A finite element method for second order nonvariational elliptic problems

SIAM J. SCI. COMPUT.
Vol. 33, No. 2, pp. 786–801

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 quasi-linear PDE, all in nonvariational form.

Key words. finite element method, nonvariational form second order elliptic PDE, Hessian recovery, Schur complement

AMS subject classifications. 65N30, 65Y20, 65F99

DOI. 10.1137/100787672

1. Introduction.

Finite element methods (FEMs) arguably constitute one of the most successful method families in numerically approximating elliptic partial differential equations (PDEs) 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 $\Omega \subset \mathbb{R}^d$ 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 its inner product $\langle v, w\rangle := \int_\Omega vw$ and norm $\|v\| := \|v\|_{\mathrm{L}_2(\Omega)} = \langle v, v\rangle^{1/2}$. 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 also denote by $\langle f\rangle_\omega$ the integral of a function f over the domain ω and drop the subscript for $\omega = \Omega$. Suppose $f, a_{\alpha,\beta} = a_{\beta,\alpha} : \Omega \to \mathbb{R}$ are given functions with the appropriate regularity (resp., $\mathrm{L}_2(\Omega)$ and $\mathrm{L}_\infty(\Omega)$) such that the operator $\operatorname{div}(\boldsymbol{A}\nabla u)$, for $\boldsymbol{A} := [a_{\alpha,\beta}]_{\alpha,\beta=1,\dots,d}$, makes sense and is elliptic, and such that there is a unique function $u : \Omega \to \mathbb{R}$ satisfying $\operatorname{div}(\boldsymbol{A}\nabla u) = f$ with $u = 0$ on $\partial\Omega$ (see [GT01] for details). (The boundary condition $u = 0$, taken in this paragraph for simplicity, is generalizable to $u = g$.) 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, $U \in \mathbb{Y}$

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

*Submitted to the journal's Methods and Algorithms for Scientific Computing section March 5, 2010; accepted for publication (in revised form) January 10, 2011; published electronically April 5, 2011.

http://www.siam.org/journals/sisc/33-2/78767.html

†Department of Mathematics, University of Sussex, Brighton, UK, GB-BN1 9QH (O.Lakkis@sussex.ac.uk).

‡School of Mathematics, Statistics & Actuarial Science, University of Kent, Canterbury, UK, GB-CT2 7NF (T.Pryer@kent.ac.uk).

where $\mathbb{Y}$ and $\mathbb{X}$ are finite-dimensional "counterparts" (usually subspaces, but not necessarily) of $\mathscr{Y}$ and $\mathscr{X}$ and the bilinear form A is 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 degree p polynomial functions (also known as conforming $\mathbb{P}^p$ elements) on a partition of Ω, when p is a fixed number, we obtain the standard *conforming mesh-refinement (h-version) FEM 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, e.g.), and (2) the discrete functions need to be differentiated at most once, whence weak smoothness requirements on the "elements."

In this paper 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 by $\mathrm{D}^2 u$ (instead of the correct $\nabla \mathrm{D}u$) and is represented by a $d \times d$ matrix.

We depart from the basis of the standard FEM and consider second order elliptic boundary value problems (BVPs) in nonvariational form:

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

for which one may not always be successful in applying the standard FEM.

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 a convection (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 the coefficient matrix $\boldsymbol{A}$ is differentiable on Ω, this procedure could result in the problem becoming convection-dominated and unstable for conforming FEM, as we demonstrate numerically using problem (4.4).

Our main motivation for studying linear elliptic BVPs 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 section 2.1. The finite element Hessian has been used earlier in different contexts, such as anisotropic mesh generation [AV02, CSX07, VMD+07] and *finite element convexity* [AM09]. 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 does not seem to be available in the literature. It consists in discretizing, via a Galerkin procedure, the BVP (1.3) *directly without writing it in divergence form*. In this paper, we will consider only conforming finite element spaces, while noting that there is potential to expand the method to nonconforming spaces.

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

BVPs). We overcame this difficulty in section 3, by combining the definition of u's Hessian as a tempered (i.e., up to the boundary) distribution,

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

with the nonvariational problem (1.3) into a system of equations that are larger, but easier to handle numerically, once discretized. We thus call this method the *nonvariational finite element method* (NVFEM).

To clarify our nonstandard notation, the second term on the right-hand side of (1.5) is

$$\left\langle \nabla u \otimes \boldsymbol{n}\ \phi \right\rangle_{\partial\Omega} := \int_{\partial\Omega} \nabla u \otimes \boldsymbol{n}\ \phi, \tag{1.6}$$

for example.

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 or higher).

With our approach we leave the door open to potential adaptive methods, by simply modifying appropriately 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].

As noted later in Theorem 3.5 (see [Pry10] for a proof), this method can be seen as an extension of the standard FEM. If the problem's coefficient matrix is (piecewise) constant, then

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div}(\boldsymbol{A}\nabla u), \tag{1.7}$$

and the finite element solution generated by this method coincides with the standard conforming finite element solution [Cia78].

This paper focuses mainly on the algorithmic and linear algebraic aspects of the method and is organized as follows. In section 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 section 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 section 4 we summarize extensive numerical experiments on model linear BVPs in nonvariational form and an application to a quasi-linear BVP in nonvariational form.

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

$$\begin{aligned} \mathscr{L}u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{2.1}$$

where the data $f \in \mathrm{L}_2(\Omega)$ is prescribed and $\mathscr{L}$ is a general linear, second order, uniformly elliptic partial differential operator. Let $\boldsymbol{A} \in \mathrm{L}_\infty(\Omega)^{d\times d}$, and for each $\boldsymbol{x} \in \Omega$ let $\boldsymbol{A}(\boldsymbol{x}) \in \mathrm{Sym}^+(\mathbb{R}^{d\times d})$, the space of bounded, symmetric, positive definite $d \times d$ matrices. We then define

$$\begin{array}{rccc} \mathscr{L}: & \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega) & \to & \mathrm{L}_2(\Omega) \\ & u & \mapsto & \mathscr{L}u := \boldsymbol{A}{:}\mathrm{D}^2 u. \end{array} \tag{2.2}$$

We use $\boldsymbol{X}{:}\boldsymbol{Y} := \operatorname{trace}(\boldsymbol{X}^{\intercal}\boldsymbol{Y})$ to denote the Frobenius inner product between two matrices.

2.1. Discretization. Let $\mathscr{T}$ be a conforming, shape regular 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) $\bigcup_{K\in\mathscr{T}} \overline{K} = \overline{\Omega}$.

We use the convention where $h : \Omega \to \mathbb{R}$ denotes the *meshsize function* of $\mathscr{T}$, i.e.,

$$h(\boldsymbol{x}) := \max_{\overline{K}\ni\boldsymbol{x}} h_K. \tag{2.3}$$

We introduce the *finite element spaces*

$$\mathbb{V} := \left\{\Phi \in \mathrm{H}^1(\Omega) :\ \Phi|_K \in \mathbb{P}^p\ \forall\, K \in \mathscr{T}\right\}, \tag{2.4}$$

$$\mathring{\mathbb{V}} := \mathbb{V} \cap \mathrm{H}^1_0(\Omega) = \{\Phi \in \mathbb{V} :\ \Phi|_{\partial\Omega} = 0\}, \tag{2.5}$$

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 $\mathring{N} := \dim \mathring{\mathbb{V}}$ and $N = \mathring{N} + N_\partial := \dim \mathbb{V}$. Let $\mathring{\boldsymbol{\Phi}} = (\mathring{\Phi}_1, \ldots, \mathring{\Phi}_{\mathring{N}})^{\intercal}$ and $\boldsymbol{\Phi} = (\mathring{\Phi}_1, \ldots, \mathring{\Phi}_{\mathring{N}}, \Phi_1, \ldots, \Phi_{N_\partial})^{\intercal}$, where $\{\mathring{\Phi}_1, \ldots, \mathring{\Phi}_{\mathring{N}}\}$ and $\{\mathring{\Phi}_1, \ldots, \mathring{\Phi}_{\mathring{N}}, \Phi_1, \ldots, \Phi_{N_\partial}\}$ form a basis of $\mathring{\mathbb{V}}$, $\mathbb{V}$, respectively.

Testing the model problem (2.1) with $\phi \in \mathrm{H}^1_0(\Omega)$ gives

$$\langle \mathscr{L}u, \phi\rangle = \left\langle \boldsymbol{A}{:}\mathrm{D}^2 u, \phi\right\rangle = \langle f, \phi\rangle. \tag{2.6}$$

2.2. Finite element Hessian. In order to discretize (2.6) with a Galerkin procedure over $\mathbb{V}$, we will introduce an appropriate definition of a "Hessian" of a finite element function. One prerequisite for this "Hessian," which we will call a *finite element Hessian*, is that it extends in some sense the usual distributional definition of Hessian, which for a function $v \in \mathrm{H}^1(\Omega)$ is defined as

$$\left\langle \mathrm{D}^2 v \,|\, \phi\right\rangle = -\langle \nabla v \otimes \nabla\phi\rangle \qquad \forall\, \phi \in \mathrm{C}^\infty_0(\Omega), \tag{2.7}$$

where $\mathrm{C}^\infty_0(\Omega)$ denotes the Schwartz class of functions on Ω:

$$\mathrm{C}^\infty_0(\Omega) := \{\phi \in \mathrm{C}^\infty(\Omega) :\ \operatorname{supp}\phi \text{ compact in } \Omega\}. \tag{2.8}$$

The first step towards a Galerkin method is to specialize (2.7) to test functions $\phi \in \mathbb{V}$, which is not a subset of $\mathrm{C}^\infty_0(\Omega)$. *We stress that using $\phi \in \mathring{\mathbb{V}}$ is not enough because, roughly speaking, we lose too much information on the boundary and the finite element Hessian is not necessarily zero at the boundary.* So, for a given v we would like to extend (or replace) the domain of the functional $\mathrm{D}^2 v$ in (2.7) so as to include some test functions which *are not compactly supported*.

Finding a candidate to "extend" $\mathrm{D}^2 v$ is immediate: Letting $\boldsymbol{n} : \partial\Omega \to \mathbb{R}^d$ be the outward pointing normal of Ω, and taking $v \in \mathrm{C}^2(\Omega) \cap \mathrm{C}^1(\overline{\Omega})$, an integration by parts gives

$$\left\langle \mathrm{D}^2 v, \phi\right\rangle = -\langle \nabla v \otimes \nabla\phi\rangle + \langle \nabla v \otimes \boldsymbol{n}\ \phi\rangle_{\partial\Omega} \qquad \forall\, \phi \in \mathrm{C}^1(\Omega) \cap \mathrm{C}^0(\overline{\Omega}). \tag{2.9}$$

Now if $v \in \mathbb{V}$, i.e., v is a piecewise polynomial on the triangulation $\mathscr{T}$, continuous but not necessarily differentiable, then its gradient, ∇v, is a function in $\mathbb{P}^{p-1}(\mathscr{T})$ (possibly

discontinuous across the edges) and its limit at the boundary, $\nabla v|_{\partial\Omega}$, is well defined $(d-1)$-almost everywhere; hence *the expression on the right-hand side of* (2.9) *is well defined for* $\phi \in \mathbb{V}$*, and it coincides with the right-hand side in* (2.7) *when* ϕ *has zero boundary values.*

DEFINITION 2.1 (finite element Hessian). *Inspired by section* 2.2*, and using Riesz representation in* $\mathbb{V}$*, we define the* finite element Hessian *of* $V \in \mathring{\mathbb{V}}$ *as the unique* $\boldsymbol{H}[V] \in \mathbb{V}$ *such that*

$$\langle \boldsymbol{H}[V], \Phi \rangle := -\langle \nabla V \otimes \nabla \Phi \rangle + \langle \nabla V \otimes \boldsymbol{n}\, \Phi \rangle_{\partial\Omega} \qquad \forall\, \Phi \in \mathbb{V}. \tag{2.10}$$

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

Taking the model problem (2.6), 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

$$\left\langle \boldsymbol{A} : \boldsymbol{H}[U], \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle \qquad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{2.11}$$

THEOREM 2.2 (nonvariational finite element method (NVFEM)). *The* nonvariational finite element solution *for the model problem's discretization* (2.11) *is given as* $U = \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \mathbf{u}$*, where* $\mathbf{u} \in \mathbb{R}^{\mathring{N}}$ *is the solution to the linear system*

$$\mathbf{D}\mathbf{u} := \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} \mathbf{B}^{\alpha,\beta} \mathbf{M}^{-1} \mathbf{C}_{\alpha,\beta} \mathbf{u} = \mathbf{f}. \tag{2.12}$$

The components of (2.12) *are given by*

$$\mathbf{B}^{\alpha,\beta} := \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{A}^{\alpha,\beta} \boldsymbol{\Phi}^{\mathsf{T}} \right\rangle \in \mathbb{R}^{\mathring{N} \times N}, \tag{2.13}$$

$$\mathbf{M} := \langle \boldsymbol{\Phi}, \boldsymbol{\Phi}^{\mathsf{T}} \rangle \in \mathbb{R}^{N \times N}, \tag{2.14}$$

$$\mathbf{C}_{\alpha,\beta} := -\left\langle \partial_\beta \boldsymbol{\Phi}, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle + \left\langle \boldsymbol{\Phi} n_\beta, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle_{\partial\Omega} \in \mathbb{R}^{N \times \mathring{N}}, \tag{2.15}$$

$$\mathbf{f} := \left\langle f, \mathring{\boldsymbol{\Phi}} \right\rangle \in \mathbb{R}^{\mathring{N}}. \tag{2.16}$$

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 (2.11) with $\mathring{\boldsymbol{\Phi}}$,

$$\begin{aligned} \left\langle f, \mathring{\boldsymbol{\Phi}} \right\rangle &= \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} \left\langle \boldsymbol{A}^{\alpha,\beta} \boldsymbol{H}_{\alpha,\beta}[U], \mathring{\boldsymbol{\Phi}} \right\rangle \\ &= \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{A}^{\alpha,\beta} \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{h}_{\alpha,\beta} \right\rangle \\ &= \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{A}^{\alpha,\beta} \boldsymbol{\Phi}^{\mathsf{T}} \right\rangle \mathbf{h}_{\alpha,\beta} \\ &= \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} \mathbf{B}^{\alpha,\beta} \mathbf{h}_{\alpha,\beta}. \end{aligned} \tag{2.17}$$

Utilizing Definition 2.1 for each $\alpha, \beta = 1, \ldots, d$, we can compute $\mathbf{h}_{\alpha,\beta} \in \mathbb{R}^N$, noting $U = \mathring{\boldsymbol{\Phi}}^{\mathsf{T}}\mathbf{u}$,

$$\begin{aligned}\langle \boldsymbol{\Phi}, \boldsymbol{\Phi}^{\mathsf{T}} \rangle \, \mathbf{h}_{\alpha,\beta} &= \langle \boldsymbol{\Phi}, \boldsymbol{H}_{\alpha,\beta}[U] \rangle \\ &= -\langle \partial_\beta \boldsymbol{\Phi}, \partial_\alpha U \rangle + \langle \boldsymbol{\Phi} \boldsymbol{n}_\beta, \partial_\alpha U \rangle_{\partial\Omega} \\ &= \left(-\left\langle \partial_\beta \boldsymbol{\Phi}, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle + \left\langle \boldsymbol{\Phi} \boldsymbol{n}_\beta, \partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle_{\partial\Omega} \right) \mathbf{u}.\end{aligned} \tag{2.18}$$

Using the definitions of $\mathbf{C}_{\alpha,\beta}$ (2.15) and $\mathbf{M}$ (2.14) we see that for each $\alpha, \beta = 1, \ldots, d$

$$\begin{aligned}\mathbf{M}\mathbf{h}_{\alpha,\beta} &= \mathbf{C}_{\alpha,\beta}\mathbf{u}, \\ \mathbf{h}_{\alpha,\beta} &= \mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}\mathbf{u}.\end{aligned} \tag{2.19}$$

Substituting $\mathbf{h}_{\alpha,\beta}$ from (2.19) into (2.17), we obtain the desired result. ∎

Example 2.3 (for $d = 2$). For a general elliptic operator in two dimensions, the formulation (2.12) takes the form

$$\left(\mathbf{B}^{1,1}\mathbf{M}^{-1}\mathbf{C}_{1,1} + \mathbf{B}^{2,2}\mathbf{M}^{-1}\mathbf{C}_{2,2} + \mathbf{B}^{1,2}\mathbf{M}^{-1}\mathbf{C}_{1,2} + \mathbf{B}^{2,1}\mathbf{M}^{-1}\mathbf{C}_{2,1}\right)\mathbf{u} = \mathbf{f}. \tag{2.20}$$

3. Solving the linear system (2.12). Most Galerkin and FEMs feature sparse matrices that allow the use of efficient iterative methods for their solution. In this section, after noting that this sparsity is lost in setting up system (2.12), we show how to recover a solver that takes advantage of the sparsity of these matrices by augmenting the discrete finite element space with a "Hessian" space and using a generalized Schur complement approach.

Remark 3.1 (system (2.12) is difficult to solve). By choosing appropriate basis functions, e.g., standard conforming $\mathbb{P}^p$ elements, the matrices $\mathbf{B}^{\alpha,\beta}$, $\mathbf{M}$, and $\mathbf{C}_{\alpha,\beta}$ are sparse. However, the full system matrix $\mathbf{D} = \sum\sum \mathbf{B}^{\alpha,\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}$ is generally not sparse.

Remark 3.2 (mass lumping). 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 matrices $\mathbf{B}^{\alpha,\beta}$ and $\mathbf{C}_{\alpha,\beta}$). Hence the system (2.12) can be solved using standard iterative methods for sparse matrices. 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 matrices that make up the system.

3.1. A generalized Schur complement. We observe that the matrix $\mathbf{D}$ in the system (2.12) is a sum of Schur complements $\mathbf{B}^{\alpha,\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha,\beta}$. With this in mind we introduce the $(d^2+1)^2$ block matrix

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

LEMMA 3.3 (generalized Schur complement). *Given*

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

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

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 ith row by the ith 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:

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

∎

Remark 3.4 (structure of the block matrix). In fact, this method for the solution of the system $\mathbf{Du} = \mathbf{f}$ is not surprising given that the discretization presented in the proof of Theorem 2.2 is equivalent to the following system:

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

THEOREM 3.5 (equivalence to the standard FEM [Pry10]). *In the case that the problem coefficients in* (2.6) *are piecewise constant, then the problem*

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div}(\boldsymbol{A}\nabla u) \tag{3.10}$$

and the nonvariational finite element solution coincides with that of the standard FEM. That is, $\mathbf{u}$ *solves both*

$$\mathbf{D}\mathbf{u} = \mathbf{f} \tag{3.11}$$

and

$$\mathbf{S}\mathbf{u} = \mathbf{f}, \tag{3.12}$$

where

$$\mathbf{S} = \sum_{\alpha,\beta=1}^{d} \left\langle \partial_\beta \mathring{\boldsymbol{\Phi}}, a^{\alpha,\beta}\partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle \tag{3.13}$$

is the standard finite element stiffness matrix.

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

$$\begin{aligned} \mathscr{L}u &= f \text{ in } \Omega, \\ u &= g \text{ on } \partial\Omega, \end{aligned} \tag{3.14}$$

it is not immediate how to enforce the boundary conditions. If we were solving the full system $\mathbf{D}\mathbf{u} = \mathbf{f}$, 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,

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{E}_\partial & \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{v}_\partial \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_\partial \\ \mathbf{b} \end{bmatrix}, \tag{3.15}$$

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:

$$\mathbf{E}_\partial = \begin{bmatrix} \mathbf{M} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{C}^\partial_{1,1} \\ \mathbf{0} & \mathbf{M} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{C}^\partial_{1,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M} & \mathbf{0} & -\mathbf{C}^\partial_{d,d-1} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{M} & -\mathbf{C}^\partial_{d,d} \\ \mathbf{B}^{1,1} & \mathbf{B}^{1,2} & \ldots & \mathbf{B}^{d,d-1} & \mathbf{B}^{d,d} & \mathbf{0} \end{bmatrix}, \tag{3.16}$$

$$\mathbf{v}_\partial = \left[\mathbf{h}^\partial_{1,1}, \mathbf{h}^\partial_{1,2}, \ldots, \mathbf{h}^\partial_{d,d-1}, \mathbf{h}^\partial_{d,d}, \mathbf{u}^\partial\right]^{\mathsf{T}}, \tag{3.17}$$

$$\mathbf{b}_\partial = [\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, \mathbf{0}, \mathbf{g}]^{\mathsf{T}}. \tag{3.18}$$

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

$$\mathbf{C}^\partial_{\alpha,\beta} = -\langle \partial_\beta \boldsymbol{\Phi}, \partial_\alpha \boldsymbol{\Phi}_\partial{}^{\mathsf{T}} \rangle + \langle \boldsymbol{\Phi}\boldsymbol{n}_\beta, \partial_\alpha \boldsymbol{\Phi}_\partial{}^{\mathsf{T}} \rangle_{\partial\Omega} \in \mathbb{R}^{N \times N_\partial}, \tag{3.19}$$

$$\mathbf{g}_j = g(x_j)\Phi_j \in \mathbb{R}^{N_\partial}, \tag{3.20}$$

TABLE 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 $\kappa(\mathbf{E}) \approx Ch^{-2}\log h$. *The problem data are taken from the test given in section* 4.1.

$\dim \mathbb{V}$	h	$\kappa(\mathbf{E})$	$h^2\kappa(\mathbf{E})$
16	0.4714	4.904×10^1	10.898
64	0.2020	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	40.430

where x_j is the Lagrange node associated with Φ_j.

The block matrix (3.15) can then be trivially solved:

$$\mathbf{E}\mathbf{v} = \mathbf{b} - \mathbf{E}_\partial \mathbf{b}_\partial. \tag{3.21}$$

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

Remark 3.8 (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 $\mathbf{N}$:

$$\kappa(\mathbf{N}) = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{3.22}$$

Numerically we observe for $p = 1$ the condition number of the block matrix $\kappa(\mathbf{E}) \leq Ch^{-2}\log h$ (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 four 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 $\mathscr{L}$ introduced in (2.2) with diffusion matrix

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

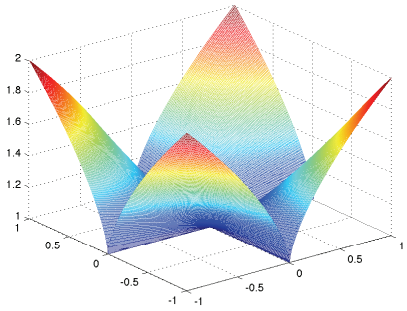
varying the coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x})$.

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 not differentiable on Ω, and in this case we take the diffusion matrix $\boldsymbol{A}$ in (4.1) with

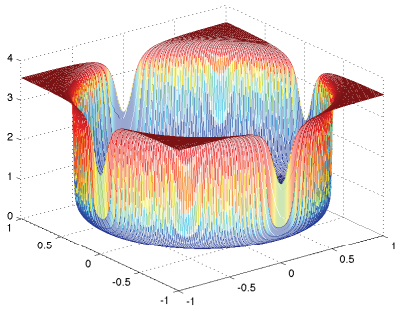
$$\begin{aligned} a(\boldsymbol{x}) &= (x_1^2 x_2^2)^{1/3} + 1, \\ b(\boldsymbol{x}) &= 0. \end{aligned} \tag{4.2}$$

A visualization of the operator (4.2) is given in Figure 1. We choose our problem data f such that the exact solution to the problem is given by

$$u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|^2). \tag{4.3}$$

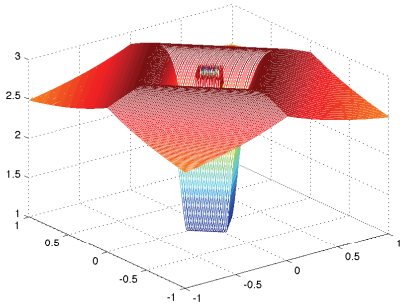


(a) The function $(x_1^2 x_2^2)^{1/3} + 1$ over Ω. Note that the derivatives are singular at $x_1 = 0$ and $x_2 = 0$.

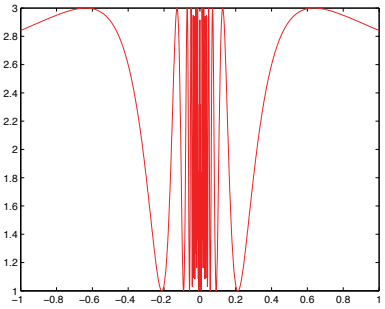


(b) The function $\arctan\big(5000(|x|^2 - 1)\big)$ over Ω. Note that the derivatives are very large on the unit circle.

FIG. 1. *A visualization of the coefficient of the operators* (4.2) *(on the left) and* (4.4) *(on the right).*



(a) The function $\sin\big(\frac{1}{|x_1|+|x_2|+10^{-15}}\big)$ over Ω. Note that the function oscillates heavily near $\mathbf{0}$.



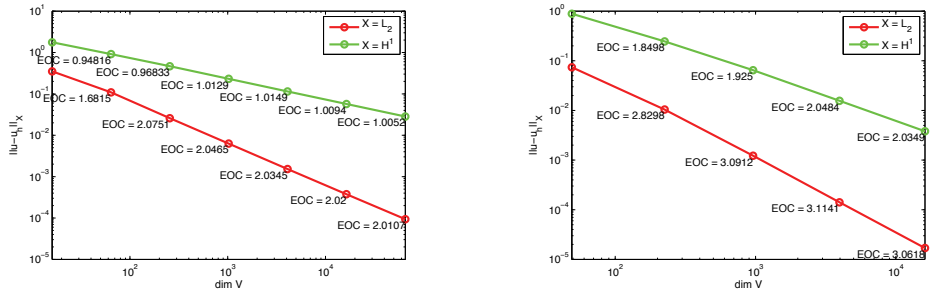
(b) A cross section through the first coordinate axis. Take note of the wild oscillations around the origin.

FIG. 2. *A visualization of the coefficient of the operator in Example* 4.3 *and a cross section through the coordinate axis.*

We discretize the problem given by (4.2) under the algorithm set out in section 2.1. In Figure 3 we test the method for $p = 1, 2$ and numerically show that $\|u - U\| = \mathrm{O}(h^{p+1})$ and $|u - U|_1 = \mathrm{O}(h^p)$.

4.2. Test problem with convection dominated operator. The second test problem is designed for demonstrating the ability of NVFEM to avoid oscillations introduced into the standard finite element when rewriting the operator in divergence form. Take the matrix $\boldsymbol{A}$ in (4.1) with

$$
\begin{aligned}
a(\boldsymbol{x}) &= \arctan\Big(K(|\boldsymbol{x}|^2 - 1)\Big) + 2, \\
b(\boldsymbol{x}) &= 0
\end{aligned}
\tag{4.4}
$$



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

FIG. 3. *Test problem given in section* 4.1. *We plot the* log *of the error together with its estimated order of convergence. We study both* $\mathrm{L}_2(\Omega)$ *and* $\mathrm{H}^1(\Omega)$ *norms of the error for the NVFEM applied to a nondivergence form operator whose coefficients are given by* (4.2). *Here* f *is chosen appropriately such that* $u(\boldsymbol{x}) = \exp\left(-10\left|\boldsymbol{x}\right|^2\right)$. *The convergence rates are optimal, in the sense that 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* $|u - U|_1 = \mathrm{O}(h^2)$.

and $K \in \mathbb{R}^+$. We rewrite the problem in divergence form (1.4) and take notice that the derivatives

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

can be made arbitrarily large on the unit circle by choosing K appropriately (see Figure 1).

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

$$u(\boldsymbol{x}) = \sin\left(\pi x_1\right)\sin\left(\pi x_2\right). \tag{4.6}$$

We then construct the standard FEM around (1.4), that is, find $U \in \mathring{\mathbb{V}}$ such that

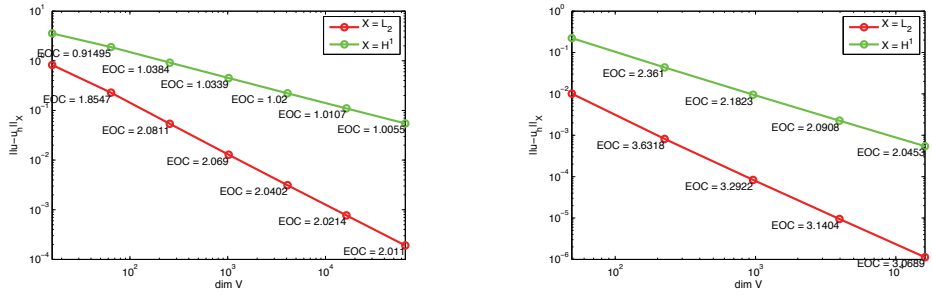
$$\left\langle \boldsymbol{A}\nabla U, \nabla\mathring{\Phi}\right\rangle - \left\langle \mathrm{div}\left(\boldsymbol{A}\right)\nabla U, \mathring{\Phi}\right\rangle = \left\langle f, \mathring{\Phi}\right\rangle \qquad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.7}$$

If K is chosen small enough, the standard FEM converges optimally. If we increase the value of K, oscillations become apparent in the finite element solution along the unit circle and the FEM no longer converges. Fixing $K = 5000$, in Figure 5 we show the oscillations arising from this method compared to discretizing using the NVFEM. In Figure 4 we test the method for $p = 1, 2$ and numerically show that $\|u - U\| = \mathrm{O}(h^{p+1})$ and $|u - U|_1 = \mathrm{O}(h^p)$.

4.3. Test problem with a singular solution. In this test we choose the matrix $\boldsymbol{A}$ in (4.1) with

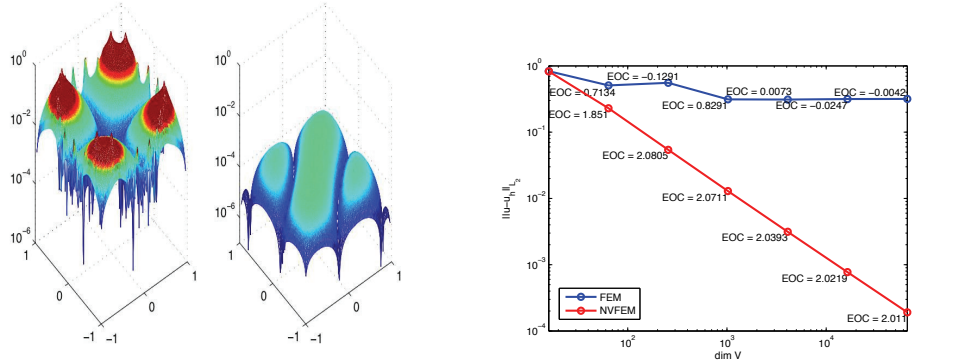
$$\begin{aligned} a(\boldsymbol{x}) &= \sin\left(\frac{1}{|x_1| + |x_2| + 10^{-15}}\right) + 2, \\ b(\boldsymbol{x}) &= 0. \end{aligned} \tag{4.8}$$

Notice that the operator oscillates heavily near $\mathbf{0}$. Figure 2 shows a surface plot of the operator (4.8) and a cross section through $x_1 = 0$ illustrating the oscillations near the origin.



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

FIG. 4. *Test problem given in section* 4.2. *We plot the* log *of the error together with its estimated order of convergence. We study both* $\mathrm{L}_2(\Omega)$ *and* $\mathrm{H}^1(\Omega)$ *norms of the error for the NVFEM applied to a nondivergence form operator whose coefficients are given by* (4.4) *with* $K = 5000$. *Here* f *is chosen appropriately such that* $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$. *The convergence rates are optimal as in Figure* 3.



(a) On the left we present $\|u - \tilde{U}\|_{\mathrm{L}_\infty(K)}$ plotted on a logarithmic scale as a function over Ω. On the right we show $\|u - U\|_{\mathrm{L}_\infty(K)}$ plotted on a logarithmic scale as a function over Ω both with 16384 DOFs ($h = 1/32$).

(b) We plot the log of the error together with its estimated order of convergence. We study the $\mathrm{L}_2(\Omega)$ error of the standard FEM and the NVFEM.

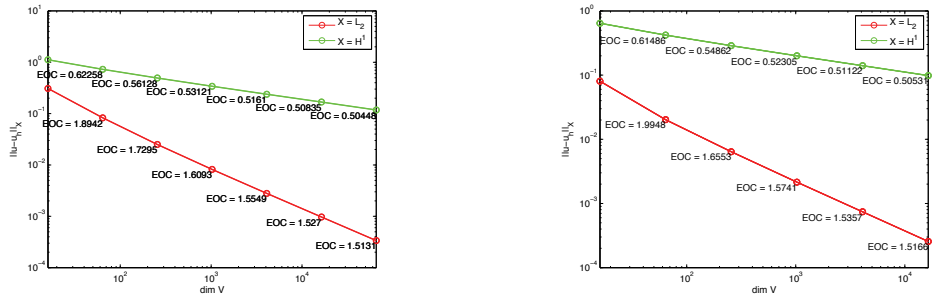
FIG. 5. *Test problem given in section* 4.2. *We compare the convergence of the two methods. We fix* $p = 1$ *and* $k = 5000$. *Notice the oscillations apparent on the unit circle when using the standard FEM. The standard FEM fails to converge as the mesh is refined. The NVFEM converges optimally, in the sense that* $\|u - U\| = \mathrm{O}(h^2)$.

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

$$u(\boldsymbol{x}) = \left(2 - x_1^2 - x_2^2\right)^{1/2}. \tag{4.9}$$

The solution $u \in \mathrm{H}^1(\Omega)$ but $u \notin \mathrm{H}^2(\Omega)$. The singularities occur in the gradient on the corners of Ω, and the convergence rates are slowed, as can be seen in Figure 6; $\|u - U\| = \mathrm{O}(h^{1.5})$ and $|u - U| = \mathrm{O}(h^{0.5})$ regardless of p.

4.4. Test problem choosing a solution with nonsymmetric Hessian. In this test we choose the operator such that $b(\boldsymbol{x})$ is nonzero. To maintain ellipticity in this problem we must choose $a(\boldsymbol{x})$ such that the trace of $\boldsymbol{A}$ dominates its determinant.



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

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

FIG. 6. *Test problem given in section* 4.3. $\mathrm{L}_2(\Omega)$ *and* $\mathrm{H}^1(\Omega)$ *errors and convergence rates for the NVFEM on an operator whose coefficients are given by* (4.8)*, choosing* f *appropriately such that* $u(\boldsymbol{x}) = (2 - x_1^2 - x_2^2)^{1/2}$*. The convergence rates are suboptimal due to the solution's gradient singularity at the corners of* $\Omega = (-1, 1)^2$*, that is, for both* $\mathbb{P}^1$ *(on the left) and* $\mathbb{P}^2$ *elements (on the right)* $\|u - U\| = \mathrm{O}(h^{1.5})$ *and* $|u - U|_1 = \mathrm{O}(h^{0.5})$.

We choose

$$\begin{aligned} a(\boldsymbol{x}) &= 2, \\ b(\boldsymbol{x}) &= (x_1^2 x_2^2)^{1/3}. \end{aligned} \tag{4.10}$$

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

$$u(\boldsymbol{x}) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2}, & \boldsymbol{x} \neq \mathbf{0}, \\ 0, & \boldsymbol{x} = \mathbf{0}. \end{cases} \tag{4.11}$$

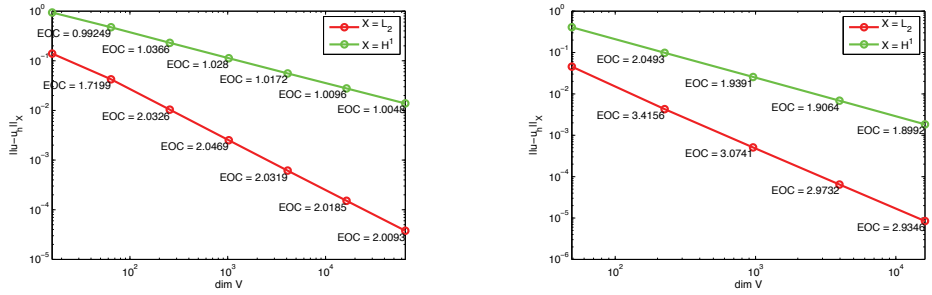
This function has a nonsymmetric Hessian at the point $\mathbf{0}$. (Schwarz's theorem, which allows the interchange of mixed derivatives from calculus, is not applicable for lack of continuity in the second derivatives.) The nontrivial Dirichlet boundary is dealt with using Remark 3.6. In Figure 7 we test the method for $p = 1, 2$ and numerically show that $\|u - U\| = \mathrm{O}(h^{p+1})$ and $|u - U|_1 = \mathrm{O}(h^p)$.

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

$$\operatorname{div}\left(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}}\right) = \frac{f}{\sqrt{1 + |\nabla u|^2}}, \tag{4.12}$$

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

$$\operatorname{div}\left(\frac{\nabla u^n}{\sqrt{1 + |\nabla u^{n-1}|^2}}\right) = \frac{f}{\sqrt{1 + |\nabla u^{n-1}|^2}}. \tag{4.13}$$



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

FIG. 7. *Test problem given in section* 4.4. *We plot the* log *of the error together with its estimated order of convergence. We study both* $\mathrm{L}_2(\Omega)$ *and* $\mathrm{H}^1(\Omega)$ *norms of the error for whose coefficients are given by the NVFEM applied to a nondivergence form operator* (4.10). *Here* f *is chosen 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 as in Figure* 3.

Applying a standard finite element discretization of (4.13) yields the following: Given $U^0 \in \mathring{\mathbb{V}}$, for each $n \in \mathbb{N}$ find $U^n \in \mathring{\mathbb{V}}$ such that

$$\left\langle \frac{\nabla U^n}{\sqrt{1 + |\nabla U^{n-1}|^2}}, \nabla \mathring{\Phi} \right\rangle = \left\langle \frac{f}{\sqrt{1 + |\nabla U^{n-1}|^2}}, \mathring{\Phi} \right\rangle \qquad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.14}$$

In fact we can work on this problem combining the two nonlinear terms. To do so we must first rewrite (4.12) in the form $A(u, \nabla u){:}\mathrm{D}^2 u = f$:

$$\begin{aligned} f &= \sqrt{1 + |\nabla u|^2}\, \mathrm{div} \left(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}} \right) \\ &= \Delta u + \frac{\mathrm{D}u \mathrm{D}^2 u \nabla u}{1 + |\nabla u|^2} \\ &= \left(\boldsymbol{I} + \frac{\nabla u \mathrm{D}u}{1 + |\nabla u|^2} \right) {:}\mathrm{D}^2 u. \end{aligned} \tag{4.15}$$

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

$$\left(\boldsymbol{I} + \frac{\nabla u^{n-1} \mathrm{D}u^{n-1}}{1 + |\nabla u^{n-1}|^2} \right) {:}\mathrm{D}^2 u^n = f. \tag{4.16}$$

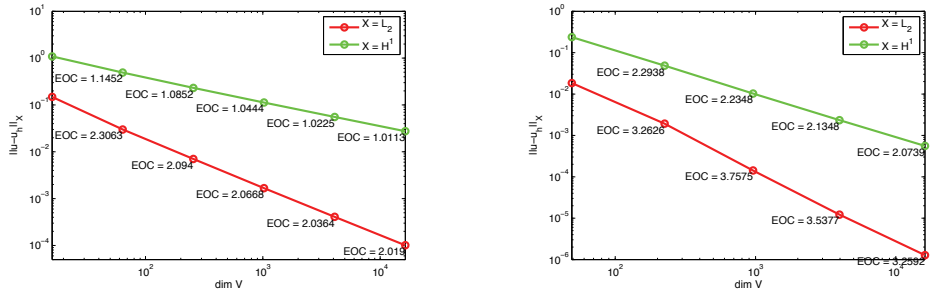
Discretizing the problem is then similar to that set out in section 2.1. The component matrices $\mathbf{M}$ and $\mathbf{C}_{\alpha,\beta}$ are problem independent, and $\mathbf{B}^{\alpha,\beta}$ are defined as

$$\mathbf{B}^{\alpha,\beta} = \begin{cases} \left\langle \mathring{\boldsymbol{\Phi}}, 1 + \frac{\partial_\alpha U^{n-1} \partial_\beta U^{n-1}}{1 + |\nabla U^{n-1}|^2} \boldsymbol{\Phi} \right\rangle & \text{for } \alpha = \beta, \\ \left\langle \mathring{\boldsymbol{\Phi}}, \frac{\partial_\alpha U^{n-1} \partial_\beta U^{n-1}}{1 + |\nabla U^{n-1}|^2} \boldsymbol{\Phi} \right\rangle & \text{for } \alpha \neq \beta. \end{cases} \tag{4.17}$$

TABLE 2

Test problem given in section 4.5. *Comparison of the fixed point linearization in variational form* (4.13) *and in nonvariational form* (4.16). *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* (4.12) *using a standard FEM and using the NVFEM. Denoting* U_i *and* $\tilde{U}_i$ *to be the nonvariational, respectively, standard finite element solution, we run both linearizations until a tolerance* $\|U_{n+1} - U_n\|$ *(resp.,* $\|\tilde{U}_{n+1} - \tilde{U}_n\|$*)* $\le h^2$ *is achieved. We compute both the stagnation point—defined as the iteration at which the prescribed tolerance is achieved—and the total CPU time. Notice the significant savings in the number of iterations required to reach the stagnation point using the NVFEM over the standard FEM. On the other hand, 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
NVFEM	Stag. point	4	6	7	8	10	12
	CPU time	0.72	3.40	16.49	97.93	838.8	5256.84



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

FIG. 8. *Test problem given in section* 4.5. *We plot the* log *of the error together with its estimated order of convergence. We study both* $\mathrm{L}_2(\Omega)$ *and* $\mathrm{H}^1(\Omega)$ *norms of the error for the NVFEM applied to* (4.12)*, a quasi-linear PDE under a fixed point linearization* (4.16). *Here* f *is chosen appropriately such that* $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$. *The convergence rates are optimal as in Figure* 3.

Table 2 compares the two linearizations (4.13) and (4.16). In Figure 8 we test the method for $p = 1, 2$ and numerically show that $\left\|u - U^N\right\| = \mathrm{O}(h^{p+1})$ and $\left|u - U^N\right|_1 = \mathrm{O}(h^p)$, where N is the final iteration of the discretized problem.

5. Conclusions. We have proposed a novel method, called NVFEM, for the approximation of strong solutions to general second order elliptic BVPs in nonvariational form. This method, based on a concept of "recovered Hessian," consists in applying a Galerkin procedure on the strong form of the elliptic PDE, rather than the weak form, as is usually done.

The NVFEM, after using a generalized Schur complement, results in a linear system larger than (but with similar sparsity to) the standard FEM. This apparent disadvantage is compensated for by its ability to include problems in nondivergence form where the standard conforming FEM underperforms or simply fails to converge due to instabilities. The NVFEM uses conforming elements and needs no stabilization.

We have demonstrated the above facts by implementing NVFEM with MATLAB (code freely available on request) and testing the code as follows:

(a) NVFEM was run on a set of elliptic problems given in strong form and whose variational version involves singular coefficients. While the conforming FEM cannot be used (due to singularities of coefficients), the NVFEM (whose main feature avoids the variational version) performs with optimal experimental orders of convergence.
(b) NVFEM and conforming FEM were both run on problems whose variational form has dominant convection terms. This prevents a successful use of conforming elements, unless one uses stabilization techniques or, as we did, the NVFEM *without the need of any stabilization mechanism*.
(c) Finally we outlined the potential of NVFEM for nonlinear problems. Although this is the objective of a companion paper of ours [LP10], here we have implemented a fixed point linearization of the nonparametric prescribed mean curvature problem without writing it in variational form. Computer tests show that the fixed point NVFEM is robust.

REFERENCES

[AM09] N. E. Aguilera and P. Morin, *On convex functions and the finite element method*, SIAM J. Numer. Anal., 47 (2009), pp. 3139–3157.

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

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

[Cia78] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.

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

[CSX07] L. Chen, P. Sun, and J. Xu, *Optimal anisotropic meshes for minimizing interpolation errors in L^p-norm*, Math. Comp., 76 (2007), pp. 179–204.

[GT01] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Classics Math., Springer-Verlag, Berlin, 2001; reprint of the 1998 edition.

[KT92] H.-J. Kuo and N. S. Trudinger, *Discrete methods for fully nonlinear elliptic equations*, SIAM J. Numer. Anal., 29 (1992), pp. 123–135.

[LP10] O. Lakkis and T. Pryer, *A Finite Element Method for Fully Nonlinear Elliptic Equations*, Preprint, 2011; available online from http://arxiv.org/abs/1103.2970.

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

[Obe08] A. 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 (2008), pp. 221–238.

[Pry10] T. Pryer, *Recovery Methods for Evolution and Nonlinear Problems*, DPhil Thesis, University of Sussex, 2010.

[Tho06] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer Ser. Comput. Math. 25, Springer-Verlag, Berlin, 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 (2007), pp. 987–1007.