A FINITE ELEMENT METHOD FOR SECOND ORDER NONVARIATIONAL ELLIPTIC PROBLEMS

OMAR LAKKIS AND TRISTAN PRYER

Abstract. We propose a numerical method to approximate the solution of second order elliptic problems in nonvariational form. The method is of Galerkin type using conforming finite elements and applied directly to the nonvariational (nondivergence) form of a second order linear elliptic problem. The key tools are an appropriate concept of "finite element Hessian" and a Schur complement approach to solving the resulting linear algebra problem. The method is illustrated with computational experiments on three linear and one quasilinear PDE, all in nonvariational form.

1. Introduction

Finite element methods (FEM) arguably constitute one of the most successful method families in numerically approximating elliptic partial differential equations (PDE's) that are given in variational (also known as divergence) form.

For the reader's appreciation of this statement we briefly introduce standard FEM concepts. Let Ω be a given domain (open and bounded set) in $\mathbb{R}^d$, $d \in \mathbb{N}$, $f, a_{\alpha,\beta} = a_{\beta,\alpha} : \Omega \to \mathbb{R}$, be given functions with the appropriate regularity such that the operator $\operatorname{div}(\boldsymbol{A}\nabla u)$, for $\boldsymbol{A} := [a_{\alpha,\beta}]_{\alpha,\beta=1,\dots,d}$, makes sense, is elliptic and there is a unique function $u : \Omega \to \mathbb{R}$ satisfying $\operatorname{div}(\boldsymbol{A}\nabla u) = f$ with $u = 0$ on $\partial\Omega$ [GT83, for details]. The *classical solution*, u, of this problem can be characterized by first writing the PDE in *weak* (also known as *variational*) *form* using Green's formula:

$$u \in \mathscr{Y} \text{ and satisfies } a(u,v) := \int_\Omega \nabla u^\intercal \boldsymbol{A} \nabla v = \int_\Omega fv \quad \forall\, v \in \mathscr{X}, \tag{1.1}$$

where $\mathscr{X}$ and $\mathscr{Y}$ are appropriate (infinite dimensional) function spaces. A (finite) Galerkin procedure consists in finding an *approximation* of u, $U \in \mathbb{Y}$

$$A(U,V) = \langle f, V\rangle \quad \forall\, V \in \mathbb{X}, \tag{1.2}$$

where $\mathbb{Y}$ and $\mathbb{X}$ are finite dimensional "counterparts" (usually subspaces, but may be not) of $\mathscr{Y}$ and $\mathscr{X}$ and the bilinear form A an approximation of a. For example, when $a = A$ (modulo quadrature) $\mathscr{X} = \mathscr{Y} = \mathrm{H}^1_0(\Omega)$ and $\mathbb{X} = \mathbb{Y}$ are a space of continuous piecewise p-degree polynomial functions on a partition of Ω, we obtain the standard *conforming mesh-refinement (h-version) finite element method of degree p.*

The reason behind the FEM's success in such a framework is twofold: (1) the weak form is suitable to apply functional analytic frameworks (Lax–Milgram Theorem or Babuška–Brezzi–Ladyženskaya condition, e.g.), and (2) the discrete functions need to be differentiated at most once, whence weak smoothness requirements on the "elements".

In this article, we depart from this basis by considering second order elliptic boundary value problems (BVP's) in nonvariational form

$$\text{find } u \text{ such that } \boldsymbol{A}{:}\mathrm{D}^2 u = f \text{ in } \Omega \text{ and } \left.u\right|_{\partial\Omega} = g, \tag{1.3}$$

Date: November 3, 2018.

for which one may not always be successful in applying the standard FEM (with reference to §2 for the notation). Indeed, the use of the standard FEM requires (1) the coefficient matrix $\boldsymbol{A} : \Omega \to \mathbb{R}^{d\times d}$ to be (weakly) differentiable and (2) the rewriting of the second order term in divergence form, an operation which introduces an advection (first order) term:

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div}(\boldsymbol{A}\nabla u) - (\operatorname{div}(\boldsymbol{A}))\,\nabla u. \tag{1.4}$$

Even when coefficient matrix $\boldsymbol{A}$ is differentiable on Ω, this procedure could result in the problem becoming advection–dominated and unstable for conforming FEM, as we demonstrate numerically using Problem (4.5).

Our main motivation for studying linear elliptic BVP's in nonvariational form is their important role in pure and applied mathematics. An important example of nonvariational problems is the fully nonlinear BVP that is approximated via a Newton method which becomes an infinite sequence of linear nonvariational elliptic problems [Böh08].

In this article, we propose and test a direct discretization of the *strong form* (1.3) that makes no special assumption on the derivative of $\boldsymbol{A}$. The main idea, is an appropriate definition of a *finite element Hessian* given in §2.5. The finite element Hessian has been used earlier in different contexts, such as anisotropic mesh generation [AV02, CSX07, VMD$^+$07] and *finite element convexity* [AM08]. The finite element Hessian is related also to the finite element (discrete) elliptic operator appearing in the analysis of evolution problems [Tho06].

The method we propose is quite straightforward, and we are surprised that it is not easily available in the literature. It consists in discretizing, via a Galerkin procedure, the BVP (1.3) *directly without writing it in divergence form.*

The main difficulty of our approach is having to deal with a somewhat involved linear algebra problem that needs to be solved as efficiently as possible (this is especially important when we apply this method in the linearization of nonlinear elliptic BVP's). We overcame this difficulty in §3, by combining the definition of u's distributional Hessian,

$$\left\langle \mathrm{D}^2 u \,|\, \phi \right\rangle = -\left\langle \nabla u \otimes \nabla \phi \right\rangle + \left\langle \nabla u \otimes \boldsymbol{n}\ \phi \right\rangle_{\partial\Omega} \qquad \forall\, \phi \in \mathrm{C}^\infty(\Omega), \tag{1.5}$$

with equation (1.3) into a system of equations that are larger, but easier to handle numerically, once discretized.

It is worth noting that there are alternatives to our approach, most notably the standard finite difference method and its variants. The reason we are interested in a Galerkin procedure is the ability to use an unstructured mesh, essential for complicated geometries where the finite difference method leads to complicated, and sometimes prohibitive, modifications (especially in dimension 3 and higher), and the potential of dealing with adaptive methods, using available finite element code. Furthermore, our method has the potential to approach the iterative solution fully nonlinear problems where finite difference methods can become clumsy and demanding [KT92, LR05, Obe08, CS08].

This paper focuses mainly on the algorithmic and linear algebraic aspects of the method and is set out as follows. In §2 we introduce some notation and set out the model problem. We then present a discretization scheme for the model problem using standard conforming finite elements in $\mathrm{C}^0(\Omega)$. In §3 we present a linear algebra technique, inspired by the standard *Schur complement* idea, for solving the linear system arising from the discretization. Finally, in §4 we summarize extensive numerical experiments on model linear boundary value problems (BVPs) in nonvariational form and an application to quasilinear BVP in nonvariational form.

2. Set up

2.1. **Notation.** Let $\Omega \subset \mathbb{R}^d$ be an open and bounded Lipschitz domain. We denote $\mathrm{L}_2(\Omega)$ to be the space of square (Lebesgue) integrable functions on Ω together with it's inner product $\langle v, w\rangle := \int_\Omega vw$ and norm $\|v\| := \|v\|_{\mathrm{L}_2(\Omega)} = \langle v, v\rangle^{1/2}$. We also denote by $\langle f\rangle_\omega$ the integral of a function f over the domain ω and drop the subscript for $\omega = \Omega$.

We use the convention that the derivative $\mathrm{D}u$ of a function $u : \Omega \to \mathbb{R}$ is a row vector, while the gradient of u, ∇u is the derivative's transpose, i.e., $\nabla u = (\mathrm{D}u)^\intercal$. We will make use of the slight abuse of notation, following a common practice, whereby the Hessian of u is denoted as D^2u (instead of the correct $\nabla\mathrm{D}u$) and is represented by a $d \times d$ matrix.

The Sobolev spaces [Cia78, Eva98]

$$\mathrm{H}^k(\Omega) := \mathrm{W}^k_2(\Omega) = \left\{ \phi \in \mathrm{L}_2(\Omega) : \sum_{|\boldsymbol{\alpha}|\le k} \mathrm{D}^{\boldsymbol{\alpha}}\phi \in \mathrm{L}_2(\Omega) \right\}, \tag{2.1}$$

are equipped with norms and semi-norms

$$\|v\|_k^2 := \|v\|^2_{\mathrm{H}^k(\Omega)} = \sum_{|\boldsymbol{\alpha}|\le k} \|\mathrm{D}^{\boldsymbol{\alpha}} v\|^2 \tag{2.2}$$

$$\text{and } |v|_k^2 := |v|^2_{\mathrm{H}^k(\Omega)} = \sum_{|\boldsymbol{\alpha}|= k} \|\mathrm{D}^{\boldsymbol{\alpha}} v\|^2 \tag{2.3}$$

respectively, where $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_d\}$ is a multi-index, $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$ and derivatives $\mathrm{D}^{\boldsymbol{\alpha}}$ are understood in a weak sense. We pay particular attention to the cases $k = 1, 2$,

$$\mathrm{H}^1_0(\Omega) := \text{closure of } \mathrm{C}^\infty_0(\Omega) \text{ in } \mathrm{H}^1(\Omega) \tag{2.4}$$

$$\text{and } \mathrm{H}^{-1}(\Omega) := \text{dual}\left(\mathrm{H}^1_0(\Omega)\right). \tag{2.5}$$

We denote by $\langle v \,|\, w\rangle$ the action of a distribution v on the function w. If both $v, w \in \mathrm{L}_2(\Omega)$ then $\langle v \,|\, w\rangle = \langle v, w\rangle$.

We consider the following problem: Find $u \in \mathrm{H}^1_0(\Omega)$ such that

$$\begin{aligned} \mathscr{L}u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{2.6}$$

where the data $f : \Omega \to \mathbb{R}$ is prescribed and $\mathscr{L}$ is a general linear, second order, uniformly elliptic partial differential operator. Let $\boldsymbol{A} \in \mathrm{L}_\infty(\Omega)^{d\times d} \cap \mathrm{Sym}(\mathbb{R}^{d\times d})$, the space of bounded, symmetric, positive definite, $d \times d$ matrixes.

$$\begin{array}{rccc} \mathscr{L}: & \mathrm{H}^1_0(\Omega) & \to & \mathrm{H}^{-1}(\Omega) \\ & u & \mapsto & \mathscr{L}u := \boldsymbol{A}{:}\mathrm{D}^2u, \end{array} \tag{2.7}$$

we use $\boldsymbol{X}{:}\boldsymbol{Y} := \mathrm{trace}\,(\boldsymbol{X}^\intercal\boldsymbol{Y})$ to denote the Frobenius inner product between two matrixes.

2.2. **Discretization.** Let $\mathscr{T}$ be a conforming triangulation of Ω, namely, $\mathscr{T}$ is a finite family of sets such that

(1) $K \in \mathscr{T}$ implies K is an open simplex (segment for $d = 1$, triangle for $d = 2$, tetrahedron for $d = 3$),
(2) for any $K, J \in \mathscr{T}$ we have that $\overline{K} \cap \overline{J}$ is a full subsimplex (i.e., it is either $\emptyset$, a vertex, an edge, a face, or the whole of $\overline{K}$ and $\overline{J}$) of both $\overline{K}$ and $\overline{J}$ and
(3) $\bigcup_{K\in\mathscr{T}} \overline{K} = \overline{\Omega}$.

The *shape regularity* of $\mathscr{T}$ is defined as

$$\mu(\mathscr{T}) := \inf_{K\in\mathscr{T}} \frac{\rho_K}{h_K}, \tag{2.8}$$

where ρ_K is the radius of the largest ball contained inside K and h_K is the diameter of K. We use the convention where $h : \Omega \to \mathbb{R}$ denotes the *meshsize function* of $\mathscr{T}$, i.e.,

$$h(\boldsymbol{x}) := \max_{K\ni\boldsymbol{x}} h_K. \tag{2.9}$$

We introduce the *finite element spaces*

$$\mathbb{V} := \left\{\Phi \in \mathrm{H}^1(\Omega) :\ \Phi|_K \in \mathbb{P}^p\ \forall\, K \in \mathscr{T}\right\}, \tag{2.10}$$

$$\mathring{\mathbb{V}} := \mathbb{V} \cap \mathrm{H}^1_0(\Omega), \tag{2.11}$$

where $\mathbb{P}^k$ denotes the linear space of polynomials in d variables of degree no higher than a positive integer k. We consider $p \geq 1$ to be fixed and denote by $\mathring{N} := \dim \mathring{\mathbb{V}}$ and $N = \mathring{N} + N_\partial := \dim \mathbb{V}$. Let $\mathring{\boldsymbol{\Phi}} = (\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}})^\intercal$ and $\boldsymbol{\Phi} = (\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}}, \Phi_1, ..., \Phi_{N_\partial})^\intercal$ where $\{\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}}\}$ and $\{\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}}, \Phi_1, ..., \Phi_{N_\partial}\}$ form a basis of $\mathring{\mathbb{V}}$, $\mathbb{V}$ respectively.

Testing the model problem (2.6) with $\phi \in \mathrm{H}^1_0(\Omega)$ gives

$$\langle \mathscr{L}u, \phi\rangle = \left\langle \boldsymbol{A}{:}\mathrm{D}^2 u, \phi\right\rangle = \langle f, \phi\rangle\,. \tag{2.12}$$

In order to discretize (2.12) with $\mathbb{V}$ we use an appropriate definition of a Hessian of a finite element function. Such a function may not admit a Hessian in the classical sense, so we consider it as a distribution (or generalized function) which we recall the definition.

2.3. **Definition** (generalized Hessian)**.** Let $\boldsymbol{n} : \partial\Omega \to \mathbb{R}^d$ be the outward pointing normal of Ω. Given $v \in \mathrm{H}^1_0(\Omega)$ its *generalized Hessian* defined in the standard distributional sense is given by

$$\left\langle \mathrm{D}^2 v \,|\, \phi\right\rangle = -\left\langle \nabla v \otimes \nabla\phi\right\rangle + \left\langle \nabla v \otimes \boldsymbol{n}\ \phi\right\rangle_{\partial\Omega} \qquad \forall\, \phi \in \mathrm{C}^\infty(\Omega), \tag{2.13}$$

where we are using $\boldsymbol{x}\otimes\boldsymbol{y} := \boldsymbol{x}\boldsymbol{y}^\intercal$ to denote the tensor product between two geometric vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

2.4. **Theorem** (finite element Hessian)**.** *For each $V \in \mathring{\mathbb{V}}$ there exists a unique $\boldsymbol{H}[V] \in \mathbb{V}^{d\times d}$ such that*

$$\langle \boldsymbol{H}[V], \Phi\rangle = \left\langle \mathrm{D}^2 V \,|\, \Phi\right\rangle \qquad \forall\, \Phi \in \mathbb{V}. \tag{2.14}$$

Proof . Given a finite element function $V \in \mathring{\mathbb{V}}$, Definition 2.3 implies

$$\left\langle \mathrm{D}^2 V \,|\, \phi\right\rangle = -\left\langle \nabla V \otimes \nabla\phi\right\rangle + \left\langle \nabla V \otimes \boldsymbol{n}\ \phi\right\rangle_{\partial\Omega} \qquad \forall\, \phi \in \mathrm{C}^\infty(\Omega). \tag{2.15}$$

We fix V and let

$$\begin{array}{rccl} G: & \mathrm{C}^\infty(\Omega) & \to & \mathbb{R}^{d\times d} \\ & \phi & \mapsto & -\left\langle \nabla V \otimes \nabla\phi\right\rangle + \left\langle \nabla V \otimes \boldsymbol{n}\ \phi\right\rangle_{\partial\Omega}. \end{array} \tag{2.16}$$

Notice that G is a bounded linear functional on $\mathrm{C}^\infty(\Omega)$ in the $\mathrm{H}^1(\Omega)$-norm as,

$$|G(\phi)| = |\langle \nabla V \otimes \nabla\phi\rangle| + |\langle \nabla V \otimes \boldsymbol{n}\ \phi\rangle_{\partial\Omega}| \leq C(d,\Omega)\,\|V\|_1\,\|\phi\|_1\,. \tag{2.17}$$

Thus, due to the density of $\mathrm{C}^\infty(\Omega)$ in $\mathrm{H}^1(\Omega)$, G admits a unique extension, $\tilde{G}$.

Let $R = \tilde{G}\Big|_{\mathbb{V}}$ be the restriction of $\tilde{G}$ to $\mathbb{V}$. Since $\tilde{G}$ is linear and bounded on $\mathrm{H}^1(\Omega)$ it follows that R is linear and bounded on $\mathbb{V}$ in the $\mathrm{H}^1(\Omega)$-norm. Hence by Riesz's Representation Theorem there exists an $\boldsymbol{H}[V] \in \mathbb{V}^{d\times d}$ such that for each $\Phi \in \mathbb{V}$

$$\langle \boldsymbol{H}[V], \Phi\rangle := R(\Phi) = -\left\langle \nabla V \otimes \nabla\Phi\right\rangle + \left\langle \nabla V \otimes \boldsymbol{n}\ \Phi\right\rangle_{\partial\Omega}, \tag{2.18}$$

which coincides with the generalized Hessian (cf. Definition 2.3) on $\mathbb{V}$. □

2.5. **Definition** (finite element Hessian)**.** From Theorem 2.4 we define the *finite element Hessian* as follows. Let $V \in \mathring{\mathbb{V}}$ then

$$\langle \boldsymbol{H}[V], \Phi \rangle := -\langle \nabla V \otimes \nabla \Phi \rangle + \langle \nabla V \otimes \boldsymbol{n}\ \Phi \rangle_{\partial\Omega} \quad \forall\, \Phi \in \mathbb{V}. \tag{2.19}$$

It follows that $\boldsymbol{H}$ is a linear operator on $\mathring{\mathbb{V}}$.

Taking the model problem (2.12) we substitute the finite element Hessian directly, reducing the space of test functions to $\mathring{\mathbb{V}}$, we wish to find $U \in \mathring{\mathbb{V}}$ such that

$$\left\langle \boldsymbol{A}{:}\boldsymbol{H}[U], \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{2.20}$$

2.6. **Theorem** (nonvariational finite element method (NVFEM))**.** *The* nonvariational finite element solution *for the model problem's discretization (2.20) is given as* $U = \mathring{\boldsymbol{\Phi}}^{\mathsf{T}}\mathbf{u}$, *where* $\mathbf{u} \in \mathbb{R}^{\mathring{N}}$ *is the solution to the following linear system*

$$\mathbf{D}\mathbf{u} := \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathbf{B}^{\alpha,\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}\mathbf{u} = \mathbf{f}. \tag{2.21}$$

The components of (2.21) are given by

$$\mathbf{B}^{\alpha,\beta} := \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{A}^{\alpha,\beta}\boldsymbol{\Phi}^{\mathsf{T}} \right\rangle \in \mathbb{R}^{\mathring{N}\times N}, \tag{2.22}$$

$$\mathbf{M} := \langle \boldsymbol{\Phi}, \boldsymbol{\Phi}^{\mathsf{T}} \rangle \in \mathbb{R}^{N\times N}, \tag{2.23}$$

$$\mathbf{C}_{\alpha,\beta} := -\left\langle \partial_\beta \boldsymbol{\Phi}, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle + \left\langle \boldsymbol{\Phi} n_\beta, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle_{\partial\Omega} \in \mathbb{R}^{N\times \mathring{N}}, \tag{2.24}$$

$$\mathbf{f} := \left\langle f, \mathring{\boldsymbol{\Phi}} \right\rangle \in \mathbb{R}^{\mathring{N}}. \tag{2.25}$$

Proof . Since $\boldsymbol{H}[U] \in \mathbb{V}^{d\times d}$ for each $\alpha, \beta = 1, \ldots, d$, $\boldsymbol{H}_{\alpha,\beta}[U] = \boldsymbol{\Phi}^{\mathsf{T}}\mathbf{h}_{\alpha,\beta}$. Then, testing (2.20) with $\mathring{\boldsymbol{\Phi}}$,

$$\begin{aligned}
\left\langle f, \mathring{\boldsymbol{\Phi}} \right\rangle &= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \left\langle \boldsymbol{A}^{\alpha,\beta}\boldsymbol{H}_{\alpha,\beta}[U], \mathring{\boldsymbol{\Phi}} \right\rangle \\
&= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{A}^{\alpha,\beta}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{h}_{\alpha,\beta} \right\rangle \\
&= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{A}^{\alpha,\beta}\boldsymbol{\Phi}^{\mathsf{T}} \right\rangle \mathbf{h}_{\alpha,\beta}. \\
&= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathbf{B}^{\alpha,\beta}\mathbf{h}_{\alpha,\beta}
\end{aligned} \tag{2.26}$$

Utilizing Definition 2.5 for each $\alpha, \beta = 1 \ldots d$ we can compute $\mathbf{h}_{\alpha,\beta} \in \mathbb{R}^N$, noting $U = \mathring{\boldsymbol{\Phi}}^{\mathsf{T}}\mathbf{u}$,

$$\begin{aligned}
\langle \boldsymbol{\Phi}, \boldsymbol{\Phi}^{\mathsf{T}} \rangle\, \mathbf{h}_{\alpha,\beta} &= \langle \boldsymbol{\Phi}, \boldsymbol{H}_{\alpha,\beta}[U] \rangle \\
&= -\langle \partial_\beta \boldsymbol{\Phi}, \partial_\alpha U \rangle + \langle \boldsymbol{\Phi}\boldsymbol{n}_\beta, \partial_\alpha U \rangle_{\partial\Omega} \\
&= \left(-\left\langle \partial_\beta \boldsymbol{\Phi}, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle + \left\langle \boldsymbol{\Phi}\boldsymbol{n}_\beta, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle_{\partial\Omega} \right) \mathbf{u}.
\end{aligned} \tag{2.27}$$

Using the definition of $\mathbf{C}_{\alpha,\beta}$ (2.24) and $\mathbf{M}$ (2.23) we see for each $\alpha, \beta = 1 \ldots d$

$$\begin{aligned}
\mathbf{M}\mathbf{h}_{\alpha,\beta} &= \mathbf{C}_{\alpha,\beta}\mathbf{u} \\
\mathbf{h}_{\alpha,\beta} &= \mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}\mathbf{u}.
\end{aligned} \tag{2.28}$$

Substituting $\mathbf{h}_{\alpha,\beta}$ from (2.28) into (2.26) we obtain the desired result. □

2.7. **Example** (for $d = 2$)**.** For a general elliptic operator in 2-D, the formulation (2.21) takes the form

$$\left(\mathbf{B}^{1,1}\mathbf{M}^{-1}\mathbf{C}_{1,1} + \mathbf{B}^{2,2}\mathbf{M}^{-1}\mathbf{C}_{2,2} + \mathbf{B}^{1,2}\mathbf{M}^{-1}\mathbf{C}_{1,2} + \mathbf{B}^{2,1}\mathbf{M}^{-1}\mathbf{C}_{2,1}\right)\mathbf{u} = \mathbf{f} \tag{2.29}$$

3. Solving the linear system

3.1. **Remark** ((2.21) is difficult to solve)**.** Looking at the full system setting $\mathbf{D} = \sum\sum \mathbf{B}^{\alpha,\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}$ multiplying out each of the matrixes and proceeding to solve $\mathbf{Du} = \mathbf{f}$ the resulting system would not be sparse forcing the use of direct solvers.

In this section we will present a method to solve formulation (2.21) in a general setting. This method makes use of the sparsity of the component matrixes $\mathbf{B}^{\alpha,\beta}, \mathbf{C}^{\alpha,\beta}$ and $\mathbf{M}$.

3.2. **Remark.** An interesting point of note is that if the mass matrix $\mathbf{M}$ were diagonalized, by mass lumping, then for each α and β the matrix $\mathbf{B}^{\alpha,\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}$ would still be sparse (albeit less so than the individual matrixes $\mathbf{B}^{\alpha,\beta}$ and $\mathbf{C}_{\alpha,\beta}$). Hence the system can be easily solved using existing sparse methods. However mass lumping is only applicable to $\mathbb{P}^1$ finite elements. For higher order finite elements it would be desirable to exploit the sparse structure of the component matrixes that make up the system.

3.3. **A generalized Schur complement.** We observe the matrix $\mathbf{D}$ in the system (2.21) is a sum of Schur complements $\mathbf{B}^{\alpha,\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}$. With that in mind we introduce the $(d^2+1)^2$ block matrix

$$\mathbf{E} = \begin{bmatrix} \mathbf{M} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{C}_{1,1} \\ \mathbf{0} & \mathbf{M} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{C}_{1,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M} & \mathbf{0} & -\mathbf{C}_{d,d-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{M} & -\mathbf{C}_{d,d} \\ \mathbf{B}^{1,1} & \mathbf{B}^{1,2} & \dots & \mathbf{B}^{d,d-1} & \mathbf{B}^{d,d} & \mathbf{0} \end{bmatrix}. \tag{3.1}$$

3.4. **Lemma** (generalized Schur complement)**.** *Given*

$$\mathbf{v} = (\mathbf{h}_{1,1}, \mathbf{h}_{1,2}, \dots, \mathbf{h}_{d,d-1}, \mathbf{h}_{d,d}, \mathbf{u})^{\mathsf{T}}, \tag{3.2}$$

$$\mathbf{b} = (\mathbf{0}, \mathbf{0} \dots, \mathbf{0}, \mathbf{0}, \mathbf{f})^{\mathsf{T}}, \tag{3.3}$$

solving the system

$$\mathbf{Du} = \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathbf{B}^{\alpha,\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}\mathbf{u} = \mathbf{f}, \tag{3.4}$$

is equivalent to solving

$$\mathbf{Ev} = \mathbf{b}. \tag{3.5}$$

for $\mathbf{u}$.

Proof . The proof is just block Gaussian elimination on $\mathbf{E}$. Left-multiplying the first d^2 rows by $\mathbf{M}^{-1}$ yields

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{M}^{-1}\mathbf{C}_{1,1} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{M}^{-1}\mathbf{C}_{1,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} & -\mathbf{M}^{-1}\mathbf{C}_{d,d-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{I} & -\mathbf{M}^{-1}\mathbf{C}_{d,d} \\ \mathbf{B}^{1,1} & \mathbf{B}^{1,2} & \dots & \mathbf{B}^{d,d-1} & \mathbf{B}^{d,d} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{1,1} \\ \mathbf{h}_{1,2} \\ \vdots \\ \mathbf{h}_{d,d-1} \\ \mathbf{h}_{d,d} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{f} \end{bmatrix}. \tag{3.6}$$

Multiplying the i-th row by the i-th entry of the (d^2+1)-th row for $i = 1, \dots, d^2$

$$\begin{bmatrix} \mathbf{B}^{1,1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{B}^{1,1}\mathbf{M}^{-1}\mathbf{C}_{1,1} \\ \mathbf{0} & \mathbf{B}^{1,2} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{B}^{1,2}\mathbf{M}^{-1}\mathbf{C}_{1,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}^{d,d-1} & \mathbf{0} & -\mathbf{B}^{d,d-1}\mathbf{M}^{-1}\mathbf{C}_{d,d-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{B}^{d,d} & -\mathbf{B}^{d,d}\mathbf{M}^{-1}\mathbf{C}_{d,d} \\ \mathbf{B}^{1,1} & \mathbf{B}^{1,2} & \dots & \mathbf{B}^{d,d-1} & \mathbf{B}^{d,d} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{1,1} \\ \mathbf{h}_{1,2} \\ \vdots \\ \mathbf{h}_{d,d-1} \\ \mathbf{h}_{d,d} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{f} \end{bmatrix}. \tag{3.7}$$

Subtracting each of the first d^2 rows from the (d^2+1)-th row reduces the system into row echelon form.

$$\begin{bmatrix} \mathbf{B}^{1,1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{B}^{1,1}\mathbf{M}^{-1}\mathbf{C}_{1,1} \\ \mathbf{0} & \mathbf{B}^{1,2} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{B}^{1,2}\mathbf{M}^{-1}\mathbf{C}_{1,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}^{d,d-1} & \mathbf{0} & -\mathbf{B}^{d,d-1}\mathbf{M}^{-1}\mathbf{C}_{d,d-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{B}^{d,d} & -\mathbf{B}^{d,d}\mathbf{M}^{-1}\mathbf{C}_{d,d} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{1,1} \\ \mathbf{h}_{1,2} \\ \vdots \\ \mathbf{h}_{d,d-1} \\ \mathbf{h}_{d,d} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{f} \end{bmatrix}. \tag{3.8}$$

□

3.5. **Remark** (structure of the block matrix)**.** In fact this method for the solution of the system $\mathbf{Du} = \mathbf{f}$ is not surprising given the discretization presented in the proof of Theorem 2.6 is equivalent to the following system:

$$\text{Find } U \in \mathring{\mathbb{V}} \text{ such that } \begin{cases} \langle \boldsymbol{H}[U], \Phi \rangle = -\langle \nabla U \otimes \nabla \Phi \rangle + \langle \nabla U \otimes \boldsymbol{n}\ \Phi \rangle_{\partial\Omega} & \forall\, \Phi \in \mathbb{V} \\ \left\langle \boldsymbol{A} : \boldsymbol{H}[U], \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle & \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \end{cases} \tag{3.9}$$

3.6. **Remark** (enforcing non-trivial Dirichlet boundary values)**.** Given additional problem data $g \in \mathrm{H}^{1/2}(\Omega)$, to solve

$$\begin{aligned} \mathscr{L} u &= f \text{ in } \Omega, \\ u &= g \text{ on } \partial\Omega, \end{aligned} \tag{3.10}$$

it is not immediate how to enforce the boundary conditions. If we were solving the full system $\mathbf{Du} = \mathbf{f}$, we could directly enforce them into the system matrix.

Since $g \in \mathrm{H}^{1/2}(\Omega)$ by an embedding it is continuous and can be approximated by the Lagrange interpolant with optimal order. To enforce the Dirichlet boundaries we introduce a further block representation

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{E}_\partial & \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{v}_\partial \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_\partial \\ \mathbf{b} \end{bmatrix}, \tag{3.11}$$

where $\mathbf{E}, \mathbf{v}$ and $\mathbf{b}$ are defined as before and $\mathbf{E}_\partial, \mathbf{v}_\partial$ and $\mathbf{b}_\partial$ are defined as follows

$$\mathbf{E}_\partial = \begin{bmatrix} \mathbf{M} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{C}^\partial_{1,1} \\ \mathbf{0} & \mathbf{M} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{C}^\partial_{1,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M} & \mathbf{0} & -\mathbf{C}^\partial_{d,d-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{M} & -\mathbf{C}^\partial_{d,d} \\ \mathbf{B}^{1,1} & \mathbf{B}^{1,2} & \dots & \mathbf{B}^{d,d-1} & \mathbf{B}^{d,d} & \mathbf{0} \end{bmatrix}, \tag{3.12}$$

$$\mathbf{v}_\partial = \left[\mathbf{h}^\partial_{1,1}, \mathbf{h}^\partial_{1,2}, \dots, \mathbf{h}^\partial_{d,d-1}, \mathbf{h}^\partial_{d,d}, \mathbf{u}^\partial\right]^\mathsf{T}, \tag{3.13}$$

$$\mathbf{b}_\partial = [\mathbf{0}, \mathbf{0}, \dots, \mathbf{0}, \mathbf{0}, \mathbf{g}]^\mathsf{T}. \tag{3.14}$$

Let $\mathbf{\Phi}_\partial = \{\Phi_1, \dots, \Phi_{N_\partial}\}$, then the components of $\mathbf{E}_\partial$ and $\mathbf{b}_\partial$ are defined as follows

$$\mathbf{C}^\partial_{\alpha,\beta} = -\langle \partial_\beta \mathbf{\Phi}, \partial_\alpha \mathbf{\Phi}_\partial{}^\mathsf{T} \rangle + \langle \mathbf{\Phi} \boldsymbol{n}_\beta, \partial_\alpha \mathbf{\Phi}_\partial{}^\mathsf{T} \rangle_{\partial\Omega} \in \mathbb{R}^{N \times N_\partial}, \tag{3.15}$$

$$\mathbf{g}_j = g(x_j)\Phi_j \in \mathbb{R}^{N_\partial}, \tag{3.16}$$

where x_j is the Lagrange node associated with Φ_j.

The block matrix (3.11) can then be trivially solved

$$\mathbf{E}\mathbf{v} = \mathbf{b} - \mathbf{E}_\partial \mathbf{b}_\partial. \tag{3.17}$$

3.7. **Remark** (storage issues)**.** We will be using the generalized minimal residual method (GMRES) to solve this system. The GMRES, as with any iterative solver, only requires an algorithm to compute a matrix-vector multiplication. Hence we are only required to store the component matrixes $\mathbf{B}^{\alpha,\beta}, \mathbf{C}_{\alpha,\beta}$ and $\mathbf{M}$.

3.8. **Remark** (condition number)**.** The convergence rate of an iterative solver applied to a linear system $\mathbf{N}\mathbf{v} = \mathbf{g}$ will depend on the condition number $\kappa(\mathbf{N})$, defined as the ratio of the maximum and minimum eigenvalues of $\mathbf{N}$:

$$\kappa(\mathbf{N}) = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{3.18}$$

Numerically we observe the condition number of the block matrix $\kappa(\mathbf{E}) \leq Ch^{-2}$ (see Table 1).

4. Numerical applications

In this section we study the numerical behavior of the scheme presented above. All our computations were carried out in Matlab® (code available on request).

We present two linear benchmark problems, for which the solution is known. We take Ω to be the square $S = (-1, 1) \times (-1, 1) \subset \mathbb{R}^2$ and in the first two tests consider the operator

$$\boldsymbol{A}(\boldsymbol{x}) = \begin{bmatrix} 1 & b(\boldsymbol{x}) \\ b(\boldsymbol{x}) & a(\boldsymbol{x}) \end{bmatrix} \tag{4.1}$$

varying the coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x})$.

4.1. **Test problem with a nondifferentiable operator.** For the first test problem we choose the operator in such a way that (1.4) does not hold, that is the components of $\boldsymbol{A}$ are non-differentiable on Ω, in this case we take

$$a(\boldsymbol{x}) = (x_1^2 x_2^2)^{1/3} + 1 \tag{4.2}$$

$$b(\boldsymbol{x}) = 0. \tag{4.3}$$

A visualization of the operator (4.2) is given in Figure 2(a). We choose our problem data f such that the exact solution to the problem is given by:

$$u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|^2). \tag{4.4}$$

We discretize the problem given by (4.2) under the algorithm set out in §2.2, numerical convergence results are shown in Figure 2.

4.2. **Test problem with convection dominated operator.** The second test problem demonstrates the ability to overcome oscillations introduced into the standard finite element when rewriting the operator in divergence form. Take

$$a(\boldsymbol{x}) = \arctan\left(K(|\boldsymbol{x}|^2 - 1)\right) + 2 \tag{4.5}$$

$$b(\boldsymbol{x}) = 0. \tag{4.6}$$

with $K \in \mathbb{R}^+$. Rewriting in divergence form gives

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div}(\boldsymbol{A}\nabla u) - \operatorname{div}(\boldsymbol{A})\,\nabla u. \tag{4.7}$$

The derivatives

$$\partial_\alpha a(\boldsymbol{x}) = \frac{dKx_\alpha}{1 + K\left(|\boldsymbol{x}|^2 - 1\right)} \tag{4.8}$$

can be made arbitrarily large on the unit circle by choosing K appropriately (see Figure 2(b)).

We choose our problem data f such that the exact solution to the problem is given by:

$$u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2). \tag{4.9}$$

We then construct the standard finite element method around (4.7), that is find $U \in \mathring{\mathbb{V}}$ such that for each $\mathring{\Phi} \in \mathring{\mathbb{V}}$

$$\left\langle \boldsymbol{A}\nabla U, \nabla\mathring{\Phi}\right\rangle - \left\langle \operatorname{div}(\boldsymbol{A})\,\nabla U, \mathring{\Phi}\right\rangle = \left\langle f, \mathring{\Phi}\right\rangle. \tag{4.10}$$

If K is chosen small enough the standard finite element method converges optimally. If we increase the value of K oscillations become apparent in the finite element solution along the unit circle. Figure 4 demonstrates the oscillations arising from this method compared to discretizing using the nonvariational finite element method.

Figure 3 shows the numerical convergence rates of the nonvariational finite element method applied to this problem.

4.3. **Test problem choosing a solution with nonsymmetric Hessian.** In this test we choose the operator such that $b(\boldsymbol{x})$ is non-zero. To maintain ellipticity in this problem we must choose $a(\boldsymbol{x})$ such that the trace of $\boldsymbol{A}$ dominates it's determinant. We choose

$$a(\boldsymbol{x}) = 2 \tag{4.11}$$

$$b(\boldsymbol{x}) = (x_1^2 x_2^2)^{1/3}. \tag{4.12}$$

We choose the problem data such that the exact solution is given by

$$u(\boldsymbol{x}) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2} & \boldsymbol{x} \neq \mathbf{0} \\ 0 & \boldsymbol{x} = \mathbf{0}. \end{cases} \tag{4.13}$$

This function has a nonsymmetric Hessian at the point $\mathbf{0}$. The nontrivial Dirichlet boundary is dealt with using Remark 3.6. Figure 5 shows numerical results for this problem.

4.4. **Test problem with quasilinear PDE in nondivergence form.** The problem under consideration in this test is the following quasi-linear PDE arising from differential geometry:

$$\operatorname{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right) = \frac{f}{\sqrt{1+|\nabla u|^2}}, \tag{4.14}$$

where $\sqrt{1+|\nabla u|^2}$ is the area element. Here we are using $|\nabla u|^2 = \mathrm{D}u\nabla u$. Applying a fixed point linearization given an initial guess u^0 for each $n \in \mathbb{N}$ we seek u^n such that

$$\operatorname{div}\left(\frac{\nabla u^n}{\sqrt{1+|\nabla u^{n-1}|^2}}\right) = \frac{f}{\sqrt{1+|\nabla u^{n-1}|^2}}. \tag{4.15}$$

Applying a standard finite element discretization of (4.15) yields: Given $U^0 \in \mathring{\mathbb{V}}$, for each $n \in \mathbb{N}$ find $U^n \in \mathring{\mathbb{V}}$ such that for each $\mathring{\Phi} \in \mathring{\mathbb{V}}$

$$\left\langle \frac{\nabla U^n}{\sqrt{1+|\nabla U^{n-1}|^2}}, \nabla\mathring{\Phi}\right\rangle = \left\langle \frac{f}{\sqrt{1+|\nabla U^{n-1}|^2}}, \mathring{\Phi}\right\rangle. \tag{4.16}$$

In fact we can work on this problem combining the two nonlinear terms. To do so we must first rewrite (4.14) into the form $A(u,\nabla u){:}\mathrm{D}^2u = f$.

$$\begin{aligned} f =& \sqrt{1+|\nabla u|^2}\operatorname{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right)\\ =& \sqrt{1+|\nabla u|^2}\left(\frac{\Delta u}{\sqrt{1+|\nabla u|^2}} + \frac{\mathrm{D}\left(1+|\nabla u|^2\right)}{2\left(1+|\nabla u|^2\right)^{3/2}}\nabla u\right)\\ =& \Delta u + \frac{\mathrm{D}u\mathrm{D}^2u\nabla u}{1+|\nabla u|^2}\\ =& \left(\boldsymbol{I} + \frac{\nabla u\mathrm{D}u}{1+|\nabla u|^2}\right){:}\mathrm{D}^2u. \end{aligned} \tag{4.17}$$

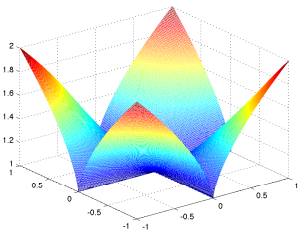
Applying a similar fixed point linearization given an initial guess u^0 for each $n \in \mathbb{N}$ we seek u^n such that

$$\left(\boldsymbol{I} + \frac{\nabla u^{n-1}\mathrm{D}u^{n-1}}{1+|\nabla u^{n-1}|^2}\right){:}\mathrm{D}^2u^n = f \tag{4.18}$$

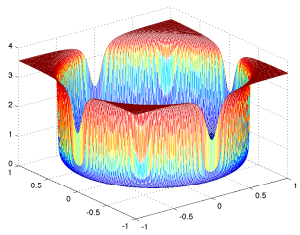
Discretizing the problem is then similar to that set out in Section 2.2. The component matrixes M and $\mathsf{C}_{\alpha,\beta}$ are problem independent, $\mathsf{B}^{\alpha,\beta}$ are defined as

$$\mathsf{B}^{\alpha,\beta} = \begin{cases} \left\langle \mathring{\boldsymbol{\Phi}}, 1 + \frac{\partial_\alpha U^{n-1}\partial_\beta U^{n-1}}{1+|\nabla U^{n-1}|^2}\boldsymbol{\Phi}\right\rangle, & \text{for } \alpha = \beta,\\ \left\langle \mathring{\boldsymbol{\Phi}}, \frac{\partial_\alpha U^{n-1}\partial_\beta U^{n-1}}{1+|\nabla U^{n-1}|^2}\boldsymbol{\Phi}\right\rangle, & \text{for } \alpha \neq \beta. \end{cases} \tag{4.19}$$

FIGURE 1. A visualization of the coefficient of the operators (4.2) (on the left) and (4.5) (on the right).

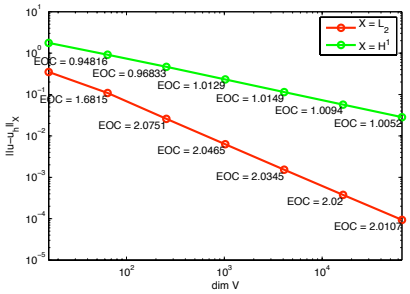


(a) The function $(x_1^2 x_2^2)^{1/3} + 1$ over Ω. Note the derivatives are singular at $x_1 = 0$ and $x_2 = 0$.

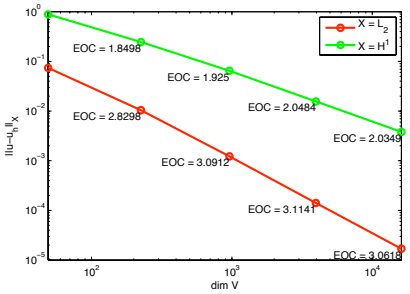


(b) The function $\arctan\left(5000(|x|^2 - 1)\right)$ over Ω. Note the derivatives are very large on the unit circle.

FIGURE 2. Test 4.1. Errors and convergence rates for the NVFEM applied to a non-divergence form operator (4.2), choosing f appropriately such that $u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|)$. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements

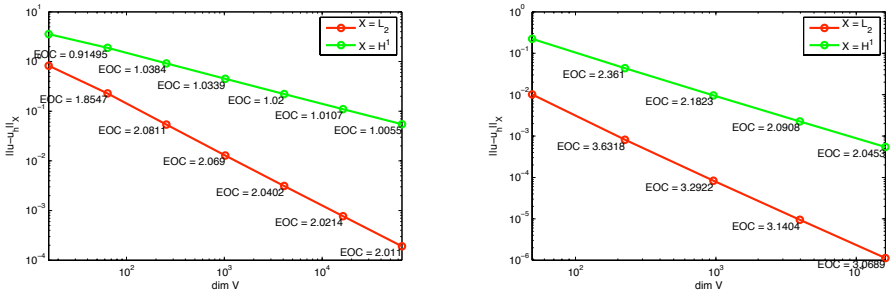


(b) $\mathbb{P}^2$-elements

TABLE 1. Test 4.1. On the condition number of $\mathbf{E}$ upon discretizing problem (4.2) using $\mathbb{P}^1$ finite elements. As claimed in Remark 3.8 $\kappa(\mathbf{E}) \approx Ch^{-2}$.

$\dim \mathbb{V}$	h	$\kappa(\mathbf{E})$	$h^{-2}\kappa(\mathbf{E})$
16	0.4714	4.904×10^1	10.898
64	0.202	6.594×10^2	26.952
256	0.0943	3.665×10^3	32.633
1024	0.0456	1.722×10^4	35.833
4096	0.0224	6.894×10^4	34.737
16384	0.0111	3.383×10^5	41.949
65536	0.0055	1.337×10^6	40.43

FIGURE 3. Test 4.2. Errors and convergence rates for the NVFEM applied to a non-divergence form operator (4.5) with $K = 5000$, choosing f appropriately such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

FIGURE 4. Test 4.2. On the left we present $\left\|u - \tilde{U}\right\|_{\mathrm{L}_\infty(K)}$ plotted on a logarithmic scale as a function over Ω. This represents the maximum error of the standard FE-solution, $\tilde{U}$, to problem (4.5) with 16384 DOF's ($h = 1/32$). Notice the oscillations apparent on the unit circle. On the right we show $\|u - U\|_{\mathrm{L}_\infty(K)}$ plotted on a logarithmic scale as a function over Ω, the maximum error of the NVFE-solution, U, to problem (4.5) with 16384 DOF's ($h = 1/32$).

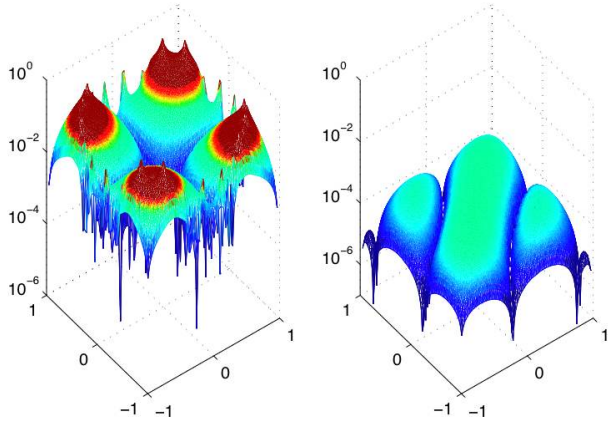


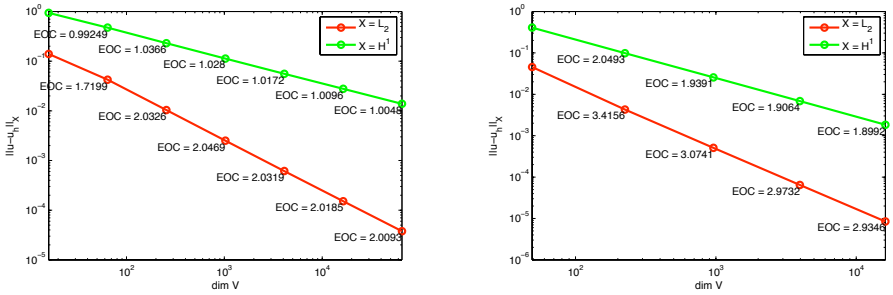
Table 2 compares the two linearizations (4.15) and (4.18). Figure 6 show asymptotic numerical convergence results for NVFEM applied to (4.18).

REFERENCES

[AM08] Néstor E. Aguilera and Pedro Morin. On convex functions and the finite element method. online preprint arXiv:0804.1780v1, arXiv.org, Apr 2008.

[AV02] A. Agouzal and Yu. Vassilevski. On a discrete Hessian recovery for P_1 finite elements. *J. Numer. Math.*, 10(1):1–12, 2002.

FIGURE 5. Test 4.3. Errors and convergence rates for the NVFEM on an operator (4.11), choosing f appropriately such that $u(\boldsymbol{x}) = \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2}$ if $\boldsymbol{x} \neq \mathbf{0}$, or $u(\boldsymbol{x}) = 0$ otherwise. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

TABLE 2. Test 4.4. Comparison of the fixed point linearization in variational form (4.15) and in nonvariational form (4.18). We fix f appropriately such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$. Taking initial guesses $U^0 = \tilde{U}^0 = 0$ we discretize problem (4.14) using a standard FEM and using the NVFEM. Denoting U_i and $\tilde{U}_i$ to be the NVFE-solution FE-solution respectively we run both linearizations for until a tolerance $\|U_{n+1} - U_n\|$ (resp. $\left\|\tilde{U}_{n+1} - \tilde{U}_n\right\|$) $\leq h^2$ is achieved. We compute both the stagnation point—which is the iteration at which the prescribed tolerance is achieved—and the total CPU time. Notice there is significant savings in the number of iterations required to reach the stagnation point using the NVFEM over the standard FEM, however each iteration is computationally more costly using the NVFEM since the system is larger and more complicated to solve. The CPU cost for the entire algorithm is comparable for each fixed h.

	h	$\sqrt{2}/5$	$\sqrt{2}/10$	$\sqrt{2}/20$	$\sqrt{2}/40$	$\sqrt{2}/80$	$\sqrt{2}/160$
FEM	Stag. Point	5	13	16	26	32	36
	CPU Time	0.50	4.02	17.51	117.58	796.58	5308.81
NDFEM	Stag. Point	4	6	7	8	10	12
	CPU Time	0.72	3.40	16.49	97.93	838.8	5256.84

[Böh08] Klaus Böhmer. On finite element methods for fully nonlinear elliptic equations of second order. *SIAM J. Numer. Anal.*, 46(3):1212–1249, 2008.

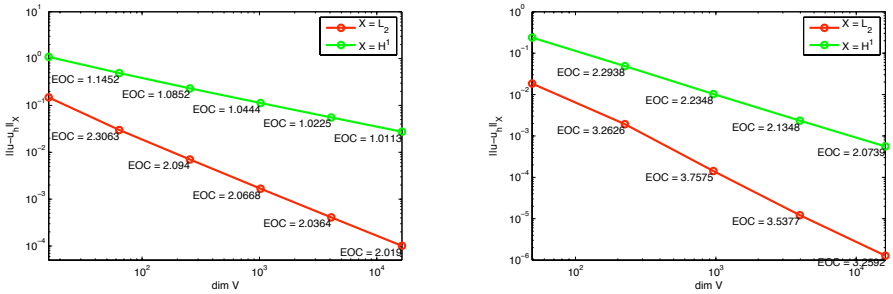
[Cia78] Philippe G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.

[CS08] Luis A. Caffarelli and Panagiotis E. Souganidis. A rate of convergence for monotone finite difference approximations to fully nonlinear, uniformly elliptic PDEs. *Comm. Pure Appl. Math.*, 61(1):1–17, 2008.

[CSX07] Long Chen, Pengtao Sun, and Jinchao Xu. Optimal anisotropic meshes for minimizing interpolation errors in L^p-norm. *Math. Comp.*, 76(257):179–204 (electronic), 2007.

[Eva98] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

FIGURE 6. Test 4.4. Errors and convergence rates for NVFEM applied to (4.14), a quasi-linear PDE under a fixed point linearization. We fix f appropriately such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$, taking an initial guess $u^0 = 0$. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

[GT83] David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, Berlin, second edition, 1983.

[KT92] Hung Ju Kuo and Neil S. Trudinger. Discrete methods for fully nonlinear elliptic equations. *SIAM J. Numer. Anal.*, 29(1):123–135, 1992.

[LR05] Grégoire Loeper and Francesca Rapetti. Numerical solution of the Monge-Ampère equation by a Newton's algorithm. *C. R. Math. Acad. Sci. Paris*, 340(4):319–324, 2005.

[Obe08] Adam M. Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. *Discrete Contin. Dyn. Syst. Ser. B*, 10(1):221–238, 2008.

[Tho06] Vidar Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[VMD+07] M.-G. Vallet, C.-M. Manole, J. Dompierre, S. Dufour, and F. Guibault. Numerical comparison of some Hessian recovery techniques. *Internat. J. Numer. Methods Engrg.*, 72(8):987–1007, 2007.

OMAR LAKKIS
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SUSSEX
BRIGHTON
UK-BN1 9RF, UNITED KINGDOM
E-mail address: o.lakkis@sussex.ac.uk
URL: http://www.maths.sussex.ac.uk/Staff/OL

TRISTAN PRYER
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SUSSEX
BRIGHTON
UK-BN1 9RF, UNITED KINGDOM
E-mail address: tmp20@sussex.ac.uk
URL: http://www.maths.sussex.ac.uk/~tristan