ seems to depend on the four linearly independent components of $\mathbf{\Phi}_n$. However, under an appropriate change of variables, the eigenvalues can be exactly expressed in terms of just two scalar functions over Ω:

$$\begin{aligned} \sigma &= \Phi_{11}^2 + \Phi_{12}^2 + \Phi_{21}^2 + \Phi_{22}^2, \\ \delta &= \Phi_{11}\Phi_{22} - \Phi_{12}\Phi_{21}. \end{aligned} \tag{6.1}$$

In fact, the eigenvalues of $\tilde{A}_n$ are as follows:

$$\begin{aligned} \mathbf{\Lambda}_1 &= \frac{1}{2}(\lambda+2)\sigma - \delta - \lambda, \\ \mathbf{\Lambda}_2 &= \frac{1}{2}(\lambda+2)\sigma + \delta - \lambda - 2, \\ \mathbf{\Lambda}_3 &= (\lambda + \frac{3}{2})\sigma - (\lambda+1) - \sqrt{\frac{1}{4}(\lambda+3)^2\sigma^2 - (6\lambda+9)\delta^2 + 2\lambda\delta + 1}, \\ \mathbf{\Lambda}_4 &= (\lambda + \frac{3}{2})\sigma - (\lambda+1) + \sqrt{\frac{1}{4}(\lambda+3)^2\sigma^2 - (6\lambda+9)\delta^2 + 2\lambda\delta + 1}. \end{aligned} \tag{6.2}$$

That the spectrum can be represented by only two independent quantities is surprising, but that the two quantities have such an obvious physical meaning is remarkable. For example, δ, the determinant of the Jacobian of the mapping of the current approximation, is a local measure of change in volume: $\delta > 1$ indicates areas under tension and $\delta < 1$ indicates areas under compression. Similarly, $\sigma < 2$ when there is significant local compression. In general, we know that in the small strains regime $\sigma \approx 2$ and $0 < \delta \approx 1$.

Since the model for the geometrically nonlinear elasticity equations assumes a deformed configuration with small strains, we may assume small strains of the solution. We show in Section 8 that, for an initial guess sufficiently close to the solution, each iterate remains bounded near the solution and Newton's method converges. Under these constraints, we take each iterate to satisfy some small strain condition of the form $\|\mathbf{E}\| \ll 1$. We now choose the norm to enforce this condition.

Define the following *Frobenius* norm for tensor-valued quantities:

$$\|\mathbf{X}\|_{Fr}^2 = \sup_{\Omega} \sum_{ij} \left(\mathbf{X}_{ij}\right)^2.$$

Thus, we may write $\|\mathbf{\Phi}\|_{Fr} = \|\sigma\|_\infty$. We can also express the Frobenius norm of the strain tensor exactly in terms of variables σ and δ. We now establish bounds on

the strain that guarantee that the modified system matrix is uniformly symmetric positive definite.

Recall that the strain tensor is given by $\mathbf{E}(\mathbf{U}) = \frac{1}{2}(\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U})$. Define

$$\mathcal{S}_\lambda = \left\{ \mathbf{U} : \|\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U}\|_{Fr} < \frac{\sqrt{2}}{\lambda + 3} \right\}$$

as the set of all displacement gradients corresponding to deformations with "small strains." We may choose $\mathcal{Q}_0^m$ small enough to ensure that $\mathbf{f} \in \mathcal{Q}_0^m$ guarantees $\mathbf{U} \in \mathcal{S}_\lambda$. Thus, the condition of small strains follows from the assumptions in Theorem 4.2. We explore the regime of small strains in more detail in Section 11.

Theorem 6.1 *For all $\mathbf{U} \in \mathcal{S}_\lambda$, matrix $\tilde{A} = A(\mathbf{U}) + B(1)$ is uniformly positive definite over Ω.*

Proof. We directly compute positive lower bounds on each eigenvalue of $\tilde{A}$. For convenience, we work with $\mathbf{\Phi} = \mathbf{I} + \mathbf{U}$, where $\mathbf{U} + \mathbf{U}^t + \mathbf{U}^t\mathbf{U} = \mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}$. Let $\varepsilon = \|\mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}\|_{Fr}$. By direct computation, we write

$$\varepsilon^2 = (\sigma - 1)^2 - 2\delta^2 + 1. \tag{6.3}$$

We also have

$$\sigma \geq 2\delta \tag{6.4}$$

because $\sigma - 2\delta = (\Phi_1 - \Phi_{22})^2 + (\Phi_{12} + \Phi_{21})^2 \geq 0$. Using (6.4), we can also establish upper and lower bounds on σ in terms of ε. Specifically, $\varepsilon^2 = (\sigma - 1)^2 - 2\delta^2 + 1 \geq (\sigma - 1)^2 - \frac{1}{2}\sigma^2 - 1 = \frac{1}{2}(\sigma - 2)^2$, so

$$2 - \sqrt{2}\varepsilon \leq \sigma \leq 2 + \sqrt{2}\varepsilon. \tag{6.5}$$

Expressions for the eigenvalues of $\tilde{A}$ are given in (6.2). Starting with $\mathbf{\Lambda}_1$ and using (6.4) and (6.5), we obtain

$$\begin{aligned}
\mathbf{\Lambda}_1 &= \tfrac{1}{2}(\lambda + 2)(\sigma - 2) - \delta + 2 \\
&\geq \tfrac{1}{2}(\lambda + 2)(\sigma - 2) - \tfrac{1}{2}\sigma + 2 \\
&= \tfrac{1}{2}(\lambda + 1)\sigma - \lambda \\
&\geq \tfrac{1}{2}(\lambda + 1)(2 - \sqrt{2}\varepsilon) - \lambda \\
&= 1 - \tfrac{\sqrt{2}}{2}(\lambda + 1)\varepsilon,
\end{aligned}$$

which is strictly positive when $\varepsilon < \frac{\sqrt{2}}{\lambda+1}$.

Again, using (6.4) and (6.5) along with (6.3), the second eigenvalue satisfies

$$\begin{aligned}
\mathbf{\Lambda}_2 &= \tfrac{1}{2}(\lambda + 2)(\sigma - 2) + \delta \\
&\geq \tfrac{1}{2}(\lambda + 2)(2\delta - 2) + \delta \\
&= \delta(\lambda + 3) - (\lambda + 2) \\
&= \frac{\sqrt{2}}{2}((\sigma - 1)^2 + 1 - \varepsilon^2)^{\frac{1}{2}}(\lambda + 3) - (\lambda + 2) \\
&\geq \frac{\sqrt{2}}{2}((1 - \sqrt{2}\varepsilon)^2 + 1 - \varepsilon^2)^{\frac{1}{2}}(\lambda + 3) - (\lambda + 2) \\
&= ((1 - \sqrt{2}\varepsilon + \tfrac{1}{2}\varepsilon^2)^{\frac{1}{2}}(\lambda + 3) - (\lambda + 2),
\end{aligned}$$

which is strictly positive when $f(\varepsilon) = \frac{1}{2}\varepsilon^2 - \sqrt{2}\varepsilon + 1 - \left(\frac{\lambda+2}{\lambda+3}\right)^2 > 0$. Solving for the roots of $f(\varepsilon)$, we see that $f(\varepsilon)$ is positive for $\varepsilon < \frac{\sqrt{2}}{\lambda+3}$.

The third eigenvalue is more cumbersome to treat and requires a bit more care than the first two. Write $\mathbf{\Lambda}_3 = R - \sqrt{Z}$, where $R = (\lambda + \frac{3}{2})\sigma - (\lambda + 1)$ and $Z = \frac{1}{4}(\lambda+3)^2\sigma^2 - (6\lambda+9)\delta^2 + 2\lambda\delta + 1$. It can be seen that Z must be nonnegative for $\lambda > 0$ by writing

$$\begin{aligned} Z &= \tfrac{1}{4}(\lambda+3)^2\sigma^2 - (6\lambda+9)\delta^2 + 2\lambda\delta + 1 \\ &= \tfrac{1}{4}\lambda^2\sigma^2 + \tfrac{1}{4}(6\lambda+9)(\sigma+2\delta)(\sigma-2\delta) + 2\lambda\delta + 1 \\ &> 0, \end{aligned}$$

since $\sigma \geq 2\delta$. From the bound on $\mathbf{\Lambda}_2$ and (6.5), we know, for $\lambda > 0$, that

$$\sigma \;\geq\; 2 - \sqrt{2}\varepsilon \;>\; 2 - \sqrt{2}\left(\frac{\sqrt{2}}{\lambda+3}\right) \;>\; \frac{4}{3}$$

and, thus, $R > 0$. Therefore, $\mathbf{\Lambda}_3$ is positive when $R^2 - Z$ is positive. But we may write

$$\begin{aligned} R^2 - Z &= (\lambda+\tfrac{3}{2})^2\sigma^2 - 2(\lambda+1)(\lambda+\tfrac{3}{2})\sigma + (\lambda+1)^2 \\ &\quad - \tfrac{1}{4}(\lambda+3)^2\sigma^2 + (6\lambda+9)\delta^2 - 2\lambda\delta - 1 \\ &\geq (\lambda+\tfrac{3}{2})^2\sigma^2 - 2(\lambda+1)(\lambda+\tfrac{3}{2})\sigma + (\lambda+1)^2 \\ &\quad - \tfrac{1}{4}(\lambda+3)^2\sigma^2 + (6\lambda+9)\delta^2 - \lambda\sigma - 1 \\ &= (\lambda+\tfrac{3}{2})^2\sigma^2 - 2(\lambda+1)(\lambda+\tfrac{3}{2})\sigma + (\lambda+1)^2 \\ &\quad - \tfrac{1}{4}(\lambda+3)^2\sigma^2 + \tfrac{1}{2}(6\lambda+9)((\sigma-1)^2 - \varepsilon^2 + 1) - \lambda\sigma - 1 \\ &= \tfrac{1}{4}(\lambda^2+6\lambda+6)(3\sigma-2)(\sigma-2) + (2\lambda+3)(1-\tfrac{3}{2}\varepsilon). \end{aligned}$$

Since $\sigma > \frac{4}{3}$ implies that the quadratic term in σ, $(3\sigma-2)(\sigma-2)$, is monotonically increasing, we can apply the lower bound in (6.5) to get

$$\begin{aligned} R^2 - Z &\geq \tfrac{1}{4}(\lambda^2+6\lambda+6)(3\sigma-2)(\sigma-2) + (2\lambda+3)(1-\tfrac{3}{2}\varepsilon) \\ &\geq \tfrac{1}{4}(\lambda^2+6\lambda+6)(-\sqrt{2}\varepsilon + \tfrac{3}{2}\varepsilon^2) + (2\lambda+3)(1-\tfrac{3}{2}\varepsilon) \\ &= \tfrac{3}{2}(\lambda^2+4\lambda+3)\varepsilon^2 - \sqrt{2}(\lambda^2+6\lambda+6)\varepsilon + 2\lambda + 3. \end{aligned}$$

Again, solving for the roots of this quadratic equation in ε, we see that $R^2 - Z$ is positive when $\varepsilon < \frac{\sqrt{2}}{\lambda+3}$.

Finally, the fourth eigenvalue, $\mathbf{\Lambda}_4$, is bounded below by $\mathbf{\Lambda}_3$ and the proof is complete. ∎

It is interesting to note that the bounds for the first three eigenvalues are of the same order (the second and third are even the exact same bound). This suggests that the modification to matrix A_n is optimally balanced with $B(c)$ for $c = 1$.

The full, modified, linearized system may now be written as

$$\begin{cases} \nabla \cdot (\tilde{A}_n \mathbf{U}) = \mathbf{f}_n, & \text{in} \quad \Omega, \\ \nabla \times \mathbf{U} = \mathbf{0}, & \text{in} \quad \Omega, \\ \boldsymbol{\tau} \cdot \mathbf{U} = \mathbf{0}, & \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot (A_n \mathbf{U}) = \mathbf{g}_n, & \text{on} \quad \Gamma_T, \end{cases}$$

where $\mathbf{f}_n = \nabla \cdot (\tilde{A}_n \mathbf{U}_n) - \mathcal{F}_n$ and $\mathbf{g}_n = \mathbf{n} \cdot (A_n \mathbf{U}) - \mathcal{T}_n$.

Define the L^2 functional

$$G(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \|\nabla \cdot (\tilde{A}_n \mathbf{U}) - \mathbf{f}_n\|^2 + \|\nabla \times \mathbf{U}\|^2, \tag{6.6}$$

and define, for any $m > 0$, the space

$$\mathcal{V}^m = \{\mathbf{V} \in H^m(\Omega)^4 : \ \mathbf{n} \cdot (A_n \mathbf{V}) = \mathbf{g}_n \text{ on } \Gamma_T, \ \boldsymbol{\tau} \cdot \mathbf{V} = \mathbf{0} \text{ on } \Gamma_D\}.$$

In the case of pure displacement boundary conditions ($\Gamma_N = \emptyset$), we denote the space by $\mathcal{V}_D^m$.

The least-squares minimization problem for each Newton step is: given $\mathbf{U}_n$, $\mathbf{f}_n$ and $\mathbf{g}_n$, find $\mathbf{U} \in \mathcal{V}^1$ such that

$$G(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \inf_{\mathbf{V} \in \mathcal{V}^1} G(\mathbf{V}; \mathbf{U}_n, \mathbf{f}_n).$$

7. Ellipticity. To use the L^2 based functional in (6.6) on each Newton step, we must assume that the previous iterate is in $H^{1+\delta}(\Omega)$ for some $\delta > 0$ because (6.6) is composed of derivatives of products of the unknown and the previous solution and, in $\mathbb{R}^2$, the space $H^{1+\delta}(\Omega)$ is closed under multiplication only for $\delta > 0$ (see Lemma 4.1). Thus, showing only H^1 ellipticity of (6.6) is not sufficient to establish a well-defined Newton iteration; we must show that each iterate remains in $H^{1+\delta}(\Omega)$. In this section, we establish H^{1+k} ellipticity of an H^k based functional for $k \geq 0$, and show that minimizing the L^2 based functional is sufficient to guarantee the required smoothness of each iterate. Our theoretical results hold for the pure displacement problem.

For clarity, we use the following conventions: $\delta > 0$ and $k \geq 0$ (our final results require the cases $k = 0$ and $k = \delta$).

In Theorem 6.1, matrix $\tilde{A} = A(\mathbf{U}) + B(1)$ is uniformly symmetric positive definite over Ω when the strain of $\mathbf{U}$ is sufficiently small, that is, for $\mathbf{U} \in \mathcal{S}_\lambda$. In this section, we assume this property holds and consider the solution of a general Newton step of the pure displacement problem by minimizing the more general H^k based functional

$$G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \|\nabla \cdot (\tilde{A}_n \mathbf{U}) - \mathbf{f}_n\|_k^2 + \|\nabla \times \mathbf{U}\|_k^2. \tag{7.1}$$

Its associated minimization problem is: given $\mathbf{U}_n \in H^{1+\delta}(\Omega)$ and $\mathbf{f}_n \in H^k(\Omega)$, find $\mathbf{U} \in \mathcal{V}_D^{1+k}$ such that

$$G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{f}_n) = \inf_{\mathbf{V} \in \mathcal{V}_D^{1+k}} G_k(\mathbf{V}; \mathbf{U}_n, \mathbf{f}_n). \tag{7.2}$$

By Lemma 4.1, it is clear that $\mathbf{U} \in H^{1+k}(\Omega)$ and $\mathbf{U}_n \in H^{1+\delta}(\Omega)$ are sufficient to ensure that $\nabla \cdot (\tilde{A}_n \mathbf{U}) \in H^k(\Omega)$.

The following series of lemmas leads to establishing equivalence of $G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{0})^{1/2}$ to the H^{1+k} norm.

Lemma 7.1 *Let Ω be a simply connected domain in $\mathbb{R}^2$ and suppose $\mathbf{V} \in L^2(\Omega)^4$. Then $\nabla \cdot \mathbf{V} = \mathbf{0}$ and $\int_{\partial\Omega} \mathbf{n} \cdot \mathbf{V} = \mathbf{0}$ if and only if there exists a function $\mathbf{r} \in H^1(\Omega)^2$ such that $\mathbf{V} = \nabla^\perp \mathbf{r}$. Furthermore, $\mathbf{r} \in H^1(\Omega)^2$ is unique up to an additive constant vector in $\mathbb{R}^2$.*

Proof. The result follows by applying Theorem 3.1 in Chapter I of [14] to each block component of $\mathbf{V}$. ∎

Lemma 7.2 *Let Ω be a simply connected domain in $\mathbb{R}^2$. Every $\mathbf{V} \in L^2(\Omega)^4$ has the orthogonal decomposition $\mathbf{V} = \nabla\mathbf{p} + \nabla^\perp\mathbf{q}$ for $\mathbf{p} \in H^1(\Omega)^2, \mathbf{q} \in H_0^1(\Omega)^2$. Furthermore, $\mathbf{q}$ is unique in $H_0^1(\Omega)^2$ and $\mathbf{p}$ is unique in $H^1(\Omega)^2$ up to an additive constant vector in $\mathbb{R}^2$.*

Proof. The result follows by applying Theorem 3.2 in Chapter I of [14] to each block component of $\mathbf{V}$. ∎

Lemma 7.3 *Assume that $\mathbf{U} \in \mathcal{S}_\lambda$ and denote $\tilde{A} = A + B$, with $A = A(\mathbf{U})$ and $B = B(1)$ as defined in Section 6. Also assume that $A\mathbf{Z}$ and $B\mathbf{Z}$ are in $L^2(\Omega)^4$. If $\mathbf{Z} \in \mathcal{V}_D^1$ satisfies the system*

$$\begin{cases} \nabla \cdot \tilde{A}\mathbf{Z} = \mathbf{0}, & \text{in} \quad \Omega, \\ \nabla \times \mathbf{Z} = \mathbf{0}, & \text{in} \quad \Omega, \end{cases} \tag{7.3}$$

then it must be the trivial solution, $\mathbf{Z} = \mathbf{0}$.

Proof. By Lemma 7.2, $\mathbf{Z} = \nabla\mathbf{p} + \nabla^\perp\mathbf{q}$ for $\mathbf{p} \in H^1(\Omega)^2, \mathbf{q} \in H_0^1(\Omega)^2$. The second equation in (7.3) implies

$$\mathbf{0} = \nabla \times \mathbf{Z} = \nabla \times \nabla\mathbf{p} + \nabla \times \nabla^\perp\mathbf{q} = -\Delta\mathbf{q},$$

and, since $\mathbf{q} \in H^1(\Omega)_0^2$, we must have $\mathbf{q} = \mathbf{0}$. Thus, $\mathbf{Z} = \nabla\mathbf{p}$.

Now, using Green's Formula with $\mathbf{1} = (1,1)^t$, we get

$$\mathbf{0} = \langle \nabla \cdot A\mathbf{Z}, \mathbf{1}\rangle + \langle A\mathbf{Z}, \nabla\mathbf{1}\rangle = \int_{\partial\Omega} \mathbf{n} \cdot A\mathbf{Z}.$$

Applying Lemma 7.1 to $A\mathbf{Z}$ yields $A\mathbf{Z} = \nabla^\perp\mathbf{r}$ for $\mathbf{r} \in H^1(\Omega)^2$. Since $\mathbf{0} = \boldsymbol{\tau}\cdot\mathbf{Z} = \boldsymbol{\tau}\cdot\nabla\mathbf{p}$ on $\partial\Omega$, we know that $\mathbf{p} = \mathbf{p}_0$ is constant on $\partial\Omega$. We thus have

$$\begin{aligned} \langle A\mathbf{Z}, \mathbf{Z}\rangle &= \langle \nabla^\perp\mathbf{r}, \nabla\mathbf{p}\rangle \\ &= \langle -\nabla \cdot \nabla^\perp\mathbf{r}, \mathbf{p}\rangle + \int_{\partial\Omega} (\mathbf{n} \cdot \nabla^\perp\mathbf{r})\mathbf{p} \\ &= \int_{\partial\Omega} (\mathbf{n} \cdot \nabla^\perp\mathbf{r})\mathbf{p} \\ &= \mathbf{p}_0 \int_{\partial\Omega} \mathbf{n} \cdot A\mathbf{Z} \\ &= \mathbf{0}. \end{aligned}$$

Since $\mathbf{Z} = \nabla\mathbf{p}$, we may then write $B\mathbf{Z} = \nabla^\perp\mathbf{s}$, where $\mathbf{s} = \begin{pmatrix} p_2 \\ -p_1 \end{pmatrix}$. Thus,

$$\begin{aligned} \langle B\mathbf{Z}, \mathbf{Z}\rangle &= \langle \nabla^\perp\mathbf{s}, \nabla\mathbf{p}\rangle \\ &= \langle \mathbf{s}, \nabla \times \nabla\mathbf{p}\rangle - \int_{\partial\Omega} (\boldsymbol{\tau} \cdot \nabla\mathbf{p})\mathbf{s} \\ &= -\int_{\partial\Omega} (\boldsymbol{\tau} \cdot \nabla\mathbf{p}_0)\mathbf{s} \\ &= \mathbf{0}, \end{aligned}$$

which implies $\langle \tilde{A}\mathbf{Z}, \mathbf{Z}\rangle = \mathbf{0}$. Since $\mathbf{U} \in \mathcal{S}_\lambda$, matrix $\tilde{A}$ is positive definite, and we must have $\mathbf{Z} = \mathbf{0}$. ∎

Now consider the following elliptic boundary value problem:

$$\nabla \cdot M\nabla \mathbf{p} = \mathbf{f} \quad \text{in} \quad \Omega, \tag{7.4}$$

satisfying either $\mathbf{p} = \mathbf{0}$ or $\mathbf{n} \cdot M\nabla \mathbf{p} = \mathbf{0}$ on $\partial\Omega$. When Ω has $C^{1+k,1}$ boundary and M is uniformly positive definite over Ω with coefficients in $C^{k,1}(\bar{\Omega})$, Problem (7.4) admits the regularity bound,

$$\|\mathbf{p}\|_{2+k} \leq C\|\mathbf{f}\|_k, \tag{7.5}$$

for $\mathbf{p} \in H^{2+k}(\Omega)$. Chapter 2 of [15] establishes this for integer values of k. For noninteger k, we may appeal to interpolation in Sobolev norms as in [17]. Similar regularity results, with different assumptions than here, are given in [1, 2, 12].

Lemma 7.4 *Assume a solution to nonlinear Problem* (5.1)*:* $\mathbf{U}^* \in \mathcal{V}_D^{2+k} \cap \mathcal{S}_\lambda$*. Assume also that* Ω *is smooth enough to admit* (7.5) *for* $k \geq 0$*. Let* $\tilde{A}_* = A(\mathbf{U}^*) + B(1)$*. Then there exists a positive constant,* c_**, independent of* $\mathbf{U}$*, such that*

$$\|\mathbf{U}\|_{1+k} \leq c_*(\|\nabla \cdot \tilde{A}_*\mathbf{U}\|_k + \|\nabla \times \mathbf{U}\|_k)$$

for all $\mathbf{U} \in \mathcal{V}_D^{1+k}$*.*

Proof. Consider the skew-symmetric orthogonal matrix

$$Q = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \tag{7.6}$$

The following relations are easily derived:

$$\begin{aligned} \nabla\times &= \nabla \cdot Q, \\ \nabla\cdot &= \nabla \times Q^t, \\ \nabla^\perp &= Q\nabla, \\ \nabla &= Q^t\nabla^\perp, \\ \mathbf{n}\cdot &= -\boldsymbol{\tau} \cdot Q, \\ \boldsymbol{\tau}\cdot &= \mathbf{n} \cdot Q^t. \end{aligned} \tag{7.7}$$

Since $\tilde{A}_*$ is uniformly positive definite over Ω, there are constants, λ_1, $\lambda_2 > 0$, such that

$$\lambda_1 \boldsymbol{\xi}^t\boldsymbol{\xi} \leq \boldsymbol{\xi}^t\tilde{A}_*\boldsymbol{\xi} \leq \lambda_2\boldsymbol{\xi}^t\boldsymbol{\xi} \tag{7.8}$$

and

$$\frac{1}{\lambda_2}\boldsymbol{\xi}^t\boldsymbol{\xi} \leq \boldsymbol{\xi}^t\tilde{A}_*^{-1}\boldsymbol{\xi} \leq \frac{1}{\lambda_1}\boldsymbol{\xi}^t\boldsymbol{\xi} \tag{7.9}$$

for any $\boldsymbol{\xi} \in \mathbb{R}^4$. Define

$$\mathcal{C} = Q^t\tilde{A}_*^{-1}Q,$$

and note that

$$\boldsymbol{\xi}^t\boldsymbol{\xi} = \boldsymbol{\xi}^t Q^t Q\boldsymbol{\xi} = (Q\boldsymbol{\xi})^t(Q\boldsymbol{\xi})$$

and

$$\boldsymbol{\xi}^t\mathcal{C}\boldsymbol{\xi} = \boldsymbol{\xi}^t Q^t \tilde{A}_*^{-1} Q\boldsymbol{\xi} = (Q\boldsymbol{\xi})^t \tilde{A}_*^{-1}(Q\boldsymbol{\xi}).$$

Now, it can easily be seen that $\mathcal{C}$ is symmetric and uniformly positive definite over Ω:

$$\frac{1}{\lambda_2}\boldsymbol{\xi}^t\boldsymbol{\xi} \leq \boldsymbol{\xi}^t\mathcal{C}\boldsymbol{\xi} \leq \frac{1}{\lambda_1}\boldsymbol{\xi}^t\boldsymbol{\xi}. \tag{7.10}$$

We also note that

$$\nabla \times \tilde{A}_*^{-1}\nabla^\perp = \nabla \cdot Q\tilde{A}_*^{-1}Q\nabla = -\nabla\cdot\mathcal{C}\nabla.$$

With $\mathbf{U}^* \in \mathcal{V}_D^{2+k}$ and $\mathbf{U} \in \mathcal{V}_D^{1+k}$, we have that $\nabla\cdot\tilde{A}_*\mathbf{U} \in H^k(\Omega)$, and thus, for any $\mathbf{U} \in \mathcal{V}_D^{1+k}$, there is a unique $\mathbf{p} \in H^{2+k}(\Omega)$ that satisfies

$$\begin{cases} \nabla\cdot\tilde{A}_*\nabla\mathbf{p} = \nabla\cdot\tilde{A}_*\mathbf{U}, & \text{in} \quad \Omega, \\ \mathbf{p} = \mathbf{0}, & \text{on} \quad \partial\Omega, \end{cases} \tag{7.11}$$

and $\mathbf{q} \in H^{2+k}(\Omega)$ that satisfies

$$\begin{cases} -\nabla\cdot\mathcal{C}\nabla\mathbf{q} = \nabla\times\mathbf{U}, & \text{in} \quad \Omega, \\ \mathbf{n}\cdot\mathcal{C}\nabla\mathbf{q} = \mathbf{0}, & \text{on} \quad \partial\Omega, \end{cases} \tag{7.12}$$

and $\int_\Omega \mathbf{q}\,dx = \mathbf{0}$. Now define $\mathbf{Z} = \mathbf{U} - \nabla\mathbf{p} - \tilde{A}_*^{-1}\nabla^\perp\mathbf{q}$. Note that $\mathbf{Z} \in H^{1+k}(\Omega)$ and, on $\partial\Omega$, that

$$\begin{aligned} \boldsymbol{\tau}\cdot\mathbf{Z} &= \boldsymbol{\tau}\cdot\mathbf{U} - \boldsymbol{\tau}\cdot\nabla\mathbf{p} - \boldsymbol{\tau}\cdot\tilde{A}_*^{-1}\nabla^\perp\mathbf{q} \\ &= -\boldsymbol{\tau}\cdot Q\mathcal{C}Q^t\nabla^\perp\mathbf{q} \\ &= \mathbf{n}\cdot\mathcal{C}\nabla\mathbf{q} \\ &= \mathbf{0}. \end{aligned}$$

Thus, $\mathbf{Z} \in \mathcal{V}_D^{1+k}$. We further see that

$$\begin{aligned} \nabla\cdot\tilde{A}_*\mathbf{Z} &= \nabla\cdot\tilde{A}_*\mathbf{U} - \nabla\cdot\tilde{A}_*\nabla\mathbf{p} - \nabla\cdot\tilde{A}_*\tilde{A}_*^{-1}\nabla^\perp\mathbf{q} \\ &= \mathbf{0} \end{aligned}$$

and

$$\begin{aligned} \nabla\times\mathbf{Z} &= \nabla\times\mathbf{U} - \nabla\times\nabla\mathbf{p} - \nabla\times\tilde{A}_*^{-1}\nabla^\perp\mathbf{q} \\ &= \nabla\times\mathbf{U} + \nabla\cdot\mathcal{C}\nabla\mathbf{q} \\ &= \mathbf{0}. \end{aligned}$$

By Lemma 7.3, we therefore conclude that $\mathbf{Z} = \mathbf{0}$. Since $\mathbf{U}^* \in H^{2+k}(\Omega) \subseteq C^{k,1}(\bar{\Omega})$, we may apply (7.5) to Problems (7.11) and (7.12). Combining these bounds with the triangle inequality and (7.9), we may write

$$\begin{aligned} \|\mathbf{U}\|_{1+k} &\leq \|\nabla\mathbf{p}\|_{1+k} + \|\tilde{A}_*^{-1}\nabla^\perp\mathbf{q}\|_{1+k} \\ &\leq \|\mathbf{p}\|_{2+k} + 1/\lambda_1\|\mathbf{q}\|_{2+k} \\ &\leq C(\|\nabla\cdot\tilde{A}_*\mathbf{U}\|_k + \|\nabla\times\mathbf{U}\|_k). \end{aligned}$$

Application of Inequality (2.1) completes the proof. ∎

Recall that we cast the solution of nonlinear Problem (3.1) as the zero of $\mathcal{F}(\mathbf{U})$, where $\mathcal{F}'(\mathbf{U})[\mathbf{V}]$ is given in Equation (5.3). We consider the boundedness of the second Fréchet derivative of $\mathcal{F}(\mathbf{U})$ in directions $\mathbf{V}$ and $\mathbf{W}$ in the following lemma.

Lemma 7.5 *For all $\mathbf{U}, \mathbf{V} \in H^{1+\delta}(\Omega)$ and $\mathbf{W} \in H^{1+k}(\Omega)$, there exists a positive constant, c_2, such that*

$$\|\mathcal{F}''(\mathbf{U})[\mathbf{V}, \mathbf{W}]\|_k \leq c_2 \|\mathbf{U}\|_{1+\delta} \|\mathbf{V}\|_{1+\delta} \|\mathbf{W}\|_{1+k}. \tag{7.13}$$

Proof. Writing the second Fréchet derivative of $\mathcal{F}$ in the directions $\mathbf{V}$ and $\mathbf{W}$ as

$$\begin{aligned}
&\mathcal{F}''(\mathbf{U})[\mathbf{V}, \mathbf{W}] = \\
&\nabla \cdot \big[\lambda tr(\mathbf{W}^t\mathbf{V})(\mathbf{I}+\mathbf{U}) + \lambda tr(\mathbf{V}^t(\mathbf{I}+\mathbf{U}))\mathbf{W} + \lambda tr(\mathbf{W}^t(\mathbf{I}+\mathbf{U}))\mathbf{V} \\
&\quad + (\mathbf{I}+\mathbf{U})(\mathbf{W}^t\mathbf{V} + \mathbf{V}^t\mathbf{W}) + \mathbf{V}(\mathbf{W} + \mathbf{W}^t + W^t\mathbf{U} + \mathbf{U}^t\mathbf{W}) \\
&\quad + \mathbf{W}(\mathbf{V} + \mathbf{V}^t + V^t\mathbf{U} + \mathbf{U}^t\mathbf{V})\big],
\end{aligned}$$

we see that each component may be written as a linear combination of terms of the form $\partial(W_i V_j U_k)$ or $\partial(W_i V_j)$, $i, j, k = 1, 2, 3, 4$, where ∂, again, represents either ∂_x or ∂_y. The lemma then follows by the triangle inequality and applying Lemma 4.1 to each term once or twice. ∎

Define the $H^{1+\delta}$ neighborhood of the solution by

$$\mathcal{B}_r = \{\mathbf{U} \in H^{1+\delta}(\Omega) : \|\mathbf{U} - \mathbf{U}^*\|_{1+\delta} < r\}.$$

We now are able to state the main result of this section.

Theorem 7.6 *Assume Ω has C^{3+k} boundary and that $\mathbf{f} \in H^{1+k}(\Omega)$ is small enough to guarantee that Problem (5.1) has solution $\mathbf{U}^* \in H^{2+k}(\Omega) \cap \mathcal{S}_\lambda$ by Theorem 4.2. Then there exists some $r > 0$ and constants $c_0, c_1 > 0$, depending only on f and Ω, such that, for all $\mathbf{U}_n \in \mathcal{V}_D^{1+\delta} \cap \mathcal{B}_r$,*

$$c_0 \|\mathbf{U}\|_{1+k}^2 \;\leq\; G_k(\mathbf{U}; \mathbf{U}_n, \mathbf{0}) \;\leq\; c_1 \|\mathbf{U}\|_{1+k}^2 \tag{7.14}$$

for every $\mathbf{U} \in \mathcal{V}_D^{1+k}$ for $k \geq 0$.

Proof. The upper bound follows from the triangle inequality and Lemma 4.1. With $\mathbf{U}_n \in H^{1+\delta}(\Omega)$, Lemma 4.1 guarantees that $A\mathbf{U}, B\mathbf{U} \in H^{1+k}(\Omega)$ when $\mathbf{U} \in H^{1+k}(\Omega)$. By Theorem 4.2 with $m = 1 + k$, there is a solution to Problem (5.1), $\mathbf{U}^* \in H^{2+k}(\Omega) \cap \mathcal{S}_\lambda$. Thus, by Lemma 7.4, we know there exists a positive constant, c_*, independent of $\mathbf{U}$, such that

$$\|\mathbf{U}\|_{1+k} \leq c_* (\|\nabla \cdot \tilde{A}(\mathbf{U}^*)\mathbf{U}\|_k + \|\nabla \times \mathbf{U}\|_k) \tag{7.15}$$

for all $\mathbf{U} \in \mathcal{V}_D^{1+k}$. We now need only to extend this result to the operator linearized about $\mathbf{U}_n$ rather than $\mathbf{U}^*$. Recall that we may denote $\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}] = \nabla \cdot A(\mathbf{U}_n)\mathbf{U}$ and $\mathcal{F}'(\mathbf{U}^*)[\mathbf{U}] = \nabla \cdot A(\mathbf{U}^*)\mathbf{U}$. By the mean value theorem, we may write

$$\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}] - \mathcal{F}'(\mathbf{U}^*)[\mathbf{U}] = \mathcal{F}''(\hat{\mathbf{U}})[\mathbf{U}, \mathbf{U}_n - \mathbf{U}^*] \tag{7.16}$$

for some $\hat{\mathbf{U}} = \theta\mathbf{U}_n + (1-\theta)\mathbf{U}^*$ with $\theta \in [0,1]$. Since $\mathbf{U}_n \in \mathcal{B}_r$, $\hat{\mathbf{U}}$ can be bounded in the $H^{1+\delta}$ norm in the following way:

$$\begin{aligned}\|\hat{\mathbf{U}}\|_{1+\delta} &= \|\theta\mathbf{U}_n + (1-\theta)\mathbf{U}^*\|_{1+\delta} \\ &\leq \|\theta(\mathbf{U}_n - \mathbf{U}^*)\|_{1+\delta} + \|\mathbf{U}^*\|_{1+\delta} \\ &\leq r + \|\mathbf{U}^*\|_{1+\delta}.\end{aligned} \tag{7.17}$$

So, by (7.16), the triangle inequality, (7.13), and (7.15), we have

$$\begin{aligned}&\|\nabla\cdot\tilde{A}(\mathbf{U}_n)\mathbf{U}\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad= \|\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}]\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad= \|\mathcal{F}'(\mathbf{U}^*)[\mathbf{U}] + \mathcal{F}''(\hat{\mathbf{U}})[\mathbf{U},\mathbf{U}_n-\mathbf{U}^*]\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad\geq \|\mathcal{F}'(\mathbf{U}^*)[\mathbf{U}]\|_k - \|\mathcal{F}''(\hat{\mathbf{U}})[\mathbf{U},\mathbf{U}_n-\mathbf{U}^*]\|_k + \|\nabla\times\mathbf{U}\|_k \\ &\quad\geq \|\nabla\cdot A(\mathbf{U}^*)\mathbf{U}\|_k + \|\nabla\times\mathbf{U}\|_k - c_2\|\hat{\mathbf{U}}\|_{1+\delta}\|\mathbf{U}\|_{1+k}\|\mathbf{U}_n-\mathbf{U}^*\|_{1+\delta} \\ &\quad\geq c_*^{-1}\|\mathbf{U}\|_{1+k} - c_2 r(r+\|\mathbf{U}^*\|_{1+\delta})\|\mathbf{U}\|_{1+k} \\ &\quad= (c_*^{-1} - c_2 r(r+\|\mathbf{U}^*\|_{1+\delta}))\|\mathbf{U}\|_{1+k} \\ &\quad\geq C\|\mathbf{U}\|_{1+k},\end{aligned} \tag{7.18}$$

where C is guaranteed to be positive for r sufficiently small. Application of Inequality (2.1) completes the proof. ∎

Corollary 7.7 *Assume that Ω, $\mathbf{f}$, $\mathbf{U}^*$ and $\mathbf{U}_n \in \mathcal{V}_D^{1+\delta} \cap \mathcal{B}_r$ satisfy the assumptions of Theorem 7.6. Then, for some r sufficiently small, the unique $\mathbf{U}$ that satisfies*

$$G_0(\mathbf{U};\mathbf{U}_n,\mathbf{f}_n) = \inf_{\mathbf{V}\in\mathcal{V}_D^1} G_0(\mathbf{V};\mathbf{U}_n,\mathbf{f}_n) \tag{7.19}$$

also satisfies

$$G_\delta(\mathbf{U};\mathbf{U}_n,\mathbf{f}_n) = \inf_{\mathbf{V}\in\mathcal{V}_D^{1+\delta}} G_\delta(\mathbf{V};\mathbf{U}_n,\mathbf{f}_n). \tag{7.20}$$

Proof. From the Riesz representation theorem and Theorem 7.6 with $k=0$, we have a unique minimizer, $\mathbf{U}$, of the L^2 based functional in (7.19) in $H^1(\Omega)$. Similarly, for $k=\delta>0$, we also have a unique minimizer, $\mathbf{U}'$, of the H^δ based functional in (7.20) in $H^{1+\delta}(\Omega)$. Since these functionals both have zero minimum, $\mathbf{U}'$ must also minimize the functional in (7.19). Thus, $\mathbf{U} = \mathbf{U}' \in H^{1+\delta}(\Omega)$. ∎

Therefore, we are able to conclude that, under sufficient smoothness requirements, minimizing the L^2 based functional is sufficient to guarantee that each Newton iterate, $\mathbf{U} = \mathbf{U}_{n+1}$, remains in $H^{1+\delta}(\Omega)$.

8. Convergence of Newton's Method. We now consider the sequence of iterates arising from the minimization of each linearized functional under the assumptions of Theorem 7.6. This section details the theory and assumptions for the convergence of Newton's method. As in Theorem 7.6, we assume the solution to the previous Newton step to be in $\mathcal{B}_r$. Here, we show convergence of the iterates in the $H^{1+\delta}$ norm and that each iterate remains in $\mathcal{B}_r$.

Consider the Taylor expansion of $\mathcal{F}(\mathbf{U}^*)$ about the current approximation $\mathbf{U}_n$:

$$\mathbf{0} = \mathcal{F}(\mathbf{U}^*) = \mathcal{F}(\mathbf{U}_n) + \mathcal{F}'(\mathbf{U}_n)[\mathbf{U}^*-\mathbf{U}_n] + \frac{1}{2}\mathcal{F}''(\tilde{\mathbf{U}})[\mathbf{U}^*-\mathbf{U}_n,\mathbf{U}^*-\mathbf{U}_n], \tag{8.1}$$

for $\tilde{\mathbf{U}} = \omega \mathbf{U}_n + (1-\omega)\mathbf{U}^*$ with $\omega \in [0,1]$. As in Equation (7.17), if $\mathbf{U}_n \in \mathcal{B}_r$, then $\tilde{\mathbf{U}}$ satisfies

$$\|\tilde{\mathbf{U}}\|_{1+\delta} \leq r + \|\mathbf{U}^*\|_{1+\delta}. \tag{8.2}$$

Recall that we may write the Newton iterate, $\mathbf{U}$, as the solution to Problem (7.2) and, thus,

$$\mathcal{F}'(\mathbf{U}_n)[\mathbf{U} - \mathbf{U}_n] = -\mathcal{F}(\mathbf{U}_n), \tag{8.3}$$

with $\nabla \times \mathbf{U} = \nabla \times \mathbf{U}_n = \nabla \times \mathbf{U}^* = 0$.

Applying (8.1), (8.3), (7.13) and (8.2) to the bound in (7.18), and recalling that $\mathbf{U}_n \in \mathcal{B}_r$, we get

$$\begin{aligned}
\|\mathbf{U}^* - \mathbf{U}\|_{1+\delta} &\leq \frac{1}{\sqrt{c_0}} \|\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}^* - \mathbf{U}]\|_\delta + \|\nabla \times (\mathbf{U}^* - \mathbf{U})\|_\delta \\
&= \frac{1}{\sqrt{c_0}} \|\mathcal{F}'(\mathbf{U}_n)[\mathbf{U}^* - \mathbf{U}_n] - \mathcal{F}'(\mathbf{U}_n)[\mathbf{U} - \mathbf{U}_n]\|_\delta \\
&= \frac{1}{2\sqrt{c_0}} \|\mathcal{F}''(\tilde{\mathbf{U}})[\mathbf{U}^* - \mathbf{U}_n, \mathbf{U}^* - \mathbf{U}_n]\|_\delta \\
&\leq \frac{c_2}{2\sqrt{c_0}} \|\tilde{\mathbf{U}}\|_{1+\delta} \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \\
&\leq \frac{c_2}{2\sqrt{c_0}} (r + \|\mathbf{U}^*\|_{1+\delta}) r \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \\
&:= c_3 \; r \; \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta}
\end{aligned} \tag{8.4}$$

which proves that Newton's method converges for r sufficiently small. Again noting that $\mathbf{U}_n \in \mathcal{B}_r$, we further note that

$$\begin{aligned}
\|\mathbf{U}^* - \mathbf{U}\|_{1+\delta} &\leq c_3 r \|\mathbf{U}^* - \mathbf{U}_n\|_{1+\delta} \\
&\leq c_3 r^2.
\end{aligned}$$

To verify that $\mathbf{U} \in \mathcal{B}_r$, we only need to show that $c_3 r^2 < r$. Substituting the definition of c_3, we see that this is satisfied for

$$r \; < \; \frac{1}{2}\left(\sqrt{\|\mathbf{U}^*\|_{1+\delta}^2 + \eta} - \|\mathbf{U}^*\|_{1+\delta}\right) \; < \; \frac{\eta}{4\|\mathbf{U}^*\|_{1+\delta}},$$

where $\eta = \frac{8\sqrt{c_0}}{c_2}$. This shows that, for guaranteed convergence, larger solutions require better initial guesses than smaller solutions (as measured in the $H^{1+\delta}$ norm). We now consider the issue of finding an appropriate "good" initial guess.

9. Multilevel Solution. As described above, the solution to nonlinear System (3.1) is generally comprised of several Newton iterations. The first few iterations are crude approximations to the true solution of the nonlinear problem. It is therefore appropriate to represent the early approximations on a mesh with fewer degrees of freedom. As the Newton iterates remove more of the error due to the nonlinearity, the approximations can be represented on increasingly finer meshes. In other words, we wish to eliminate as much of the nonlinear error as possible on coarse grids where it is less expensive.

The approach in [13] uses this multilevel nested iteration Newton idea with a FOSLS finite element discretization and a multigrid solver to achieve a robust solution strategy for a certain class of nonlinear problems. Under particular assumptions on the form of the nonlinearity, the finite element spaces used, the smoothness of the solution, and the ellipticity of the linearized equations, convergence to the solution is established with accuracy comparable to discretization error on the finest level at a cost proportional to the degrees of freedom on the finest level. We briefly summarize this Nested Iteration-Newton-FOSLS-Multigrid (NI-Newton-FOSLS-MG) algorithm and detail the additional assumptions we must make for application to the geometrically nonlinear elasticity system.

Define the hierarchy of discrete nested subspaces,

$$\mathcal{V}^{h_0} \subset \mathcal{V}^{h_1} ... \subset \mathcal{V}^{h_J} \subset \mathcal{V}_D^{1+\delta}. \tag{9.1}$$

The following algorithm describes the NI-Newton-FOSLS-MG method:

1. Begin with a zero approximation, $\mathbf{U}_0$, on coarsest level $\mathcal{V}^{h_0}$.
2. Linearize the equations about the current approximation and form the discrete least-squares minimization problem.
3. Apply m multigrid cycles to the resulting matrix equations.
4. Repeat steps 2 and 3 n times on the current level.
5. Interpolate the current approximation to the next finer level, $\mathcal{V}^{h_i}$.
6. Repeat steps 2-5 until desired accuracy is achieved.

To apply the results of [13] to the nonlinear elasticity system, we must make the following series of assumptions.

A1: Assume the existence of a solution, $\mathbf{U}^* \in H^{2+\delta}(\Omega)$, to Problem (3.1). For our problem, this is justified in Section 4. Theorem 4.2 with $m = 1 + \delta$ requires that the boundary of Ω be $C^{3+\delta}$ smooth and $\mathbf{f} \in H^{1+\delta}(\Omega)$ in order to guarantee $\mathbf{U}^* \in H^{2+\delta}(\Omega)$. In the context of elasticity, the internal forcing function, $\mathbf{f}$, is generally at least this smooth for a wide range of practical problems. Assuming a very smooth domain, however, is a stronger restriction than we generally wish to adhere to in practice. We do find that in practice this can be relaxed in some cases, but in many cases we must consider complimentary methods for dealing with nonsmooth domains.

A2: Assume the operator of linearized elasticity maps $\mathcal{V}_D^{1+k}$ into $H^k(\Omega)$. This is established in Section 7.

A3: Assume H^{1+k} ellipticity of the functional as in (7.1). Theorem 7.6 establishes this for the pure displacement problem under the small strains assumption of Theorem 6.1.

A4: Assume boundedness of the second Fréchet derivative of $\mathcal{F}$ as in (7.13). Justification of this is established in Lemma 7.5.

A5: Assume the finite element spaces in (9.1) guarantee the following approximation properties and inverse estimate. Let I_ν^h be a bounded H^ν projection onto finite element space $\mathcal{V}^h$. We assume interpolation bounds of the form

$$\|\mathbf{U} - I_{1+\delta}^h \mathbf{U}\|_\gamma \leq C h^{2+\delta-\gamma} \|\mathbf{U}\|_{2+\delta} \quad \forall \gamma \in [0, 1+\delta],$$

and the inverse estimate

$$\|\mathbf{U}\|_\beta \leq \frac{C}{h^{\beta-\gamma}} \|\mathbf{U}\|_\gamma \quad \forall \mathbf{U} \in \mathcal{V}^h,\ \beta \in [0, 1+\delta],\ \gamma \in [0, \beta].$$

We concentrate on standard finite element subspaces of H^1 (for example, bilinears on rectangles) which exhibit these properties; see [3] for details.

A6: Assume a sufficiently fine coarsest level by insisting that $\mathcal{B}_r \cap \mathcal{V}^{h_0} \neq \emptyset$ and that the initial guess is sufficiently close to the solution by choosing $\mathbf{U}_0 \in \mathcal{B}_r \cap \mathcal{V}^{h_0}$. Bounds on r can be found in the full theory in [13].

Under these assumptions, we may directly apply the theory developed in [13]. By this theory, there are values of m and n, independent of h, in the multilevel algorithm described above that result in an approximation on the finest level that is accurate to the level of discretization error at a cost proportional to the degrees of freedom on the finest level.

There are many contributions to the error in each approximation in the NI-Newton-FOSLS-MG solution process. In the innermost iteration, the multigrid solver reduces the algebraic error by performing a number of multigrid cycles before relinearizing. On each grid level, there is discretization error associated with the finite element space used. A sufficient number of Newton steps must be performed on each level to eliminate the error associated with the nonlinearity. For a truly optimal algorithm, we must consider the sources of error that contribute to the total error in the current approximation, and make decisions on how to proceed in the algorithm in order to efficiently reduce the total error to an acceptable level.

10. Computational Results. To validate the theory presented above, consider the numerical approximation to the solution of a pure displacement problem on domain $\Omega = [0,1]^2$, with Lamé constants $\lambda = 2.15$, $\mu = 1$. As a test problem, we choose the solution to nonlinear Problem (3.1) to be

$$\mathbf{u}^* = \begin{pmatrix} x(1-x)y^2(1-y)^2\sin(\pi x) \\ x^2(1-x)^2y^2(1-y)^2\cos(\pi y) \end{pmatrix},$$

and let $\mathbf{U}^* = \nabla\mathbf{u}^*$ be the exact solution for the first stage problem. The right-side function, $\mathbf{f}$, is computed accordingly.

Denote by $\mathcal{V}^h$ the space of continuous piecewise bilinear finite elements on a uniform grid of mesh size h. For convenience, we use this space for all test problems. Each step of the pure displacement problem is found by minimizing the discrete functional,

$$G(\mathbf{U}^h; \mathbf{U}_n^h, \mathbf{f}_n) = \|\nabla \cdot (\tilde{A}_n \mathbf{U}^h) - \mathbf{f}_n\|^2 + \|\nabla \times \mathbf{U}^h\|^2, \tag{10.1}$$

over the space

$$\mathcal{V}_D^h = \{\mathbf{V}^h \in \mathcal{V}^h : \boldsymbol{\tau} \cdot \mathbf{V}^h = \mathbf{0} \text{ on } \partial\Omega\}.$$

We begin with an initial guess of $\mathbf{U}_0 = \mathbf{0}$ so that the first Newton step corresponds to the linear elasticity case. Recall that we seek the solution to the original nonlinear problem as well as each linearized step. Define the following nonlinear functional to measure the convergence to nonlinear Problem (3.1):

$$\mathcal{G}(\mathbf{U}; \mathbf{f}) = \|\mathcal{F}(\mathbf{U})\|^2 + \|\nabla \times \mathbf{U}\|^2. \tag{10.2}$$

In an $H^{1+\delta}$ neighborhood near the solution, a simple computation on a Taylor series of $\mathcal{F}$ about $\mathbf{U}^*$ (invoking Lemma (7.5)) shows that $\mathcal{G}(\mathbf{U}; \mathbf{f})$ is equivalent to $G(\mathbf{U}; \mathbf{U}^*, \mathbf{0})$, indicating that the H^1 norm of the error to the nonlinear problem can be effectively monitored by $\mathcal{G}(\mathbf{U}; \mathbf{f})$. Near convergence of Newton's method, the nonlinear and linearized functionals tend to take on the same values. Thus, a practical measure of how much of the error in the approximation is due to the nonlinearity can be obtained by the difference in the linearized and nonlinear functional values.

For the test problem summarized in Table 10.1, we ensure that essentially all algebraic error is removed from each system by reducing the residual by a factor of 10^6 using $V(1,1)$ cycles. Numerical results are reported for: grid level, N ($h = (N+1)^{-1}$); Newton step, m; linearized functional norm, $G(\mathbf{U}^h; \mathbf{U}_n^h, \mathbf{f})^{1/2}$; nonlinear functional norm, $\mathcal{G}(\mathbf{U}^h; \mathbf{f})^{1/2}$; L^2 error of the solution, $\|\mathbf{U}^* - \mathbf{U}^h\|$; and asymptotic multigrid convergence factor, ρ. On each mesh size, the Newton iterations were started with initial guess $\mathbf{U}_0^h = \mathbf{0}$.

N	m	$G(\mathbf{U}^h; \mathbf{U}_n^h, \mathbf{f})^{1/2}$	$\mathcal{G}(\mathbf{U}^h; \mathbf{f})^{1/2}$	$\|\mathbf{U}^* - \mathbf{U}^h\|$	ρ
8	1	4.73e-02	4.73e-02	2.16e-03	0.70
8	2	2.58e-02	2.58e-02	1.91e-03	0.67
8	3	2.58e-02	2.58e-02	1.91e-03	0.61
16	1	1.32e-02	4.44e-02	1.31e-03	0.70
16	2	1.29e-02	1.29e-02	4.84e-04	0.69
16	3	1.29e-02	1.29e-02	4.84e-04	0.46
32	1	6.66e-03	4.38e-02	1.26e-03	0.73
32	2	6.44e-03	6.44e-03	1.22e-04	0.73
32	3	6.44e-03	6.44e-03	1.22e-04	0.70
64	1	3.38e-03	4.37e-02	1.26e-03	0.77
64	2	3.22e-03	3.22e-03	3.02e-05	0.75
64	3	3.22e-03	3.22e-03	3.04e-05	0.72
128	1	1.73e-03	4.36e-02	1.27e-03	0.80
128	2	1.61e-03	1.62e-03	7.63e-06	0.79
128	3	1.61e-03	1.61e-03	7.53e-06	0.76

TABLE 10.1

Numerical results for the pure displacement problem with known smooth solution, without using nested iteration, using $V(1,1)$ cycles.

By comparing the functional norm and L^2 error values after three Newton steps on a sequence of levels in Table 10.1, we see that the method achieves the optimal discretization accuracy of $O(h^2)$ with respect to the L^2 error norm, and $O(h)$ with respect to the linearized and nonlinear functional norms. Netwon's method essentially converges by the second iteration independent of the mesh size. But, even with such fast convergence, we see that the nonlinear functional values of the first Newton step on each level essentially stall, indicating that, even for this relatively simple problem, the linear elasticity approximation is a poor approximation to the geometrically nonlinear approximation.

We see that, for this problem, the multigrid convergence factors based on $V(1,1)$ cycles are bounded above by about 0.8. While these are acceptable convergence factors, in the remainder of the numerical test problems, we use an AMG $V(1,1)$ preconditioned conjugate gradient cycle to improve performance. We denote these accelerated cycles by $V(1,1) - pcg$, and because these cycles generally do not reduce the error by a consistent amount, we report the average convergence factor, $\bar{\rho}$, rather than the asymptotic convergence factor. Refer to [18] for complete details on such cycles.

The convergence factor does not take into account the amount of work done per cycle. For an appropriate measure of the work expended by a multigrid cycle, we define the cycle complexity as the total work per cycle relative to one fine grid relaxation sweep. As a numerical estimate of the cycle complexity, we compute the

total number of nonzero matrix entries on each level, multiplied by the number of relaxation sweeps on that level, divided by the number of nonzero matrix entries of the finest level operator. Define the work per Newton step as the work per cycle multiplied by the number of cycles per step, and the total work, W_T, as the cumulative amount of work expended relative to the current level. One such work unit is equivalent to one relaxation sweep on the finest level.

We now wish to solve the same problem as above, but in the most efficient way possible. To this end, we implement the nested iteration strategy described in Section 9. Instead of reducing the residual of each linear system by a given amount, we take only three $V(1,1) - pcg$ cycles per Newton step and one Newton step per level. Table 10.2 summarizes these results.

N	m	$\mathcal{G}(\mathbf{U}^h;\mathbf{f})^{1/2}$	$\bar{\rho}$	W_T	time (s)
8	2	2.64e-02	0.29	12.3	1
16	3	1.31e-02	0.25	16.0	4
32	4	6.61e-03	0.24	19.0	15
64	5	3.32e-03	0.24	20.8	60
128	6	1.66e-03	0.23	21.6	242

TABLE 10.2

Numerical results for the pure displacement problem with known smooth solution, using nested iteration and three $V(1,1) - pcg$ cycles per step.

As Tables 10.1 and 10.2 show, the nested iteration method achieves optimal discretization accuracy, and the nonlinear functional on the finest grid is within 5% of discretization error. The average convergence factors for the $V(1,1) - pcg$ cycles remain bounded and of very reasonable size for this problem. The total amount of work required for the solution at each level is essentially bounded at less than 25 work units, and the time to solution for each level scales almost exactly with the number of degrees of freedom of the problem.

The numerical results presented here are for the pure displacement problem with small strains. In practice, we find that the method performs similarly to the results shown here for mixed boundary conditions and for somewhat larger strains than the theory allows. In the next section, we show that the small strains assumption admits a large class of interesting problems.

11. Validating the Small-Strains Assumption. According to Ciarlet in [11], for any homogenous, isotropic, elastic material, the stress and strain tensors satisfy the relation given by

$$\mathbf{\Sigma}(\mathbf{E}) = \lambda tr(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E} + o(\mathbf{E}). \tag{11.1}$$

But the model of geometrically nonlinear elasticity uses the linear stress-strain relation given in Equation (3.2), that is, we drop the higher-order terms, $o(\mathbf{E})$, under an assumption of small strains. Thus, in analysis of the geometrically nonlinear elasticity system, we are free to impose reasonable restrictions on the size of $\|\mathbf{E}\|$ without limiting the scope of the model.

In Theorem 6.1, we assume that the strain associated with the solution of each Newton iterate satisfies

$$\|\mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}\|_{Fr} < \frac{\sqrt{2}}{\lambda + 3}, \tag{11.2}$$

	ν	λ
Rubber	0.49	33.3
Lead	0.44	7.30
Aluminum	0.34	2.15
Nickel	0.30	1.56
Steel	0.28	1.22
Glass	0.25	1.00

TABLE 11.1
Material constants of homogenous isotropic materials.

where we have scaled the problem such that $\mu = 1$. In this section, we investigate this restriction and provide examples of different configurations and their relation to (11.2) and material constant λ.

Since physically we must have $\lambda > 0$, we first see that an upper bound on the allowed strain is at $\|\mathbf{\Phi}^t\mathbf{\Phi} - \mathbf{I}\|_{Fr} = \sqrt{2}/3 \approx 0.471$. We further notice that bound (11.2) is always violated by any nonzero strain in the limit as $\lambda \to \infty$. Thus, our notion of "small strains" is coupled to the assumption of compressibility. The Poisson ratio of an elastic material is given by

$$\nu = \frac{\lambda}{2(\lambda + \mu)},$$

and we may think of the incompressible limit as $\lambda \to \infty$ (for bounded μ) or $\nu \to 0.5$. In Table 11.1, we provide a few examples of common materials and their material properties. Because we are chiefly concerned with the value of λ relative to μ, we report the unitless $\lambda \leftarrow \lambda/\mu$. For unscaled constants with meaningful physical units, consult [11]. For the numerical test problems in this paper, we uniformly choose to use $\lambda = 2.15$, that of aluminum, as the level of compressibility.

Consider the two basic modes of strain: shear and tensile strain. A unit square domain under either uniform shear or uniform tensile strain has corresponding displacements of the form

$$\mathbf{u}_{shear} = \begin{pmatrix} \beta y \\ 0 \end{pmatrix}, \text{ or } \mathbf{u}_{tensile} = \begin{pmatrix} \alpha x \\ 0 \end{pmatrix}.$$

Parameters β and α determine the extent of deformation as pictured in Figure 11.1.

Under these deformations, we may apply (6.1) and (6.3) to satisfy (11.2) for these two cases. For pure shear strain, we require β and λ to satisfy

$$(\lambda + 3)^2\beta^2(\beta^2 + 2) - 2 < 0,$$

and, for pure tensile strain, we require α and λ to satisfy

$$(\lambda + 3)^2\alpha^2(\alpha + 2)^2 - 2 < 0.$$

These relations are satisfied for the parameters in the shaded regions shown in Figure 11.2.

Now consider the following example of a deformed configuration with large displacements but small strains. The strain of the discrete approximation is computed pointwise from (6.3) for mesh size $h = 1/16$. The deformation is from a simple cantilever beam under a constant gravitational force. The max pointwise strain is 0.241

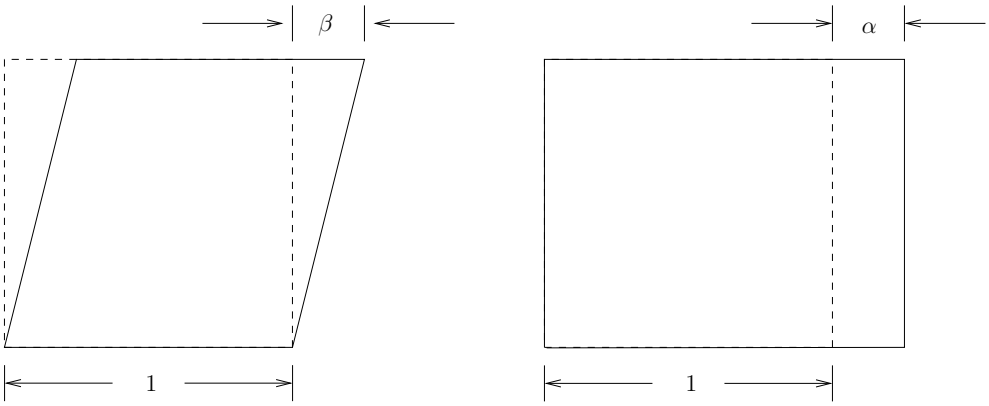

FIG. 11.1. *Pure shear and pure tensile strains.*

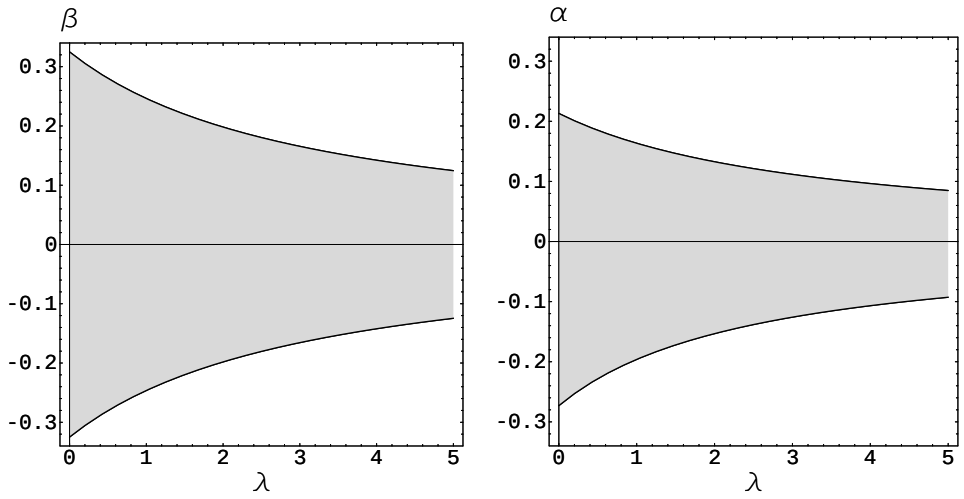

FIG. 11.2. *Shear and tensile strain limits for small strains.*

and, for this configuration to satisfy (11.2), the largest allowable λ is approximately 2.88, which corresponds to a Poisson ratio of $\nu = 0.37$. Figure 11.3 shows a plot of the deformed configuration.

REFERENCES

[1] C. BACUTA, J. BRAMBLE, AND J. XU, Regularity estimates for elliptic boundary value problems with smooth data on polygonal domains, J. Numer. Math. To Appear.

[2] ———, Regularity estimates for elliptic boundary value problems in besov spaces, Math. Comp., 72 (2003), pp. 1577–1599.

[3] S. BRENNER AND L. SCOTT, The Mathematical Theory of Finite Element Methods, Springer-Verlag, 1994.

[4] Z. CAI, J. KORSAWE, AND G. STARKE, An adaptive least squares mixed finite element method for the stress-displacement formulation of linear elasticity, Numer. Meth. PDEs, 21 (2005), pp. 132–148.

[5] Z. CAI, C.-O. LEE, T. MANTEUFFEL, AND S. MCCORMICK, First-order system least squares for linear elasticity: Further results, SIAM J. Sci. Comput., 21 (2000), pp. 1728–1739.

[6] ———, First-order system least squares for linear elasticity: Numerical results, SIAM J. Sci.

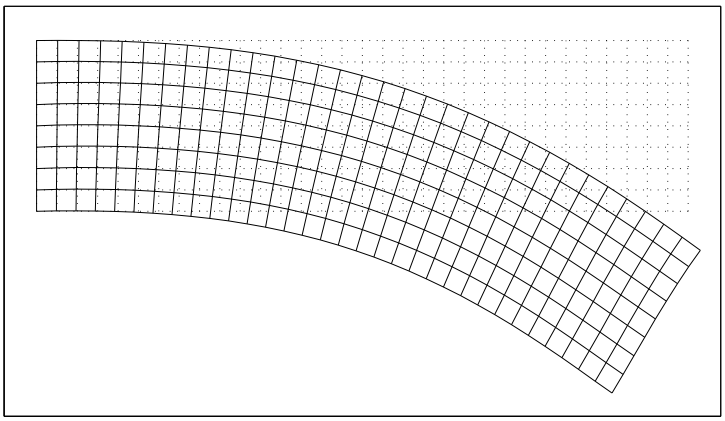

FIG. 11.3. *Displacement plot for cantilever beam displaying "large displacement and small strains".*

Comput., 21 (2000), pp. 1706–1727.
[7] Z. CAI, T. MANTEUFFEL, AND S. MCCORMICK, First-order system least squares for second-order partial differential equations: part ii, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.
[8] Z. CAI, T. MANTEUFFEL, S. MCCORMICK, AND S. PARTER, First-order system least squares (FOSLS) for planar linear elasticity: Pure traction problem, SIAM J. Numer. Anal., 35 (1998), pp. 320–335.
[9] Z. CAI AND G. STARKE, First-order system least squares for the stress-displacement formulation: Linear elasticity, SIAM J. Numer. Anal., 41 (2003), pp. 715–730.
[10] ———, Least squares methods for linear elasticity, SIAM J. Numer. Anal., 42 (2004), pp. 826–842.
[11] P. CIARLET, Mathematical Elasticity, Volume1: Three Dimensional Elasticity, North-Holland, 1988.
[12] S. CLAIN, Elliptic operators of divergence type with hölder coefficients in fractional sobolev spaces, Rend. Mat. Appl. (7), 17 (1997), pp. 207–236.
[13] A. CODD, T. MANTEUFFEL, AND S. MCCORMICK, Multilevel first-order system least squares for nonlinear elliptic partial differential equations, SIAM J. Numer. Anal., 41 (2003), pp. 2197–2209.
[14] V. GIRAULT AND P. RAVIART, Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms, Springer-Verlag, 1986.
[15] P. GRISVARD, Elliptic Problems in Nonsmooth Domains, Pitman, 1985.
[16] S. KIM, T. MANTEUFFEL, AND S. MCCORMICK, First-order system least squares (fosls) for spatial linear elasticity: Pure traction, SIAM J. Numer. Anal., 38 (2001), pp. 1454–1482.
[17] J. LIONS AND E. MAGENES, Non-Homogenous Boundary Value Problems and Applications, I, Springer-Verlag, New York, 1972.
[18] U. TROTTENBERG, C. OOSTERLEE, AND A. SCHÜLLER, Multigrid, Academic Press, 2001.
[19] C. WESTPHAL, First-Order System Least Squares (FOSLS) for Geometrically Nonlinear Elasticity in Nonsmooth Domains, PhD thesis, Univ. of Colorado, 2004.
[20] K. YOSIDA, Functional Analysis, 6th Ed., Springer-Verlag, 1980.