**Stable Finite Elements for
Problems With Wild Coefficients***

Stephen A. Vavasis**

TR 93-1364
June 1993

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*This work supported by an NSF Presidential Young Investigator grant, with matching funds received from AT&T and Xerox Corp. Part of this work was done while the author was visiting Sandia National Laboratories, supported by the U.S. Department of Energy under contract DE-AC04-76DP00789.

**Department of Computer Science, Upson Hall, Cornell University, Ithaca, NY 14853. vavasis@cs.cornell.edu

Stable finite elements for problems with wild coefficients*

Stephen A. Vavasis[†]

June 27, 1993

Abstract

We consider solving an elliptic boundary value problem in the case that the coefficients vary by many orders of magnitude over the domain. A linear finite element method is used. It is shown that the standard method for solving the resulting linear equations in finite-precision arithmetic can give an arbitrarily inaccurate answer because of ill-conditioning in the stiffness matrix. A new method for solving the linear equations is proposed. This method is based on a “mixed formulation” and gives a numerically accurate answer independent of the variation in the coefficients. The numerical error in the solution of the linear system for the new method is shown to depend on the aspect ratio of the triangulation.

1 A boundary value problem

The problem we consider is

$$\begin{aligned} \nabla \cdot (c \nabla u) &= 0 && \text{on } \Omega, \\ u &= g && \text{on } \partial\Omega. \end{aligned} \tag{1}$$

*This work supported by an NSF Presidential Young Investigator grant, with matching funds received from AT&T and Xerox Corp. Part of this work was done while the author was visiting Sandia National Laboratories, supported by the U.S. Department of Energy under contract DE-AC04-76DP00789.

[†]Department of Computer Science, Upson Hall, Cornell University, Ithaca, NY 14853.

Here, Ω is a polygonal subset (i.e., a finite union of simplices) of $\mathbb{R}^d$ that is topologically equivalent to a ball, but not necessarily convex. The dimension d of the problem is commonly 2 or 3 in applications, but our results hold in arbitrary dimensions. Scalar coefficient field c is given, and is a function from Ω to $\mathbb{R}$. This field must be positive at every point of the domain. Function u is the unknown scalar field on Ω that is to be computed. Finally, the Dirichlet boundary data g is also given; this is a scalar field defined on the boundary.

Solving this equation is a classical problem in mathematical physics and arises in many applications. For instance, u in (1) can describe the equilibrium temperature distribution a body Ω whose boundary is held at a certain constant temperature distribution g . In this thermodynamic application, c represents the thermal conductivity of the body, and $c\nabla u$ represents the heat flow at any point.

In this paper we focus on the case that c varies by many orders of magnitude over the domain, and the values of c do not follow any particular pattern. We say that such a c is “wild.” This problem arises in physical applications in which the underlying domain is composed of materials with significantly varying properties, and the arrangement of the materials has no special geometry.

Regardless of how much c varies, (1) satisfies a “maximum principle.” This principle states that the maximum and minimum value of u must occur on the boundary. See Protter and Weinberger [16] for more information about maximum principles.

Since (1) is a linear problem, the maximum principle applies also to perturbations. In particular, suppose g is perturbed by small amount to yield new boundary data $\hat{g}$. Let $\hat{u}$ be the solution to (1) for the perturbed data. Then, by linearity, $u - \hat{u}$ is the solution to (1) for boundary data $g - \hat{g}$, and thus

$$\|u - \hat{u}\|_{L_\infty(\Omega)} \leq \|g - \hat{g}\|_{L_\infty(\partial\Omega)}.$$

The observation in the last paragraph tells us that (1) is a *well-conditioned* problem mathematically, in the sense that the solution is insensitive to perturbations in the data, regardless of c . Nevertheless, the standard numerical method for obtaining u is ill-conditioned. Specifically, the condition number of the *assembled stiffness matrix* tends to ∞ as the variation in c increases, as we argue in Section 2. This situation of a mathematically well-conditioned

continuous problem whose numerical method is ill-conditioned is anomalous, and suggests that the wrong numerical method is being used. In Section 3 to 5 we propose a new method, the NSHI method, for solving the system of linear equations that has a numerical stability bound independent of c . This new method does not form the stiffness matrix, but instead approaches it implicitly via a *mixed formulation*.

It is important to stress that, in exact arithmetic, our proposed NSHI method is equivalent to the standard finite element method. Both compute the same piecewise linear approximation to u . The difference between the methods is evident only when finite-precision floating point arithmetic is introduced, in which case the ill-conditioning of the standard method magnifies roundoff errors to give the wrong answer.

The stability bounds for the new method depend on parameters $\chi_A, \bar{\chi}_A$ defined by Stewart [18], and applied to a certain “reduced geometry matrix” A that we introduce in Section 3. In Section 6 we derive bounds on these parameters in terms of the geometry of the problem; specifically, we show that the bounds depend on the *aspect ratio* of the simplices in the finite element triangulation. Finally, in Section 8 we describe our computational experiments for the case $d = 2$ comparing the standard method to the NSHI method.

2 The finite element method

In this paper we use a piecewise linear finite element method, which we briefly describe in this section. See Strang and Fix [21] for an overview of the finite element method. First, one forms a triangulation of Ω , that is, a partitioning of Ω into simplices. Let n_0 denote the number of nodes in the triangulation, of which n are interior and $n_0 - n$ lie on the boundary. Let m denote the number of simplices. Let $\mathbf{v}_1, \dots, \mathbf{v}_{n_0}$ denote the nodes of the triangulation, with the first n entries being the interior nodes. Let $T_1, \dots, T_m$ denote the simplices.

The next step is to make the assumption that the solution u to (1) is piecewise linear and continuous, with the pieces of linearity being the simplices. This assumption means that once the values of u at the n interior nodes are known, then all of u is known because the values at other points of Ω may be found with linear interpolation across simplices. (Note that at

the $n_0 - n$ boundary nodes, u is already known from the boundary data.) Let $\mathbf{y}$ be the n -vector $(y_1, \dots, y_n)^T$ encoding the nodal values of u ; it is this n -vector that we must compute.

To obtain a system of equations for $\mathbf{y}$, we introduce basis functions $\phi_1, \dots, \phi_n$, which are functions from Ω to $\mathbb{R}$. These functions, commonly known as “hat functions,” are piecewise linear and continuous, and have the property that $\phi_i(v_i) = 1$ and $\phi_i(v_j) = 0$ for $j \neq i$. This uniquely defines each ϕ_i on all of Ω by linear interpolation. These functions depend only on the triangulation, not on u or g . It is straightforward to prove that $\phi_1, \dots, \phi_n$ form a basis for all functions that are piecewise linear on Ω with respect to triangulation $T_1, \dots, T_m$ and that have boundary value identically zero.

Next, let ϕ_0 be a piecewise linear function that is zero at all interior nodes and agrees with the boundary values at the boundary nodes; again this uniquely defines ϕ_0 . The consequence of these definitions is that u may be written as

$$u = \phi_0 + \sum_{i=1}^n y_i \phi_i \quad (2)$$

where the y_i ’s are the unknown interior nodal values of u .

Next we pass to the so-called “weak form” of (1); this form is obtained by multiplying both sides of the differential equation by ϕ_i for some i , integrating over Ω , and applying Green’s theorem. The end result is that one obtains the equations

$$\int_{\Omega} c \langle \nabla \phi_i, \nabla u \rangle = 0 \quad (3)$$

for all $i = 1, \dots, n$. Here, $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^d$.

Up to now, what we have described is common to both the standard method and the NSHI method. The standard method proceeds from (3) as follows. Substituting (2) into (3) yields n equations

$$\sum_{j=1}^n y_j \int_{\Omega} c \langle \nabla \phi_i, \nabla \phi_j \rangle = - \int_{\Omega} c \langle \nabla \phi_i, \nabla \phi_0 \rangle \quad (4)$$

for $i = 1, \dots, n$.

These equations are linear in $y_1, \dots, y_n$; the coefficients on the left-hand side depend on c and on the triangulation. The quantity on the right-hand side depends on c , on the triangulation, and on the boundary data g .

Based on (4), we form an $n \times n$ linear system

$$K\mathbf{y} = -\mathbf{f}. \quad (5)$$

The (i, j) entry of K is taken from (4):

$$K_{ij} = \int_{\Omega} c \langle \nabla \phi_i, \nabla \phi_j \rangle. \quad (6)$$

This matrix K , known as the *assembled stiffness matrix*, is clearly seen to be symmetric. It is also easily seen to be sparse, since the supports of ϕ_i and ϕ_j will be disjoint or have a measure-zero intersection unless i and j are neighboring nodes in the triangulation. Finally, it can be proved that K is positive definite (this will follow from the discussion below).

Once the linear system is assembled, it is solved, for instance, with sparse Cholesky factorization as described in by George and Liu [11]. We will call this method CFSM (“Cholesky factorization of the stiffness matrix”) and regard it as the standard approach.

The difficulty with the CFSM method in the case of wildly varying coefficients is that K may be arbitrarily ill-conditioned. The ratio between diagonal entries K_{ii} and K_{jj} , for instance, depends on the ratio of c in the support of ϕ_i versus c in the support of ϕ_j . The condition number of a symmetric positive definite matrix is bounded below by the maximum ratio of any two diagonal entries. Thus, it follows that for a fixed triangulation, the condition number of K will grow like $\max(c)/\min(c)$.

The effect of this ill-conditioning is that if $K\mathbf{y} = -\mathbf{f}$ is solved using floating point arithmetic, one can expect a relative error on the order of $\epsilon \max(c)/\min(c)$, where ϵ denotes machine epsilon. Thus, in IEEE double-precision arithmetic, we would expect the CFSM method to collapse when $\max(c)/\min(c)$ is on the order of 10^{16} . Indeed, this is what we have observed in our experiments described in Section 8.

3 A mixed formulation

The first step of the NSHI method is to introduce new variables $p_{k,\mu}$, for $k = 1, \dots, m$ and $\mu = 1, \dots, d$. The vector $\mathbf{p}_{k,\bullet}$ stands for the gradient of u on simplex T_k , multiplied by the constant factor

$$\left[\int_{T_k} c \right] / s_k \quad (7)$$

Here, s_k denotes the longest side length of simplex T_k . Since u is assumed to be piecewise linear, the gradient of u on T_k is constant and therefore can be represented by a d -vector. Let us denote the md vector of scaled gradients as $\mathbf{p}$.

Notation: For the remainder of this paper, we adopt the convention that scalars are denoted with plain italic type, vectors of size d or $d + 1$ such as $\mathbf{v}_1, \dots, \mathbf{v}_{n_0}$ are denoted with bold italic type, and “long” vectors such as $\mathbf{y}$ and $\mathbf{p}$ are denoted with bold Roman type.

The side-length factor s_k that we have introduced is largely irrelevant for the developments in this section; it plays an important role in Section 6.

Note that the gradients of u can also be determined from the nodal values by subtraction and interpolation, since u is linear on each triangle. This observation implies that there is a system of equations relating the gradients $\mathbf{p}$ and the nodal values $\mathbf{y}$. Let us write this system as

$$D\mathbf{p} = A\mathbf{y} + \mathbf{b}, \quad (8)$$

where we now describe the entries of D , A and $\mathbf{b}$ in detail.

On both sides of this equation we want to have the md -vector of gradients of u , multiplied by the scalar factors s_k . First, we explain D . Since we scaled the gradients by constant factors to obtain $\mathbf{p}$, we have to unscale them with D . Thus, D is a diagonal scaling matrix with positive diagonal entries. It has m blocks of d consecutive equal diagonal entries, and the diagonal entry occurring in the k th block is equal to

$$\left[\int_{T_k} c \right]^{-1} s_k^2. \quad (9)$$

Now we explain why $A\mathbf{y} + \mathbf{b}$ represents the scaled gradients on the right-hand side. Let $M_1, \dots, M_m$ be matrices each of size $d \times (d + 1)$ such that M_k performs the following transformation. Given values $x_1, \dots, x_{d+1}$ of a linear function at the $d + 1$ vertices of T_k , $M_k \mathbf{x}$ should return the gradient of that linear function. This uniquely defines M_k ; note that it has a 1-dimensional nullspace spanned by the vector $[1, 1, \dots, 1]^T$.

An actual recipe for M_k may be described as follows. Pick a vertex of T_k arbitrarily; let this vertex be called the “origin,” and let it be numbered last.

Let M° be the $d \times (d + 1)$ matrix

$$M^\circ = \begin{pmatrix} 1 & & & -1 \\ & 1 & & -1 \\ & & \ddots & \vdots \\ & & & 1 & -1 \end{pmatrix}. \quad (10)$$

Then $M^\circ \mathbf{x}$ maps $\mathbf{x}$, a $(d + 1)$ -vector of nodal values, to the vector of nodal values minus the origin value.

Notation: For the remainder of the paper we use the superscript ‘o’ to indicate a combinatorial object, that is, a matrix or vector all of whose entries are 0, 1, or -1 .

Next, let J_k be the the matrix that transforms coordinates from the “standard” d -simplex

$$\{(x_1, \dots, x_d)^T \in \mathbb{R}^d : x_i \geq 0 \text{ for } i = 1, \dots, d \text{ and } x_1 + \dots + x_d \leq 1\}$$

to T_k ; that is,

$$J_k = \begin{pmatrix} (\mathbf{v}^1 - \mathbf{v}^{d+1})^T \\ \vdots \\ (\mathbf{v}^d - \mathbf{v}^{d+1})^T \end{pmatrix}.$$

where $\mathbf{v}^1, \dots, \mathbf{v}^{d+1}$ are the vertices of T_k with the origin numbered last. Then one can see that

$$M_k = J_k^{-1} M^\circ.$$

The choice of origin affects J_k , but M_k itself is independent of this choice (except for the ordering of columns).

Next we form a matrix A_0 of size $md \times n_0$ as follows. For $k = 1, \dots, m$, we construct the block of rows numbered $(k - 1)d + 1$ to kd based on simplex k . The entries of this block of rows are the entries of $s_k M_k$, with the columns scattered according to the global numbering of the nodes.

Now, finally, we obtain A by deleting from A_0 the columns corresponding to boundary nodes. We call A_0 the *extended geometry matrix*, and we call A the *reduced geometry matrix*.

Next, we explain where $\mathbf{b}$ in (8) comes from. This comes from the boundary data; let $\mathbf{y}_0$ be the nodal values of the boundary data. Then $\mathbf{b}$ is equal to the columns of A_0 corresponding to boundary nodes (i.e., the $n_0 - n$ columns

that were deleted to form A) multiplied by $\mathbf{y}_0$. Thus, $A\mathbf{y} + \mathbf{b}$ is the md -vector of gradients scaled by s_k .

We conclude that (8) is a system of equations that must hold relating the gradients and nodal values. Note that A depends only on the geometry; the dependence on c has been relegated to D . Vector $\mathbf{b}$ depends on the boundary data and on the geometry (but not c).

Note also that A is rank n . This follows because, assuming $\mathbf{b} = \mathbf{0}$ (zero boundary values), if any interior nodal value is nonzero, then the gradient vector on some simplex must be nonzero. This is the same as saying that whenever $\mathbf{y}$ is nonzero, so is $A\mathbf{y}$.

Now we formulate a second group of linear equations to enforce (3). We can rewrite (3) as:

$$\sum_{k=1}^m \int_{T_k} c \langle \nabla \phi_i, \nabla u \rangle = 0.$$

Because $\mathbf{p}$ is defined as a scaled version of ∇u , we can write this as:

$$\sum_{k=1}^m s_k \left[\int_{T_k} c \right]^{-1} \cdot \int_{T_k} c \langle \nabla \phi_i, \mathbf{p}_{k,\bullet} \rangle = 0.$$

Note that $\langle \nabla \phi_i, \mathbf{p}_{k,\bullet} \rangle$ is a constant on T_k so it may be pulled out of the integral (write this the constant value of $\nabla \phi_i$ as $\nabla \phi_i(T_k)$), cancelling the integration of c and leaving:

$$\sum_{k=1}^m s_k \langle \nabla \phi_i(T_k), \mathbf{p}_{k,\bullet} \rangle = 0.$$

We now try to come up with an expression for $\nabla \phi_i(T_k)$. Observe that $\nabla \phi_i(T_k)$ is zero if i is not a vertex of T_k , and otherwise it is $M_k \mathbf{e}_i$, where $\mathbf{e}_i$ is a vector with a '1' in the position corresponding to node i and zeros elsewhere. This product $M_k \mathbf{e}_i$ is equal to a column of M_k .

Thus, $\langle \nabla \phi_i(T_k), \mathbf{p}_{k,\bullet} \rangle$ is the inner product of a column of M_k with $\mathbf{p}_{k,\bullet}$. Summing over k , keeping track of the node indices, and recalling that A has the factor s_k built in, we see that in fact

$$\sum_{k=1}^m s_k \langle \nabla \phi_i(T_k), \mathbf{p}_{k,\bullet} \rangle$$

is the same as the inner product of the i th column of A with $\mathbf{p}$. Thus, (3) is equivalent to the linear system

$$A^T \mathbf{p} = \mathbf{0}. \quad (11)$$

Combining this with (8) yields

$$\begin{pmatrix} D & -A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \quad (12)$$

We call this system an “equilibrium system” following Strang [20]. Strang has observed that equilibrium systems occur in many contexts, including electrical networks, optimization, and structural analysis.

Since D is positive definite and A has full column rank, this system is invertible and hence uniquely determines $\mathbf{p}$ and $\mathbf{y}$.

This formulation is sometimes called a “mixed” formulation in the literature; see, for example, Roberts and Thomas [17]. Indeed, mixed formulations have been considered specifically for problems with ill-behaved coefficients by Babuška and Osborne [2] and by Falk and Osborne [10]. In fact, our approach is not a true mixed formulation because, as we will see in the next section, we do not solve (12); instead we reformulate the problem again so that $\mathbf{p}$ is not recovered by our method. The purpose of writing (12) is to separate A from D .

We now formally state the fact that the standard linear finite element method can be written in the form (12):

Theorem 1. *In exact arithmetic, the two linear systems (5) and (12) yield same value of $\mathbf{y}$.*

One final remark is that (12) is commonly called a “weighted least squares” problem in the linear algebra community. The case when D is ill-conditioned in this context is known as the “stiff” case. See Björck [6].

4 The NSH method

From (12) we can obtain a linear system for $\mathbf{y}$ alone. Multiply (8) by D^{-1} to obtain $\mathbf{p} = D^{-1}A\mathbf{y} + D^{-1}\mathbf{b}$, and substitute this in (11) to obtain

$$A^T D^{-1} A \mathbf{y} = -A^T D^{-1} \mathbf{b}. \quad (13)$$

and thus

$$\mathbf{y} = -(A^T D^{-1} A)^{-1} A^T D^{-1} \mathbf{b}. \quad (14)$$

It can be argued that (13) is identical to (5); indeed, $K = A^T D^{-1} A$ and $\mathbf{f} = A^T D^{-1} \mathbf{b}$. Accordingly, we do not want to try to solve (13) directly with Cholesky factorization. But (14) has a remarkable property that was established by Stewart [18]. For this theorem we imagine that $m' = md$.

Theorem 2 ([18]). *Let A be an arbitrary $m' \times n$ matrix of rank n . Let $\|\cdot\|$ be a matrix norm that is induced by a vector norm. Let $\mathcal{D}(m')$ denote the set of all positive definite $m' \times m'$ diagonal matrices. Define*

$$\chi_A = \sup\{\|(A^T D^{-1} A)^{-1} A^T D^{-1}\| : D \in \mathcal{D}(m')\}$$

and

$$\bar{\chi}_A = \sup\{\|A(A^T D^{-1} A)^{-1} A^T D^{-1}\| : D \in \mathcal{D}(m')\}.$$

Then $\chi_A, \bar{\chi}_A$ are both finite.

Further theory concerning these parameters was developed by O'Leary [14]. As a consequence of the results of Stewart and O'Leary, there exists an algorithm for computing χ_A and $\bar{\chi}_A$, although its running time is unfortunately exponential.

If we combine Theorem 1 with Theorem 2 we immediately obtain the following consequence.

Theorem 3. *Let $\hat{u}$ be the piecewise-linear finite element approximate solution to (1). Let $\|\cdot\|$ be some norm for piecewise linear functions on Ω . Then*

$$\|\hat{u}\| \leq C \|g\|_{L_\infty(\partial\Omega)}$$

where C is a constant that depends on m, n, d , on the norm, and on the triangulation, but not on g nor on the conductivity c .

This theorem is an example of a “discrete maximum principle.” Ciarlet [8] has written about discrete maximum principles, although this particular theorem does not seem to appear in [8]. Wahlbin [23] has since strengthened Theorem 3 in the $d = 2$ case. If the ∞ -norm is used, and all the triangles are nonobtuse, then Wahlbin has shown that we can take $C = 1$ in the preceding

theorem. In Section 6 we will determine an estimate for C for the general case.

The existence of this discrete maximum principle suggests that the discrete problem of solving (13) should have a stable algorithm because the solution is insensitive to perturbations.

Indeed, Vavasis [22] has proposed a stable floating-point numerical algorithm for (13). We now describe this algorithm, which is called the NSH method. NSH stands for “nullspace scaled hybrid” method.

1. Find a nullspace basis V for $A^T D^{-1}$, that is, an $md \times (md - n)$ matrix of rank $md - n$ such that $A^T D^{-1} V$ is zero. The exact procedure for finding V is described below. Matrix V must have the following properties:
 - (a) It is well-conditioned in the sense that for any $\mathbf{x}$, $\|V\mathbf{x}\|$ is bounded above and below by constant multiples of $\|\mathbf{x}\|$.
 - (b) If a numerical approximation $\hat{V}$ is computed (for instance, using floating point arithmetic), then $\hat{V}$ must be close to a true nullspace basis V of $A^T D^{-1}$ in the sense that $\|V - \hat{V}\|/\|V\|$ is small (i.e., $\hat{V}$ must have a small “relative normwise forward error”).
 - (c) Matrix V is scaled so that $\|V\| \approx \|A\|$.
2. Solve the square linear system

$$A\mathbf{y} + V\mathbf{q} = -\mathbf{b} \tag{15}$$

using any standard stable algorithm, e.g., Gaussian elimination with partial pivoting. In exact arithmetic, this system yields the same solution $\mathbf{y}$ as (13); this is seen by multiplying (15) by $A^T D^{-1}$. Vector $\mathbf{y}$ is the solution, and vector $\mathbf{q}$ is discarded.

The procedure for finding V is as follows.

1. Sort the diagonal entries of D into ascending order.
2. Form an $n \times n$ nonsingular matrix B , a submatrix of A , as follows. Loop through rows of A following the sorted order defined in the previous step. For each row, append it to B if it is linearly independent from the rows previously inserted in B . If it is dependent, then discard it. Proceed until B has n rows.

3. Solve for all the nonbasic rows of A in terms of the basis B . For each nonbasic row i , we obtain a vector in the nullspace of A^T . Specifically, this vector has a ‘ -1 ’ in position i , and the coefficients from solving for row i in the positions corresponding to B . If we concatenate these nullspace vectors together, we obtain a matrix Z that is a nullspace basis for A^T . Further, Z contains an embedded copy of $-I$, the negative identity matrix, in nonbasic rows.
4. Obtain V from the computation

$$V = \|A\| \cdot DZR,$$

where D is the original diagonal matrix from (12) and R is an $(m' - k) \times (m' - k)$ diagonal matrix uniquely determined by the requirement that we want DZR to also have an embedded copy of the negative identity matrix in nonbasic rows. In other words, R should “unscale” the embedded identity matrix that was scaled by D .

The theory supporting the NSH method appears in [22]. The main point of the method is the following theorem.

Theorem 4. *In floating point arithmetic, the NSH method returns a solution $\hat{\mathbf{y}}$ to (13) such that, when compared to the true solution $\mathbf{y}$, we have the following bound:*

$$\|\mathbf{y} - \hat{\mathbf{y}}\| \leq f(A) \cdot \epsilon \cdot \|\mathbf{b}\|,$$

where ϵ is machine-epsilon, and $f(A)$ depends on $\bar{\chi}_A$, χ_A , and $\|A\|$ but not on D .

5 Refining the theory: the NSHI method

We run into a difficulty when we try to apply the algorithm and theory described in the last section to (1). The difficulty is that the two parameters χ_A , $\bar{\chi}_A$ depend on the geometry of the triangulation in an undesirable way: these parameters depend on alignments of the simplex faces with the coordinate axes. In this section we describe a variant of the algorithm and a refinement of the theory to overcome this problem.

First, we must explain why the dependence on alignment arises. It should be observed that (1) is *isotropic*, that is, the PDE is invariant under a $d \times d$ orthogonal changes of coordinates. In contrast, our method for forming A depends on the coordinate system because the components of the gradient in Cartesian coordinates are encoded in the formation of A .

Another way to understand this dependence is to consider the two-dimensional *anisotropic* PDE

$$\frac{\partial}{\partial x} \left(a(x, y) \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(b(x, y) \frac{\partial u}{\partial y} \right) = 0 \quad (16)$$

Here, $a(x, y)$ and $b(x, y)$ are positive scalar fields. (This PDE generalizes in an obvious way to d dimensions.) If $a(x, y)$ is significantly larger or smaller than $b(x, y)$, then the solution to the PDE is sensitive to the orientation of the triangles.

One observes that (1) and (16) have exactly the same geometry matrix A , and hence the same values of $\chi_A, \bar{\chi}_A$. Therefore, we should not be surprised that the orientation of the triangulation affects these parameters.

Even though A is the same for (1) and (16), D has special structure for (1). Specifically, for (16), D may be an arbitrary matrix in $\mathcal{D}(md)$, but for (1), D has blocks of d consecutive equal entries. Let us write the set of such D formally as follows:

$$\mathcal{D}(md, d) = \{D \in \mathcal{D}(md) : D \text{ can be written as } D' \otimes I, \\ \text{where } D' \in \mathcal{D}(m) \text{ and } I \text{ is the } d \times d \text{ identity matrix.}\}$$

We can now define two new parameters for A :

$$\chi_{A,d} = \sup\{\|(A^T D^{-1} A)^{-1} A^T D^{-1}\| : D \in \mathcal{D}(md, d)\}$$

and

$$\bar{\chi}_{A,d} = \sup\{\|A(A^T D^{-1} A)^{-1} A^T D^{-1}\| : D \in \mathcal{D}(md, d)\}.$$

Obviously Stewart's theorem implies that these parameters are finite and are at most $\chi_A, \bar{\chi}_A$ respectively.

These parameters are truly isotropic as the following argument shows. Let W be an $md \times md$ rotation matrix defined to be block diagonal: $W = \text{diag}(Q, Q, \dots, Q)$, where Q is an arbitrary $d \times d$ orthogonal matrix. It is

easy to show that, in the 2-norm, $\chi_{A,d}$ and $\bar{\chi}_{A,d}$ are equal to $\chi_{WA,d}$ and $\bar{\chi}_{WA,d}$ respectively.

Most of the theory of stable algorithms described in the last section goes through when we restrict D to $\mathcal{D}(md, d)$ and when we substitute $\chi_{A,d}$ and $\bar{\chi}_{A,d}$ for χ_A and $\bar{\chi}_A$. The only step of the algorithm that must be modified is the computation of V in order to satisfy the three conditions enumerated in the last section.

It is possible to propose an algorithm for obtaining V satisfying bounds depending on $\bar{\chi}_A, \bar{\chi}_{A,d}$ for a general $m' \times n$ matrix A , but it seems that this algorithm would be very complicated and would probably not have many applications beyond (1).

Therefore, we instead explain how to form V only for the special case that A is a reduced geometry matrix arising in (1). The NSHI method is completely defined once we have explained this step, since all that remains is solving (15). NSHI stands for “NSH-isotropic.”

First, observe that A can be written as a product:

$$A = SJ^{-1}A^\circ$$

where S is a diagonal matrix with the s_k scaling factors, J is a block-diagonal matrix where each block is the Jacobian J_k defined in Section 3, and A° is an $md \times n$ matrix consisting of m copies of M° (defined by (10)) with the entries scattered into appropriate columns, and some columns deleted. To simplify the discussion in this section, we want the “origins” of the simplices to be selected as nonboundary nodes. (Recall from Section 3 that the origins may be selected arbitrarily without changing A , although J and A° will be different). The consequence of this choice of origin is that every row of A° has at least one nonzero entry.

It is not hard to see that A° is actually a *reduced node-arc incidence* (RNAI) matrix of the following directed graph G : G has n_0 nodes, one for each node of the triangulation. It has an arc from node i to i' whenever $(v_i, v_{i'})$ is an edge of some simplex T_k and i is the origin of T_k . Note that G may have multiple copies of edges; we always keep these multiple copies present.

The *node-arc incidence* matrix for G is a matrix A_0° with one row for each edge of G and one column for each node. Each row contains two nonzero entries, a ‘1’ and a ‘−1.’ The ‘−1’ entry occurs in the column corresponding

to the tail of the arc, and the ‘1’ entry occurs in the column corresponding to its head.

The *reduced node-arc incidence* (RNAI) matrix A° for G is obtained by deleting all the columns of A_0° corresponding to boundary nodes. The deleted nodes are said to be “grounded” in analogy with electrical engineering. In an RNAI matrix, each row has either a ‘1’/‘-1’ pair of entries, or a lone ‘1’ or lone ‘-1’ in rows corresponding to arcs with one end grounded. (No arc has both ends grounded by choice of origins.)

The first step toward obtaining V is to compute a nullspace basis Z° of $A^{\circ T}$. The process of finding Z° is described in [22] and is purely combinatorial. We compute a *minimum-weight spanning tree* of the graph G_1 , where weights are assigned to each arc. The graph G_1 is obtained by coalescing all the grounded nodes in G to a single node, and then removing the orientation from the arcs. Arcs without orientation are commonly called edges, but we will not change terminology. The weight to use on arc l is the value of the diagonal entry of D corresponding to arc l , that is, the quantity (9) where k is the simplex that gave rise to arc l . Each column of Z° then corresponds to the *induced cycle* for one nontree arc of G .

Basis Z° has the following properties. First, each entry is either 0, 1 or -1. (This of course implies that Z° may be computed exactly, without any floating point error.) Second, Z° has an embedded $(m' - n) \times (m' - n)$ copy of the identity matrix; the $m' - n$ rows holding the identity matrix correspond to the $m' - n$ nontree arcs.

Now we form a nullspace basis V for A by multiplying the product:

$$V = DS^{-1}J^T Z^\circ R. \quad (17)$$

In this formula, R is an $(m' - n) \times (m' - n)$ diagonal matrix that we specify below, and all the other factors have already been introduced. Since $A^{\circ T} Z^\circ = 0$, it is apparent that V with this definition is indeed a nullspace basis for $A^T D^{-1}$.

In order to define R , we need some notation to manage the subscripts. Let l be the index of a column of Z° , that is, an integer between 1 and $m' - n$. Then l corresponds to a nontree arc of G . Arcs in G are identified by two indices, $(k(l), \mu(l))$, where $k(l)$ lies between 1 and m and specifies the simplex giving rise the arc, and $\mu(l)$ lies between 1 and d . We can think of $(k(l), \mu(l))$ as a row index for the RNAI matrix A° ; it is the row index of the nontree arc

and corresponds to row number $(k(l) - 1)d + \mu(l)$ according to our standard numbering of the md rows of A° .

These indices $(k(l), \mu(l))$ also form a row index for Z° (because the rows of Z° are in correspondence with the rows of A°). This particular row of Z° has a single '1' entry (in column l), with all the other entries being zero, because Z° has an embedded identity matrix.

The diagonal matrix R is defined as follows; the l th diagonal entry is defined to be the reciprocal of (9) with k taken to be $k(l)$. Let us call this quantity $r(k(l))$. Thus, R has been defined to cancel out a certain entry of D .

We now would like show that V has the three desired properties stated in the last section. First, we consider condition (a). Let us examine the product $V\mathbf{x}$ for an arbitrary $(m' - n)$ -vector $\mathbf{x}$. The product $R\mathbf{x}$ scales the l th entry by $r(k(l))$. Next, the product $Z^\circ R\mathbf{x}$, an md -vector, has as its $(k(l), \mu(l))$ entry exactly $r(k(l))x(l)$.

For this $k = k(l)$, let us consider all the entries in block k of $Z^\circ R\mathbf{x}$. This block consists of d entries that all arise from the same simplex of the triangulation. For the entries corresponding to nontree arcs, as argued in the preceding paragraph, these entries have the form $r(k)x(l)$ for some l . For entries corresponding to tree arcs, these entries are sums of terms of the form $\pm r(k(l'))x(l')$, where l' is a nontree arc whose induced cycle contains a tree arc in block k . This is because a tree row of Z° contains entries that are ± 1 or zero, with the ± 1 entries corresponding to columns indexing nontree arcs that contain the tree arc in their induced cycles.

It is a property of minimum-weight spanning trees that the any nontree arc must have a higher weight than all of the arcs on the cycle that it induces. In particular, this means that $r(k(l')) \leq r(k)$, because $1/r(k(l'))$ is the weight on arc l' , and $1/r(k)$ is the weight on the tree arc in block k .

Thus, we see that the k th block of $Z^\circ R\mathbf{x}$ contain two kinds of entries: either it has entries of the form $r(k)x(l)$, or entries of the form $\pm r(k(l_1))x(l_1) \pm \dots \pm r(k(l_s))x(l_s)$ where each $r(k(l_p))$ is at most $r(k)$. Let $\mathbf{y}$ be this product $Z^\circ R\mathbf{x}$.

Next we consider the product $S^{-1}J^T Z^\circ R\mathbf{x}$, i.e., $S^{-1}J^T \mathbf{y}$. Since $S^{-1}J^T$ is block diagonal, it acts on each block of $\mathbf{y}$ independently. It will be shown in the next section that the norm of each diagonal block J_k^T/s_k in this block diagonal matrix is bounded above by a constant c_d depending only on the dimension d . Furthermore, the norm of the inverse of J_k^T/s_k is also bounded

above by a constant α depending on d and on the “aspect ratio” of the triangulation. Therefore, the norm of each block of $S^{-1}J^T\mathbf{y}$ is within a constant multiple of the norm of the corresponding block of $\mathbf{y}$.

Finally, when we form the product $\mathbf{z} = DS^{-1}J^T\mathbf{y}$, this has the effect of scaling the k th block of $S^{-1}J^T\mathbf{y}$ by $1/r(k)$.

We claim that the end result is a vector $\mathbf{z}$ within a constant factor of the norm of $\mathbf{x}$. Define $t = \|\mathbf{x}\|_\infty$, and let λ be chosen so that $|x_\lambda| = t$. First, we track the chain of computations for block $k = k(\lambda)$. In this block, one entry of $\mathbf{y}$ is $r(k)x_\lambda$. Therefore, the norm of this block of $\mathbf{y}$ is at least $r(k)t$. Following the arguments in the preceding paragraphs, this means that the norm of the k block of $\mathbf{z}$ is at least $r(k)t/(c_d r(k)\alpha)$, that is, at least a constant multiple t . This shows that $\|\mathbf{z}\|$ is at least a constant multiple of $\|\mathbf{x}\|$.

For the other direction, let k be the index of an arbitrary block of $\mathbf{y}$. The nontree entries of the block are $r(k)x(l)$, i.e., absolute value at most $r(k)t$. The tree entries have absolute value at most $r(k)tm'$, using the arguments above. Now, applying $S^{-1}J^T$ multiplies this by at most c_d , and applying D cancels the $r(k)$ factor. This shows that $\|\mathbf{z}\|$ is at most $c_d m'$ times $\|\mathbf{x}\|$. Let us state this as a lemma.

Lemma 1. *For an arbitrary $\mathbf{x}$, $\|V\mathbf{x}\|_\infty$ is at least $\|\mathbf{x}\|_\infty/(c_d\alpha)$ and at most $c_d m'\|\mathbf{x}\|_\infty$.*

Next we consider condition (b), which states that in floating point arithmetic, the computed $\hat{V}$ must be close to V . The computation of Z° is exact. Let us next consider the computation of $S^{-1}J^T Z^\circ$. This multiplication acts independently on columns of Z° , and furthermore, it acts independently on each block of each column. Since each diagonal block of $S^{-1}J^T$ is well-conditioned, this implies that each block of each column of $S^{-1}J^T Z^\circ$ can be computed with a small relative normwise error in floating-point arithmetic.

Next, we compute $DS^{-1}J^T Z^\circ$. Since, on each block, D has identical entries, the property that each block has a small relative forward normwise error is preserved. (Note that if D were arbitrary, this would not be true: for example, $[10^{-6}, 1]^T$ is a close relative normwise approximation to $[0, 1]^T$, but this property is not preserved under multiplication by $\text{diag}(10^{12}, 1)$.)

Thus, each block still has a small relative normwise error. Finally, we multiply on the right by R ; this does not change the property since R scales columns and hence scales each block by the same factor.

Thus, we see that V can be computed accurately in floating point arithmetic.

The last property is that $\|V\| \approx \|A\|$. This is easily achieved with one final scaling. It turns out however, that this scaling is not necessary; a consequence of the arguments of this section is that both $\|V\|$ and $\|A\|$ are close to unity—they are off from 1 by small constant factors such as c_d , m' , and α .

A conclusion of the arguments in this section, coupled with the earlier theory of the NSH method, is the following theorem.

Theorem 5. *With floating point arithmetic, the NSHI method returns a solution $\hat{\mathbf{y}}$ to (13) such that, when compared to the true solution $\mathbf{y}$, we have a bound of the following form:*

$$\|\mathbf{y} - \hat{\mathbf{y}}\| \leq f(\chi_{A,d}, \bar{\chi}_{A,d}, \alpha, m') \cdot \epsilon \cdot \|\mathbf{b}\|,$$

where ϵ is machine-epsilon.

6 Analysis of the geometric factors

In this section we try to analyze the numerical stability bound Theorem 5 to understand how it depends on the geometry. When no norm is indicated, we will generally mean the ∞ -norm in this section. The first observation is that the error bound depends on $\|\mathbf{b}\|$; we would prefer to have it depend on $\|\mathbf{y}_0\|$, the boundary data.

Recall that $\mathbf{b}$ is obtained from $\mathbf{y}_0$ by multiplication by a subset of columns of A_0 , that is, by columns of matrices of the form $s_k J_k^{-1} M^\circ$. Thus, we can assert that

$$\|\mathbf{b}\| \leq \max_k \|s_k J_k^{-1} M^\circ\| \cdot \|\mathbf{y}_0\|.$$

Note that $\|M^\circ\| = 2$ so we can ignore that factor. Thus, we must be concerned with $\|s_k J_k^{-1}\|$.

Let us define the *aspect ratio* of a simplex to be its maximum side length divided by its minimum altitude. It can be shown that, for $d = 2$, the aspect ratio is bounded above and below by constant multiples of the reciprocal of the smallest angle of the triangle. We have the following two lemmas connecting aspect ratios to $s_k J_k^{-1}$:

Lemma 2. *Let T_k be a simplex with aspect ratio a . Then*

$$\|s_k J_k^{-1}\| \leq c_d a$$

where c_d is a constant depending on d .

Notation: In this section c_d will denote a factor depending on d whose value may change from equation to equation.

Note that J_k is not uniquely defined until the origin of T_k has been selected. In order to preserve generality, the following proof, and also the proofs of the upcoming lemmas, will not depend on which node is chosen as origin.

Proof. For this proof we will switch back and forth between several norms; changing norms is an additional constant-factor contribution to c_d . Let $\mathbf{v}^1, \dots, \mathbf{v}^{d+1}$ be the vertices of T_k . To simplify notation, assume $\mathbf{v}^{d+1} = \mathbf{0}$ so that $J_k = [\mathbf{v}^1, \dots, \mathbf{v}^d]^T$. Let $\mathbf{u}$ be an arbitrary d -vector whose ∞ -norm is 1. Without loss of generality, assume the maximum entry is actually 1 (not -1). Let μ be the position of the maximum entry. Let $\mathbf{w} = J_k^T \mathbf{u}$. Then

$$\begin{aligned} \mathbf{w} &= \sum_{\nu=1}^n u_\nu \mathbf{v}^\nu \\ &= \mathbf{v}^\mu - \sum_{\substack{\nu=1 \\ \nu \neq \mu}}^n (-u_\nu) \mathbf{v}^\nu. \end{aligned}$$

Observe that the summation term in the second line lies in the plane spanned by $\mathbf{0}$ and $\mathbf{v}^1, \dots, \mathbf{v}^d$, excluding $\mathbf{v}^\mu$. Thus, this point lies in the plane that contains one of the facets of T_k (the facet going through $\mathbf{v}^1, \dots, \mathbf{v}^d$ excluding $\mathbf{v}^\mu$). Hence $\mathbf{w}$ is the vector from a vertex of T_k , namely $\mathbf{v}^\mu$, to a point on the plane containing the opposite facet. Thus, $\|\mathbf{w}\|_2 \geq t$, where t is the length of the smallest altitude of T_k .

Now we have a norm computation.

$$\begin{aligned} \|J_k^{-1}\|_2 &= \|J_k^{-T}\|_2 \\ &= \|J_k^{-T} \mathbf{w}\|_2 / \|\mathbf{w}\|_2 \\ &\leq \|\mathbf{u}\|_2 / t \\ &\leq c_d / t. \end{aligned}$$

The second line is based on the fact that $\mathbf{u}$ and hence $\mathbf{w}$ was arbitrary, and therefore may be chosen to be the d -vector that determines $\|J_k^{-T}\|_2$. The third line follows because $\mathbf{u} = J_k^{-T}\mathbf{w}$. The fourth line follows because $\|\mathbf{u}\|_\infty = 1$.

Thus, $\|s_k J_k^{-1}\|_2 \leq c_d s_k / t$. Finally, s_k / t is a by definition. ■

This shows that b is bounded in terms of y_0 , where the bound depends on the aspect ratio. From now on, we let α denote the maximum aspect ratio occurring in the triangulation.

The previous theorem establishes a bound on $\|s_k J_k^{-1}\|$. We can also establish a bound on the inverse matrix, which is independent of α .

Lemma 3.

$$\|s_k^{-1} J_k\| \leq 1.$$

Proof. This follows because each row of J_k / s_k has 2-norm at most 1, because s_k is the longest side-length. Therefore, the largest matrix entry is 1 in absolute value, so the infinity norm is at most 1. ■

These two lemmas have a converse that is also true: the aspect ratio of T_k is bounded above by $c_d \|J_k\| \cdot \|J_k^{-1}\|$. Again, assume for notational simplicity that $\mathbf{v}^{d+1} = 0$. Let a be the length of the smallest altitude. There are two subcases: either the smallest altitude is the altitude to one of $\mathbf{v}_1, \dots, \mathbf{v}_d$, or it is the altitude to $\mathbf{v}_{d+1} = 0$. In the first case, suppose e.g. that the smallest altitude connects $\mathbf{v}_1$ to a point in the plane spanned by $\mathbf{v}_2, \dots, \mathbf{v}_d, 0$; say that point is $t_2 \mathbf{v}_2 + \dots + t_d \mathbf{v}_d$. Let $\mathbf{t}$ be the vector $(1, -t_2, \dots, -t_d)^T$. Then the altitude in question is precisely $J_k^T \mathbf{t}$.

The other subcase is that the smallest altitude is from the origin to a point $t_1 \mathbf{v}_1 + \dots + t_d \mathbf{v}_d$, where $t_1 + \dots + t_d = 1$. Let $\mathbf{t}$ denote $(t_1, \dots, t_d)^T$. Then the segment again may be written $J_k^T \mathbf{t}$.

In both subcases, $\|J_k^T \mathbf{t}\|_2 = a$ and $\|\mathbf{t}\|_1 \geq 1$, hence

$$\|J_k^{-1}\|_2 \geq c_d / a.$$

Next, let b be the length of the longest edge; it is seen that b may be written as $\|J_k^T \mathbf{u}\|_2$ where $\mathbf{u}$ is a vector with either a single '1' entry, or two ' ± 1 ' entries. Thus, $\|\mathbf{u}\|_2 \leq 1.5$, so $1.5 \|J_k\|_2 \geq b$. Combining this with the inequality in the preceding paragraph yields:

$$\|J_k\|_2 \cdot \|J_k^{-1}\|_2 \geq c_d b / a.$$

Now we turn to $\chi_{A,d}$ and $\bar{\chi}_{A,d}$. Theorem 2 shows that these are finite for an arbitrary A , and therefore for our specific A . Stewart's proof, however, is nonconstructive (it involves a compactness argument), and we would like to get specific bounds on these parameters in terms of the geometry. Therefore, we need to prove new bounds on $\chi_{A,d}$ and $\bar{\chi}_{A,d}$ for a geometry matrix A . This is the content of Theorem 6 and Theorem 7. Theorem 7 below achieves a much better bound for a restricted class of triangulations.

Theorem 6. *Let α be the maximum aspect ratio among all simplices of the triangulation. Then*

$$\chi_{A,d} \leq 10(c_d \alpha)^{n_0}$$

and

$$\bar{\chi}_{A,d} \leq 10md(c_d \alpha)^{n_0+1}.$$

Proof. Let us assume that $\mathbf{b}$ is an arbitrary vector, and that D is an arbitrary symmetric positive definite matrix. Let $\mathbf{y}$ be the solution to $A^T D^{-1} A \mathbf{y} = A^T D^{-1} \mathbf{b}$. In order to bound χ_A , we need a bound on $\|\mathbf{y}\|$. Note that the solution $\mathbf{y}$ does not necessarily correspond to a solution of (1); the reason is that vector $\mathbf{b}$ arising in a solution of (1) has a special sparsity pattern (its nonzero entries occur only for simplices adjacent to the boundary). In order to obtain the needed bound, we cannot assume any special structure in $\mathbf{b}$.

Instead of working with (13), we work with (12). Specifically, assume we have vectors $\mathbf{p}$ and $\mathbf{w}$ such that $A^T \mathbf{p} = \mathbf{0}$, $D \mathbf{p} = \mathbf{w} + \mathbf{b}$, and $\mathbf{w} = A \mathbf{y}$.

Let A_0 be the extended geometry matrix defined in Section 3, that is, an extension of A to n_0 columns. Then, since $\mathbf{w} = A \mathbf{y}$, we also have $\mathbf{w} = A_0 \bar{\mathbf{y}}$, where $\bar{\mathbf{y}}$ is an extension of $\mathbf{y}$ with zeros in the boundary positions. Let us call the value of $\bar{\mathbf{y}}$ at each node v its *label* and denote it as $\bar{y}(v)$.

Now we define a new set of labels that we denote $z(v)$, which are supposed to approximate the $\bar{\mathbf{y}}$ labels in a certain sense. The $\mathbf{z}$ labels, unlike the $\bar{\mathbf{y}}$ labels, will have a discreteness property: there will be a lower bound on the difference between any two distinct $\mathbf{z}$ labels. Let $\beta > 1$ be a constant that depends on the maximum aspect ratio; the exact value for β is provided below. The procedure for determine the z 's is follows. We write the $\bar{\mathbf{y}}$ labels in ascending order; let the sorted list be written $\bar{y}_1, \dots, \bar{y}_{n_0}$. Let t denote the span of this list:

$$t = \bar{y}_{n_0} - \bar{y}_1.$$

We define a set $\mathcal{I}$ of intervals that gets repeatedly updated. Initially, $\mathcal{I}$ contains n_0 intervals $\mathcal{I} = \{[\bar{y}_1, \bar{y}_1 + \epsilon], [\bar{y}_2, \bar{y}_2 + \epsilon], \dots, [\bar{y}_{n_0}, \bar{y}_{n_0} + \epsilon]\}$, where we define $\epsilon = t(\beta + 2)^{-n_0}$.

Let $[a, b]$ and $[c, d]$ be two intervals in $\mathcal{I}$ such that $a \leq c$. We will say that these intervals are “crowded” if

$$c - b < \beta \max(b - a, d - c).$$

(In particular, if the intervals intersect, i.e., $c \leq b$, then they are certainly crowded.) We repeatedly perform the following “merge” operation on $\mathcal{I}$: if there are two intervals $[a, b], [c, d]$ that are crowded, we delete them and replace with the new interval $[\min(a, c), \max(b, d)]$. This operation can clearly be performed at most n_0 times since each time the operation is performed, the number of intervals decreases by 1. Note that the union of the intervals in $\mathcal{I}$ never decreases because $[a, b] \cup [c, d] \subset [\min(a, c), \max(b, d)]$.

When we are done, the union of the intervals in $\mathcal{I}$ contains all of the original $\bar{y}(v)$ values because of the nondecreasing property. We assign a new label $z(v)$ to each node v of the triangulation defined to be the left endpoint of the interval of the terminal $\mathcal{I}$ that contains $\bar{y}(v)$.

This construction of z has several important properties. First, not all the z ’s are equal. In particular, $\bar{y}_1$ and $\bar{y}_{n_0}$ cannot end up in the same interval. The proof of this claim is as follows. Initially, each interval has length ϵ . After merging two intervals, say of length p, q , it follows from the construction that the length of the new interval is less than $p + q + \beta \max(p, q)$, i.e., less than $(\beta + 2) \max(p, q)$. Thus, after every merge operation, the longest interval in $\mathcal{I}$ grows by less than $\beta + 2$. By choice of ϵ , the longest interval can never reach length t .

Another property of the z ’s is that any two distinct z ’s differ by at least ϵ . This follows because all the intervals in the final $\mathcal{I}$ have length at least ϵ , and no pair of intervals is crowded.

A third property is as follows. Let v_1, v_2, v_3 be three nodes such that $z(v_1) = z(v_2) \neq z(v_3)$. Then $|\bar{y}(v_3) - \bar{y}(v_1)| \geq \beta |\bar{y}(v_2) - \bar{y}(v_1)|$, and, moreover, $\bar{y}(v_3) - \bar{y}(v_1)$ is approximated by $z(v_3) - z(v_1)$:

$$\left| \frac{(\bar{y}(v_3) - \bar{y}(v_1)) - (z(v_3) - z(v_1))}{z(v_3) - z(v_1)} \right| \leq \frac{1}{\beta}. \quad (18)$$

These claims follow from the noncrowded property.

Collect all of the values of $z(v)$ into an n_0 -vector named $\mathbf{z}$. Let $\mathbf{r} = A_0 \mathbf{z}$. Since all the $\bar{\mathbf{y}}$ -labels are equal to each other at all boundary nodes (they are all zero), the $\mathbf{z}$ labels at boundary nodes are also equal. Let $\mathbf{e}$ be the vector of all 1's; note from (10) that $\mathbf{e}$ is in the nullspace of A_0 . Thus, $\mathbf{r} = A_0(\mathbf{z} + \theta \mathbf{e})$ for any θ . In this last expression we can choose θ so that $\mathbf{z} + \theta \mathbf{e}$ has zeros for all boundary nodes. The conclusion is that vector $\mathbf{r}$ lies in the image of A , as well as the image of A_0 .

Since $\mathbf{r}$ is in the image of A and $A^T \mathbf{p} = 0$, then $\mathbf{r}^T \mathbf{p} = 0$. Since $\mathbf{r}$ itself is nonzero (because not all the entries of $\mathbf{z}$ are equal to each other), this means there exists a block of d consecutive entries, say $\mathbf{r}_k$, corresponding to simplex T_k , such that $\mathbf{r}_k$ is not all zeros and such that $\mathbf{r}_k^T \mathbf{p}_k \leq 0$.

Let us also split $\mathbf{w}$ into m blocks of size d denoted $\mathbf{w}_1, \dots, \mathbf{w}_m$. We now show that $\mathbf{w}_k^T \mathbf{r}_k \geq 0.1 \|\mathbf{r}_k\|_2^2$. Let $\mathbf{y}_k$ denote the $d+1$ vector of $\bar{\mathbf{y}}$ labels on simplex T_k , and $\mathbf{z}_k$ the $\mathbf{z}$ labels. Then $\mathbf{w}_k = s_k J_k^{-1} M^\circ \mathbf{y}_k$ and $\mathbf{r}_k = s_k J_k^{-1} M^\circ \mathbf{z}_k$. We claim that $M^\circ \mathbf{z}_k$ is a close approximation to $M^\circ \mathbf{y}_k$ in the sense of relative normwise error. The argument is as follows. First, recall that $\mathbf{r}_k$ is nonzero, implying that not all the $\mathbf{z}$ labels of T_k are equal. There are two kinds of entries in $M^\circ \mathbf{z}_k$: zeros, corresponding to vertices with the same $\mathbf{z}$ label as the origin, and nonzeros. Then relation (18) implies that, componentwise, the entries of $M^\circ \mathbf{z}_k$ that are nonzero are within a factor of $(1 \pm 1/\beta)$ of the corresponding entries of $M^\circ \mathbf{y}_k$. Where $M^\circ \mathbf{z}_k$ is zero, the corresponding entry of $M^\circ \mathbf{y}_k$ is a factor β smaller than any of the nonzero entries of $M^\circ \mathbf{z}_k$.

This implies that $M^\circ \mathbf{y}_k = M^\circ \mathbf{z}_k + \mathbf{f}$, where $\mathbf{f}$ is a vector whose 2-norm is at most $(c_d/\beta) \|M^\circ \mathbf{z}_k\|_2$. Now we compute the inner product $\mathbf{w}_k^T \mathbf{r}_k$.

$$\begin{aligned} \mathbf{w}_k^T \mathbf{r}_k &= \mathbf{y}_k^T M^{\circ T} J_k^{-T} s_k^2 J_k^{-1} M^\circ \mathbf{z}_k \\ &= \mathbf{z}_k^T M^{\circ T} J_k^{-T} s_k^2 J_k^{-1} M^\circ \mathbf{z}_k + \mathbf{f}^T J_k^{-T} s_k^2 J_k^{-1} M^\circ \mathbf{z}_k \\ &\geq \|\mathbf{r}_k\|_2^2 - (c_d/\beta) \|M^\circ \mathbf{z}_k\|_2 \cdot \|J_k^{-T} s_k\|_2 \cdot \|\mathbf{r}_k\|_2 \\ &\geq \|\mathbf{r}_k\|_2^2 - (c_d^2 \alpha / \beta) \|M^\circ \mathbf{z}_k\|_2 \cdot \|\mathbf{r}_k\|_2 \\ &= \|\mathbf{r}_k\|_2^2 - (c_d^2 \alpha / \beta) \|s_k^{-1} J_k \mathbf{r}_k\|_2 \cdot \|\mathbf{r}_k\|_2 \\ &\geq \|\mathbf{r}_k\|_2^2 - (c_d^3 \alpha / \beta) \|\mathbf{r}_k\|_2 \cdot \|\mathbf{r}_k\|_2. \end{aligned}$$

Thus, we can assert that $\mathbf{w}_k^T \mathbf{r}_k \geq 0.1 \|\mathbf{r}_k\|_2^2$ provided that $\beta \geq 1.2 \alpha c_d^3$. This is how β is selected.

Now we continue the argument, recalling the equation $D\mathbf{p} = \mathbf{w} + \mathbf{b}$. For the k th block this may be written $d_k \mathbf{p}_k = \mathbf{w}_k + \mathbf{b}_k$, where d_k is a scalar positive

constant (it is a diagonal entry of D). Take the inner product with $\mathbf{r}_k$ to obtain $d_k \mathbf{p}_k^T \mathbf{r}_k = \mathbf{w}_k^T \mathbf{r}_k + \mathbf{b}_k^T \mathbf{r}_k$. Since $\mathbf{r}_k^T \mathbf{p}_k \leq 0$, $\mathbf{w}_k^T \mathbf{r}_k \leq -\mathbf{b}_k^T \mathbf{r}_k$. Next, we plug in the inequality from the last paragraph to obtain $0.1 \|\mathbf{r}_k\|_2^2 \leq -\mathbf{b}_k^T \mathbf{r}_k$. Clearly $-\mathbf{b}_k^T \mathbf{r}_k \leq \|\mathbf{b}_k\|_2 \cdot \|\mathbf{r}_k\|_2$, so

$$0.1 \|\mathbf{r}_k\|_2^2 \leq \|\mathbf{b}_k\|_2 \cdot \|\mathbf{r}_k\|_2.$$

This shows that $\|\mathbf{r}_k\|_2 \leq 10 \|\mathbf{b}_k\|_2$. Since $\|\mathbf{r}_k\|_2 \geq \|M^\circ \mathbf{z}_k\|_2 / c_d$, we have $\|M^\circ \mathbf{z}_k\|_2 \leq 10 c_d \|\mathbf{b}_k\|_2$.

Next, we have $\|M^\circ \mathbf{z}_k\|_2 \geq \epsilon$ (this is because $\mathbf{z}_k$ has differing entries, and the minimum difference among distinct $\mathbf{z}$ -labels is ϵ), so $t(\beta + 2)^{-n_0} \leq 10 c_d \|\mathbf{b}_k\|_2$, i.e.,

$$t \leq 10 c_d (\beta + 2)^{n_0} \|\mathbf{b}\|,$$

Recall that t was the maximum difference between $\bar{\mathbf{y}}$ labels. Since the boundary nodes are labeled zero, the maximum absolute $\bar{\mathbf{y}}$ label is at most t . Thus,

$$\|\mathbf{y}\|_\infty \leq 10 c_d (\beta + 2)^{n_0} \|\mathbf{b}\|.$$

This gives the bound on $\chi_{A,d}$. Recall that $\bar{\chi}_{A,d}$ is at most $\|A\|_{\chi_{A,d}}$, and $\|A\|$ is bounded by $c_d \alpha m'$. ■

The exponential dependence on the number of nodes is rather alarming. We do not know whether the bound established by this theorem is as tight as possible. In our computational experiments, we have not encountered any exponentially large values of $\|(A^T D^{-1} A)^{-1} A^T D^{-1}\|$. It seems that if the worst-case bound is indeed exponential in n_0 , it probably takes a very special choice of D to obtain this large norm.

There is a special case for which we know how to establish a much better bound.

Theorem 7. *Suppose that, in addition to having aspect ratio bound α , all the simplices T_k are **nonobtuse** in the sense that that angle between any pair of $(d - 1)$ -dimensional facets is at most $\pi/2$. Then*

$$\chi_{A,d} \leq n_0 \alpha c_d$$

and

$$\bar{\chi}_{A,d} \leq n_0 m \alpha^2 c_d.$$

Proof. Under the conditions of the theorem, it turns out that the symmetric positive definite matrix $N = J_k^{-T} J_k^{-1}$ is a Stieltjes matrix. Recall that a matrix N is said to be *Stieltjes* if it is symmetric, has positive diagonal entries and nonpositive off-diagonal entries, and if its inverse has all nonnegative entries.

The theory of Stieltjes matrices is well-known, and presented, e.g., by Ortega [15]. For our purposes, we need some variants of the standard results that we now derive.

Let the vertices of T_k be $\{v^1, \dots, v^d, 0\}$. Let $e_1, \dots, e_d$ be the standard basis, and let e be the vector of all 1's. Let J_k be the matrix $[v^1, \dots, v^d]^T$.

Consider the inner product between the normals of two facets, say u and w . Since the angle between the faces is at most $\pi/2$, the angle between u and w is at least $\pi/2$ but less than π . Thus, $u^T w \leq 0$.

In terms of J_k , the normal of the facet containing all nodes except the i th is, up to scaling, $J_k^{-1} e_i$, for $i = 1, \dots, d$. The normal of the facet not containing the origin is $-J_k^{-1} e$. Thus, the conditions that the facets make angles at most $\pi/2$ is

$$e_i^T J_k^{-T} J_k^{-1} e_j \leq 0 \quad (19)$$

for $1 \leq i, j \leq d$ and $i \neq j$, and

$$e^T J_k^{-T} J_k^{-1} e_i \geq 0 \quad (20)$$

for $1 \leq i \leq d$.

Notice that (19) immediately implies that the off-diagonal elements of N are nonpositive, because the expression on the left is precisely the formula for the (i, j) entry of N .

Next, we show that N is diagonally dominant. Because of the pattern of signs of the entries of N , we need only obtain lower bounds on each entry of Ne in order to prove diagonal dominance—the i th entry of Ne is precisely diagonal entry of N minus the absolute values of the off-diagonal elements in column i . But the i th entry of Ne is precisely the left-hand side of (20). This proves that N is diagonally dominant.

Now, we return to the problem of bounding $\chi_{A,d}$ and $\bar{\chi}_{A,d}$. We follow the proof of Theorem 6, except we choose the z labels in a different manner. Since there are n_0 nodes, there must be two consecutive labels in the sorted list $\bar{y}_1, \dots, \bar{y}_{n_0}$, say $\bar{y}_{\text{cut}}, \bar{y}_{\text{cut}+1}$ such that $\bar{y}_{\text{cut}+1} - \bar{y}_{\text{cut}} \geq t/n_0$. For nodes whose $\bar{y}$ label is $\bar{y}_{\text{cut}}$ or less, we assign a z label of 0. For the remaining

nodes we assign a $\mathbf{z}$ label of 1. We denote this entire vector by $\mathbf{z}^\circ$ since it is combinatorial.

Form the vector $\mathbf{r} = A_0 \mathbf{z}^\circ$ as in the preceding proof. Note that not all the entries of $\mathbf{r}$ are zero. As in the preceding proof, select a simplex T_k such that $\mathbf{r}_k^T \mathbf{p}_k \leq 0$ and such that $\mathbf{r}_k$ is not all zeros.

We now try to put a positive lower bound on $\mathbf{w}_k^T \mathbf{r}_k$. Renumber the vertices of T_k so that $\mathbf{v}^1, \dots, \mathbf{v}^\mu$ have $\mathbf{z}$ -label of 1, and $\mathbf{v}^{\mu+1}, \dots, \mathbf{v}^{d+1}$ have $\mathbf{z}$ -label 0, and such that the origin $\mathbf{v}^{d+1}$ is the node with the maximum $\bar{\mathbf{y}}$ -label among the nodes with $\mathbf{z}$ -label of zero. (Note that since $\mathbf{r}_k$ is nonzero, both types of $\mathbf{z}$ labels must appear on T_k .) Let $\mathbf{z}_k^\circ$ denote the $\mathbf{z}$ labels of the vertices, and let $\mathbf{y}_k$ denote the $\bar{\mathbf{y}}$ labels.

Thus $M^\circ \mathbf{z}_k^\circ$ is a d -vector whose first μ entries are 1's, and whose last $d - \mu$ entries are 0's. Vector $M^\circ \mathbf{y}_k$ has as its first μ entries numbers of size t/n_0 or more. The last $d - \mu$ entries of $M^\circ \mathbf{y}_k$ are nonpositive.

Observe that, with these definitions, $\mathbf{w}_k = s_k J_k^{-1} M^\circ \mathbf{y}_k$ and $\mathbf{r}_k = s_k J_k^{-1} M^\circ \mathbf{z}_k^\circ$, so $\mathbf{w}_k^T \mathbf{r}_k = s_k^2 (M^\circ \mathbf{y}_k)^T N (M^\circ \mathbf{z}_k^\circ)$. Let $\mathbf{u}$ denote the d -vector $s_k^2 N (M^\circ \mathbf{z}_k^\circ)$. Because N is diagonally dominant, the first μ entries of $\mathbf{u}$ are all nonnegative. Similarly, because the off-diagonal entries of N are all nonpositive, the last $d - \mu$ entries of $\mathbf{u}$ are all nonpositive.

Thus, $\mathbf{u}$ and $M^\circ \mathbf{y}_k$ have the same sign pattern. We now want to argue that, of the first μ entries of $\mathbf{u}$, at least one of them has a positive lower bound. Consider the inner product $q = (M^\circ \mathbf{z}_k^\circ)^T s_k^2 N (M^\circ \mathbf{z}_k^\circ)$, which is the same as $\mathbf{u}^T (M^\circ \mathbf{z}_k^\circ)$. This inner product is thus equal to the sum of the first μ entries of $\mathbf{u}$. On the other hand, $q \geq s_k^2 / \|N^{-1}\|_2$, because the minimum possible value of $\mathbf{x}^T P \mathbf{x}$ when $\|\mathbf{x}\|_2 \geq 1$ is $1/\|P^{-1}\|_2$ for a symmetric positive definite matrix P . As argued earlier, $s_k^2 / \|N^{-1}\|_2$, which is equal to $(s_k / \|J_k\|_2)^2$, is at least c_d .

This shows that at least one of the first μ entries of $\mathbf{u}$ is of size c_d . Finally, since $\mathbf{w}_k^T \mathbf{r}_k = (M^\circ \mathbf{y}_k)^T \mathbf{u}$, we have that $\mathbf{w}_k^T \mathbf{r}_k \geq c_d t / n_0$.

Now we continue as in the proof of Theorem 6. To summarize, since $d_k \mathbf{p}_k = \mathbf{w}_k + \mathbf{b}_k$, we have $\mathbf{w}_k^T \mathbf{r}_k \leq -\mathbf{b}_k^T \mathbf{r}_k$. Substituting the inequality from the preceding paragraph yields

$$c_d t / n_0 \leq \|\mathbf{b}_k\|_2 \cdot \|\mathbf{r}_k\|_2.$$

Note that $\|\mathbf{r}_k\|_2 \leq c_d \alpha$, hence

$$c_d t / n_0 \leq c_d \alpha \|\mathbf{b}_k\|_2,$$

thus

$$t \leq c_d n_0 \alpha \|b_k\|_2.$$

This concludes the argument. ■

The difference between Theorem 6 and Theorem 7 is very striking; Theorem 7 has a polynomial dependence. If we were to assume that both bounds are in fact tight, then there is a significant difference between obtuse and nonobtuse triangulations. This difference may be explained as follows. Recall that the parameters $\chi_{A,d}$, $\bar{\chi}_{A,d}$ correspond to constants occurring in a certain “discrete maximum” principle. It is well-known that strong discrete maximum principles are more easily established in the presence of angles bounded by 90° ; see, for example, [21, p. 78] or [8, Theorem 20.2]. Compare this also to the result of [23] mentioned in Section 4.

7 Triangulation algorithms

A consequence of the results in the last section is that, for (1) with wild coefficients, one would like a triangulation of Ω with bounded aspect ratio and with all angles bounded by $\pi/2$.

There has been considerable research on algorithms for triangulations satisfying angle bounds. An excellent source of information is the survey by Bern and Eppstein [4]. An algorithm for obtaining nonobtuse triangulations of polygonal regions in the case $d = 2$, and with bounded aspect ratio, was proposed by Baker *et al.* [3]. Their algorithm does not handle varying levels of refinement (i.e., great variation in the size of the simplices in different parts of the domain.) Bern, Eppstein and Gilbert [5] proposed a class of triangulation algorithms with bounded aspect ratio that can vary the level of refinement. Using these ideas, Melissaratos and Souvaine [12] have an algorithm for nonobtuse triangulation with bounded aspect ratio that can vary the level of refinement.

In the case $d \geq 3$, we know of no work on nonobtuse triangulation of polygonal regions. There is only one algorithm we know of that attains bounded aspect ratio in the case $d = 3$, an octree-based algorithm due to Mitchell and Vavasis [13]. This algorithm handles varying levels of refinement.

Note that for all of these algorithms, if the boundary of Ω has a sharp angle, then the aspect ratio bound must go up, because the triangle/tetrahedron that fits inside the sharp angle must necessarily have a bad aspect ratio.

For the case $d > 3$, we know of no work on triangulations with bounded aspect ratio or with a nonobtuseness condition for polygonal regions.

8 Computational experiments

In this section we describe computational experiments to compare the NSHI method to the CFSM method for solving (1). All these experiments are for $d = 2$. The integration of the c fields over triangles has been carried out using the midpoint rule. All the tests have been conducted—and all figures produced—using MATLABTM. MATLAB, an interactive software package for numerical analysis, is a trademark of The Mathworks, Inc. All tests are in double-precision, with machine epsilon approximately $2 \cdot 10^{-16}$.

For the standard method, we have used sparse Cholesky factorization, as implemented by the `chol` instruction in sparse MATLAB, on the stiffness matrix. Naturally, there are many other possible methods for solving (5), such as iterative methods. We believe, however, that any method based on (5) will be inaccurate because the ill-conditioning K will affect any numerical algorithm.

We have implemented the NSHI method exactly as described in this paper. We have used Kruskal's algorithm for the minimum weight spanning tree computation (see Aho, Hopcroft, and Ullman [1, p. 237]), which requires $O(md \log(md))$ operations.

For our first group of test problems we used the mesh Ω_1 illustrated in Fig. 1. This mesh was obtained by placing an evenly spaced mesh of points inside a polygon, displacing points that fell close to the boundary, and then triangulating with the *constrained Delauney triangulation* (see [4]). This method generally gives good aspect ratio bounds, though some angles are obtuse.

As our first test we took $c = 1$ everywhere and boundary conditions depending linearly on planar coordinates x_1, x_2 . This problem has an exact analytic solution, namely, the extension of the linear function into the domain. Furthermore, in exact arithmetic, the finite element method should give the exact analytic solution (i.e., no truncation error is present) because

Figure 1: Mesh used for several test cases.

the finite element method is piecewise linear. Thus, the only errors are round-off errors. We found that both the CFSM method and the NSHI method gave computed answers $\hat{\mathbf{y}}$ such that $\|\mathbf{y} - \hat{\mathbf{y}}\|$ was bounded by about $10^{-15}\|\mathbf{y}_0\|$. (Recall that $\mathbf{y}_0$ represents boundary values.) The spanning tree computed by the NSHI method is depicted in Fig. 2.

For our second test, we again took $c = 1$ and boundary conditions so that the exact analytic solution is the harmonic function $x_1^2 - x_2^2$. This function cannot be exactly represented with linear elements, so we would expect truncation as well as roundoff errors. The relative norm error for both methods with respect to the boundary data for this test was $3.6 \cdot 10^{-4}$ for both methods. The relative difference between the two methods was about 10^{-15} .

For the next test, we constructed a problem with wild coefficients where know an approximate analytic solution. The coefficients are set up so that c is a very large constant value c' in a disk-shaped region U strictly interior to the domain, and c is 1 everywhere else. If the conductivity were essentially infinite in U , then u would have to be constant in U , and in particular, it would be constant on the boundary of U . If $c = 1$ outside, then u is harmonic outside U . An example of a function that is harmonic on the exterior of the unit disk, and constant on the boundary of the disk is

$$u(x_1, x_2) = x_2 - \frac{x_2}{x_1^2 + x_2^2} \quad (21)$$

Furthermore, this function has the necessary compatibility property that its flux $\partial u / \partial n$ averaged around the boundary of the disk is zero. Thus, if we choose boundary conditions to agree with (21) and we choose c' very large, then the exact analytic solution of (1) should be very close to constant inside U and close to (21) outside.

We took as the boundary of the domain the same polygon used in the previous cases, with a disk inside. The boundary conditions used were defined by (21), properly scaled and recentered for the inner disk, and with a constant added.

We took five different values for the c field. In all cases, c is 1 outside the disk, and is a large value c' on the disk. The results are summarized in Table 1. Note that for small values of c' both methods return acceptable solutions. For wilder coefficients, the stiffness equation returns the wrong answer, and finally breaks down, unable to complete the Cholesky factorization.

Figure 2: An example spanning tree generated by the NSHI method.

Table 1: This table summarizes the test where the coefficient in the disk is c' . The second column is the difference between the CFSM solution and the analytic solution. The third column is for the NSHI method. Omitted entries in the second column mean that the Cholesky factorization failed.

c'	CFSM error	NSHI error
10^5	3.14%	3.14%
10^{10}	3.14%	3.14%
10^{16}	103.12%	3.14%
10^{17}	—	3.14%
10^{40}	—	3.14%

The specific solutions in the case of $c' = 10^{16}$ are illustrated in Fig. 3.

As mentioned in the last paragraph, we added a small constant to the boundary conditions. In principle, adding a constant to the boundary conditions should affect (1) only by translating the solution by the same constant. We found, however, that when we did not add this constant, the stiffness-equation method generally returned the right answer for large c' (until the point where the factorization failed to exist). Apparently the boundary conditions described in the last paragraph do not excite the unstable mode of $A^T D^{-1} A$ unless a constant is added.

We remark also that the relative differences in this example compared to the analytic solution (truncation errors) were about 0.0314, much larger than the truncation error observed with the quadratic boundary conditions described above. We believe that this is because the triangulation is not able to accurately represent the inner disk, so the c field is only zero-order accurate.

For the next test, we generated field c with two distinct large values, 10^{150} and 10^{300} . The results are illustrated in Fig. 4. To generate this mesh, we have used a two-dimension version of the algorithm in [13] implemented by S. Mitchell. This mesh generator guarantees a bound on the aspect ratio, but the triangles may be obtuse. Unlike the Delauney method described above, this method has the ability to generate triangulations with varying degrees of refinement.

The CFSM method failed for this problem. We do not know the ana-

Figure 3: The upper left figure shows the field c ; the blue region is where $c = 10^{16}$. The upper right figure shows the analytic solution. The lower left figure shows the solution computed by the NSHI method, and the lower right figure shows the solution computed by the CFSM method. Note that the CFSM solution appears to have a local maximum, namely, the large red region in the center. The color scale of the last three figures has been equalized.

Figure 4: A problem with coefficients of size 10^{150} and 10^{300} . The top figure shows the coefficient field: blue indicates 10^{300} , green indicates 10^{150} , and red indicates 1. The bottom figure shows the solution computed by the NSHI method.

lytic solution, so the NSHI solution presented in the figure can be judged only qualitatively. For instance, we would expect that in the region of conductivity 10^{300} , the temperature should be nearly constant since this region is isolated from the boundary, and the computed solution does indeed have this property. The boundary conditions are a linear function of the (x_1, x_2) coordinates.

For our last experiment on wild c 's, we used the same domain, but we let c vary from 10^{30} up to 10^{186} smoothly. In particular, c was defined to be an exponential function of the x_1 coordinate. Again, linear boundary conditions were used.

Surprisingly, the CFSM method ran to completion and returned an answer that had a relative error of only $5.6 \cdot 10^{-16}$ compared to the NSHI method. MATLAB warned that the Cholesky factors were probably singular. The CFSM method can sometimes be robust in the face of extreme ill-conditioning; we do not know why this is so.

As our final battery of tests, we compared the running time of the NSHI method to that of the standard method. It is not completely obvious how to make this comparison since the NSHI method requires combinatorial as well as numerical computation, and MATLAB is not efficient for combinatorial computation. Accordingly, we compared the two methods based only on the number of flops, as measured by MATLAB. There is a further issue that clouds the comparison. It is well-known that the number of flops for sparse factorization methods depends heavily on the ordering of the equations and the unknowns. There are very efficient orderings known for CFSM method; see [11]. For the NSHI method, on the other hand, we have not studied the ordering issue. We opted to simply use the native ordering in both algorithms, that is, the ordering produced by the Delauney mesh generator. It turns out that our implementation produces a “band ordering” for the the CFSM method, which is favorable for this method (though not optimal). We do not know whether this is a favorable ordering for the NSHI method. The sparsity pattern of the stiffness matrix K is depicted in Fig. 5. The sparsity pattern of the matrix $[A, V]$ is depicted in Fig. 6.

We generated three problems based on the same domain but with different mesh spacings. The running time figures for three problems are given in Table 2. Notice that the NSHI method took significantly longer times. The primary reason for this gap is that the system K is $n \times n$, whereas $[A, V]$ is $md \times md$, in general, much larger. We observe also from the table that

Figure 5: Nonzero structure of the stiffness matrix, $n = 1165$.

Figure 6: Nonzero structure of the matrix $[A, V]$, $md = 4984$.

the gap between the two methods appears to be widening as the problems get larger. We suspect that this is because the band ordering asymptotically improves the Cholesky running time but apparently does not help the NSHI method as much.

Table 2: Comparison of the number of flops required by the two methods.

n	md	CFSM flops	NSHI flops
73	373	18,880	952,669
292	1328	153,438	14,272,869
1165	4984	1,761,780	476,627,626

9 Conclusions and future directions

We have shown that boundary value problem (1) with wild coefficients may be solved stably by the NSHI method, even when the condition number of the stiffness matrix is well beyond the reciprocal of machine-epsilon. It would be interesting to generalize this work to other boundary value problems. Here are some examples of possible generalizations.

1. A straightforward generalization would be to use finite element basis functions of order higher than linear. We do not anticipate any difficulties in the theory, but we do not know how $\chi_{A,d}$ and $\bar{\chi}_{A,d}$ might change.
2. A more general problem is (1) with a nonzero right-hand side, i.e., $\nabla \cdot (c\nabla u) = -f$. In this case, we believe that the NSHI method and the theory developed in this paper will break down. There will be no maximum principle in this case; there is no bound on u in terms of f and g independently of c .

There is one special case of a nonzero right-hand side that could be partially handled, namely, the case that $g = 0$:

$$\begin{aligned} \nabla \cdot (c\nabla u) &= -f && \text{on } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

In this case, we believe that the RSH method (dual to the NSH method; see [22]) can accurately recover the vector $\mathbf{p}$, that is, the vector representing $c\nabla u$ (but not $\mathbf{y}$). There appears to be a maximum principle bounding $c\nabla u$ in terms of f independently of c when g is zero.

3. Another more general problem is the anisotropic version of (1), that is, $\nabla \cdot (G\nabla u) = 0$, where G is a function $\Omega \rightarrow \mathbb{R}^{d \times d}$ that is a symmetric positive definite matrix at each point.

Let $\lambda_{\max}$, $\lambda_{\min}$ denote the maximum and minimum eigenvalues of G . Thus, $\lambda_{\max}$, $\lambda_{\min}$ are also functions of the spatial coordinates. There are two different ways in which G could be wild. The first case is that $\lambda_{\max} \leq k\lambda_{\min}$ where k is a small constant at every point on the domain, but $\lambda_{\min}$ and $\lambda_{\max}$ could vary significantly as a function of the coordinates. This case might be called the “weakly” anisotropic case. We believe that the NSHI method, as described in this paper, will be adequate for this case.

The other case, the “strongly” anisotropic case, is when $\lambda_{\max}$ may be far larger than $\lambda_{\min}$ at some points, and both may vary by orders of magnitude over the domain. In this case, it seems that the plain NSH method should be used (not the NSHI method), because the isotropic versions of χ_A , $\bar{\chi}_A$ are irrelevant. The plain versions of χ_A , $\bar{\chi}_A$ for a strongly anisotropic problem will depend significantly on the orientation of the faces of the triangulation with respect to the eigenvectors of $G(x_1, \dots, x_d)$. We do not know the nature of this dependence, but it will probably be necessary for the triangulation to have very special properties in order to get reasonable bounds on χ_A , $\bar{\chi}_A$.

4. Nonlinear boundary value problems could also be considered. The most straightforward nonlinear generalization of (1) is the case that c depends on u .
5. Stokes flow could be considered. Stokes flow in a fluid with highly varying viscosity can be written:

$$\begin{aligned} -\nabla \cdot (G(x, y, z)\nabla \mathbf{v}) + \nabla p &= \mathbf{f} \\ \nabla \cdot \mathbf{v} &= 0 \end{aligned} \tag{22}$$

Strang [19] observes that this can be written in the form of (12). The difficulty is that D will be in general nondiagonal, so none of Stewart's results underlying the NSH method could be applied.

By factoring this D , say as $Q^T D' Q$ where Q is orthogonal and D' is diagonal, we may rewrite (22) with a diagonal D' by changing variables. This new diagonal matrix D' presumably would encode the variation in viscosity. The change of variables would replace the original A in (12) with QA . Then the theory can be applied. Of course, the interesting question is how to obtain bounds on $\chi_A, \bar{\chi}_A$ in this case; indeed, it is not clear even what QA represents (mathematically or physically).

By writing Stokes flow in the form (12), Wathen *et. al.* [24] have derived a method for obtaining optimal preconditioners for iterative linear solvers. This method as presented does not apply to the case when G is highly varying.

There are other future directions for this research besides widening the scope of applications. Here are some of the more interesting research questions.

- Perhaps it is possible to simplify the NSHI method, which currently involves a variety of numerical and combinatorial tasks, or to derive a completely different algorithm satisfying a bound like Theorem 5.

Indeed, the CFSM method showed surprising robustness on many test cases as mentioned in Section 8, and appears to be far more efficient. Perhaps there is some simple modification of the CFSM method that guarantees stability.

- If, on the other hand, no stable method simpler than the NSHI method is found, then the question of sparse orderings and efficiency in the NSHI method needs attention.
- A very interesting question is whether iterative methods can be applied to problems with wild coefficients. Many efficient iterative methods for boundary value problems with nonconstant coefficients have been proposed; see, e.g., Concus and Golub [9]. We do not know whether these methods are numerically stable when c is wild. Another iterative approach to problems with nonconstant coefficients is based in domain

decomposition: see Bramble, Pasciak and Schatz [7]. These methods have a convergence rate bounded independently of the variation c , but c must follow a certain pattern (it must be constant or nearly constant on substructures). Furthermore, the connection between convergence rate results (which assume exact arithmetic) and numerical stability is unclear.

- Do $\chi_{A,d}, \bar{\chi}_{A,d}$ have better bounds in the case of obtuse simplices? If $\chi_{A,d}, \bar{\chi}_{A,d}$ can indeed be exponentially large for obtuse triangulations in cases of interest, then it seems that nonobtuse triangulation algorithms must be used. Can the algorithms described in Section 7 be extended to three dimensions or higher?

10 Acknowledgements

The author benefitted greatly from discussions with James Bramble, Mark Hanisch, Alfred Schatz, and Lars Wahlbin of Cornell, Gilbert Strang of MIT, Scott Mitchell of Sandia National Laboratories, and Marshall Bern of Xerox PARC.

References

- [1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *Data Structures and Algorithms*. Addison-Wesley, 1983.
- [2] I. Babuška and J. E. Osborne. Generalized finite element methods: their performance and their relation to mixed methods. *SIAM Journal on Numerical Analysis*, 20:510–536, 1983.
- [3] B. S. Baker, E. Grosse, and C. S. Rafferty. Nonobtuse triangulation of polygons. *Discrete and Computational Geometry*, 3:147–168, 1988.
- [4] M. Bern and D. Eppstein. Mesh generation and optimal triangulation. In D. Z. Du and F. Hwang, editors, *Computing in Euclidean Geometry*. World Scientific, 1992. Also Xerox Palo Alto Research Center Technical report CSL-92-1.

- [5] M. Bern, D. Eppstein, and J. Gilbert. Provably good mesh generation. In *Proc. 31st Symp. Foundations of Computer Science*, pages 231–241, 1990.
- [6] A. Björck. Least squares methods. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume I. North-Holland, Amsterdam, 1990.
- [7] J. H. Bramble, J. E. Pasciak, and A. H. Schatz. An iterative method for elliptic problems on regions partitioned into substructures. *Mathematics of Computation*, 46:361–369, 1986.
- [8] P. G. Ciarlet. Basic error estimates for elliptic problems. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume II. North-Holland, Amsterdam, 1991.
- [9] P. Concus and G. Golub. Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations. *SIAM J. of Numerical Analysis*, 10:1103–1120, 1973.
- [10] R. S. Falk and J. E. Osborn. Remarks on mixed finite element methods for problems with rough coefficients (preprint). 1992.
- [11] A. George and J. W. H. Liu. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Englewood Cliffs, N.J., 1981.
- [12] E. Melissaratos and D. Souvaine. Coping with inconsistencies: A new approach to produce quality triangulations of polygonal domains with holes. In *Proc. 8th Symp. on Computational Geometry*, pages 202–211, 1992.
- [13] S. A. Mitchell and S. A. Vavasis. Quality mesh generation in three dimensions. In *Proc. 8th Symp. on Computational Geometry*, pages 212–221. ACM Press, 1992.
- [14] D. P. O’Leary. On bounds for scaled projections and pseudoinverses. *Linear Algebra and its Applications*, 132:115–117, 1990.
- [15] J. Ortega. *Numerical Analysis: A Second Course*. Academic Press, 1972.

- [16] M. H Protter and H. F. Weinberger. *Maximum principles in differential equations*. Springer Verlag, New York, 1984.
- [17] J. E. Roberts and J. M. Thomas. Mixed and hybrid methods. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume II. North-Holland, Amsterdam, 1991.
- [18] G. W. Stewart. On scaled projections and pseudoinverses. *Linear Algebra and its Applications*, 112:189–193, 1989.
- [19] G. Strang. *Introduction to Applied Mathematics*. Wellesley–Cambridge Press, 1986.
- [20] G. Strang. A framework for equilibrium equations. *SIAM Review*, 30:283–297, 1988.
- [21] G. Strang and G. J. Fix. *An analysis of the finite element method*. Prentice Hall, Englewood Cliffs, N.J., 1973.
- [22] S. A. Vavasis. Stable numerical algorithms for equilibrium systems. Technical Report 92-1280, Department of Computer Science, Cornell University, Ithaca, NY, 1992. To appear in *SIAM J. Matrix Analysis*.
- [23] L. Wahlbin. Private communication. 1993.
- [24] A. J. Wathen, B. Fischer, and D. J. Silvester. The convergence rate of the minimum residual method for the Stokes problem (in preparation). 1993.