UNSTRUCTURED ADDITIVE SCHWARZ - CG METHOD FOR ELLIPTIC PROBLEMS WITH HIGHLY DISCONTINUOUS COEFFICIENTS *

I.G. GRAHAM† AND M.J. HAGGER‡

Abstract. This paper concerns the iterative solution of symmetric elliptic problems with piecewise constant coefficients in two or three space dimensions discretised by linear finite element methods on unstructured triangular or tetrahedral meshes. We consider the performance of additive Schwarz domain decomposition preconditioners constructed from local solves on automatically generated subdomains together with a global solve on some coarser grid. Assuming no relationship between the coefficient function and either the subdomains or the coarse grid, we show that the preconditioned congugate gradient method converges (in both the energy and the Euclidean norms) with a number of iterations which grows only logarithmically in the size of the maximum jump in the coefficient function. This result is obtained by a careful analysis of the preconditioned matrix. It cannot be obtained by the usual procedure of estimating condition numbers: a simple example is given in which the condition number of both the original and the preconditioned stiffness matrices degrade linearly in the coefficient jump. It follows from our results and recent papers of T.F. Chan, B. Smith and J. Zou that unstructured additive Schwarz methods are both optimal with respect to the discretisation and resilient to highly discontinuous coefficients even if no attention is paid to the coefficient discontinuity in the construction of the solver.

Keywords: Domain Decomposition, Additive Schwarz, overlapping subdomains, diagonal scaling, discontinuous coefficients, conjugate gradient method.
Mathematics Subject Classification: 65F10, 65N30

1. Introduction. In this paper we consider the performance of various preconditioners for the linear finite element approximation of the model problem

$$- \nabla . a \nabla u = f \text{ on } \Omega \ , \tag{1.1}$$

$$u = g \text{ on } \partial\Omega_D \ , \qquad \frac{\partial u}{\partial n} = \tilde{g} \text{ on } \partial\Omega_N \ , \tag{1.2}$$

where $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ is a polygonal or polyhedral domain with boundary $\partial\Omega$ consisting of two pieces $\partial\Omega_N$ and $\partial\Omega_D$ ($\partial\Omega_D \neq \emptyset$) and a is a positive valued piecewise constant function on Ω.

We assume that the regions on which a is constant are given and fixed but may possibly have complicated geometry and we are concerned with the robustness of the preconditioned conjugate gradient method to both the fineness of the discretisation of the overall problem and to the severity of the discontinuities in a. This is relevant to many applications such as groundwater flow, electromagnetics and semiconductor modelling (e.g. [1, 2, 7, 8, 14, 15, 16, 17]) where the coefficients have large discontinuities across interfaces between (often complicated) regions with differing material properties.

Much work has been done on domain decomposition methods for such problems and there are many preconditioners which remove (or almost remove) ill-conditioning due to both discretisation and discontinuity - see, for example [3, 22, 10, 24, 20, 21]. More extensive surveys are in [4] and [9]. However almost all of these methods require that the

* This work was supported by EPSRC Grant GR/J88616 and British Council ARC Grant 590.

† (igg@maths.bath.ac.uk) School of Mathematical Sciences, University of Bath, Bath, BA2 7AY, U.K.

‡ (mjh@maths.bath.ac.uk) School of Mathematical Sciences, University of Bath, Bath, BA2 7AY, U.K.

mesh refinement is done in some structured (quasi-uniform or locally uniform) way and that the coarse grid or substructures resolve the discontinutiy in a. (See also [19] for a result of this type which is even independent of the aspect ratios of the substructures.)

In many practical cases an unstructured mesh is constructed using a mesh generator (perhaps together with some adaptive refinement). The discrete problem can still be preconditioned by a domain decomposition method but now the mesh is usually partitioned using some automatic load-balancing strategy and a coarse grid is constructed often by some coarsening technique. While it reasonable to assume, as we shall do in this paper, that the fine mesh provides a good approximation to the geometry of the problem, it is unreasonable to assume the same about the subdomains and/or coarse grid obtained by this automatic technique.

Recently a number of domain decomposition methods have been extended in a non-trivial way to cover the unstructured case (see, e.g., [6], [5] and the references given there). However the theory given there does not discuss the outcome when problems with very discontinuous coefficients are tackled. It is the purpose of this paper to show that *provided the conjugate gradient method is used as a solver* then the unstructured additive Schwarz preconditioners discussed above are both optimal with respect to the mesh parameter h and only very weakly dependent on the discontinuity in a even if the position of the discontinuity is ignored when the solver is constructed. In the case of unstructured meshes the meaning of "h " needs to be made precise. Here we follow [6, 5]: h is the maximum mesh diameter and we assume that the meshes remain shape regular as they are refined but are otherwise unstructured.

Our main result, obtained in Section 6, is to prove that the number of conjugate gradient iterations to obtain convergence to within a given tolerance grows only logarithmically in the coefficient discontinuity. This is in contrast to the linear growth which can (as an example in Section 3 shows) be experienced by the *condition number* of the preconditioned problem. A precise statement is given in Theorem 6.1 and the discussion following it. While Theorem 6.1 discusses convergence in the energy norm (the usual setting for conjugate gradients), it may be remarked that this setting is unsatisfactory due to the strong dependence of the energy norm on the coefficient a. Fortunately in Corollary 6.2 we are able by a simple argument to show that convergence in the Euclidean norm follows the same qualitative pattern as far as dependence on the discontinuity is concerned.

The proof of Theorem 6.1 is obtained in three stages. First it is shown in Theorem 4.4 that the eigenvalues of the additive Schwarz preconditioned matrix may be estimated below in terms of the eigenvalues of the diagonally scaled matrix. Diagonal scaling has been observed in practical computations to be very effective as a preconditioner for discontinuous problems (e.g. [12], [14, §2.6]) but no rigorous proof has ever been given of this. This lack of interest may be because diagonal scaling is of no general utility: In general it has no effect on the conditioning with respect to the mesh diameter h (see, e.g.[23]). However here we use it as a theoretical device: to estimate the properties of the additive Schwarz preconditioner as the discontinuity varies, we reduce the problem to the same question for diagonal scaling. This latter question turns out to be tractable: In Section 5 we are able to give a complete analysis of the spectral properties of the diagonally scaled matrix for a wide class of problems with discontinuous coefficients. In order to make the results easy to write down we consider sequences of discontinuous coefficients and classify the spectral properties in terms of the asymptotics of the maximum jumps of these sequences across interior interfaces. The main result of Section 5, proved in Theorem 5.5, is to show that at worst only a fixed number of eigenvalues of the diagonally scaled matrix may approach zero as the jumps worsen. This number depends on the geometry of the regions on which the coefficient is constant, but not on the mesh or the values of the coefficients. The rest of the eigenvalues are bounded above and below by positive numbers independent of the size of the jumps. Theorem 4.4 implies that the same statement is true for the additive Schwarz preconditioned matrix.

The proof of Theorem 6.1 then uses the spectral bounds established in Sections 4 and 5 together with an extension of the usual conjugate gradient convergence theory to show that, provided we discount a fixed number of initial iterations, then the overall number of iterations for convergence grows logarithmically in the smallest eigenvalue and linearly in the square root of the "reduced condition number" which neglects the bad eigenvalues. This reduced condition number can be bounded optimally with respect to h (using the results of [5]) and independently of the jumps in the coefficient by the results proved here. Corollary 6.2 bounds the number of iterations needed to achieve the same error in the Euclidean norm. Apart from the introduction of an additional term (depending logarithmically on the condition number of the discrete Laplacian on the underlying mesh) this bound is very similar to that proved in Theorem 6.1 for the energy norm. The paper concludes with a 3D numerical illustration of the theory.

We finish this section by making precise the notation which we shall need later. For problem (1.1), (1.2) we assume that $\partial\Omega_N,\ \partial\Omega_D$ are unions of segments of polygonal planar pieces in $\mathbb{R}^3$ (or straight line segments in $\mathbb{R}^2$) and that f, g and $\tilde{g}$ are suitably smooth given data. For the coefficient function a we assume that there are d open disjoint polyhedral or polygonal regions $\Lambda_k,\ k = 1,\ldots,d$, satisfying $\cup_{k=1}^{d}\bar{\Lambda}_k = \bar{\Omega}$ and with the property

$$a|_{\Omega_k} = a_k, \quad k = 1,\ldots,d\ , \tag{1.3}$$

where each a_k is positive. Then a can be identified with the vector $\mathbf{a} = (a_1,\ldots,a_d)^T \in \mathbb{R}_+^d$ where $\mathbb{R}_+ = (0,\infty)$. The value of a at the interfaces between the Λ_k is unimportant. We subdivide Ω into a mesh $\mathcal{T}$ of open triangles or tetrahedra T such that $\bar{\Omega} = \cup\{\bar{T} : T \in \mathcal{T}\}$. We assume that this "fine mesh" resolves the discontinuity in a, i.e. we assume that the boundaries $\partial\Lambda_k$ consist entirely of edges of triangles (or faces of tetrahedra). For convenience we also assume that on $\partial\Omega$ the mesh resolves the interface between $\partial\Omega_D$ and $\partial\Omega_N$.

Let $\mathcal{N}$ denote the set of all triangle or tetrahedra nodes in $\Omega \cup \partial\Omega_N$ and let n be the number of such nodes. Introduce the usual nodal basis functions $\{\phi_p : p \in \mathcal{N}\}$ of piecewise linear functions on $\mathcal{T}$ satisfying $\phi_p(q) = \delta_{pq}$ for $p, q \in \mathcal{N}$. Then the finite element method for (1.1), (1.2) yields the system

$$K(\mathbf{a})\mathbf{x} = \mathbf{b}(\mathbf{a}), \quad \text{where} \quad K(\mathbf{a})_{pq} = \int_\Omega a\nabla\phi_p\cdot\nabla\phi_q\ , \quad p, q \in \mathcal{N}. \tag{1.4}$$

$K(\mathbf{a})$ and $\mathbf{b}(\mathbf{a})$ depend implicitly on the geometry of Ω and $\mathcal{T}$ and (explicitly) on $\mathbf{a} \in \mathbb{R}_+^d$.

If $\mathcal{S}$ is any subset of $\mathcal{N}$ we write $[\mathcal{S}]$ to mean the space of vectors which have a unique real entry for each node in $\mathcal{S}$. Then (1.4) is to be solved for $\mathbf{x}$ in the space $[\mathcal{N}]$ (which is isomorphic to $\mathbb{R}^n$). The matrix $K(\mathbf{a})$ and right-hand side $\mathbf{b}(\mathbf{a})$ are assembled in accordance with some numbering of the nodes in $\mathcal{N}$ but this has no bearing on the present discussion.

2. Unstructured Additive Schwarz Methods. Here we follow [6], [5] and first partition Ω into s non-overlapping subdomains Ω_j such that $\bar{\Omega} = \cup_{j=1}^{s}\bar{\Omega}_j$. This may be done in practice by graph partitioning techniques such as [25]. Then extend each Ω_j to a domain Ω_j' in such a way that $\partial\Omega_j' \cap \Omega$ contains only triangle edges (or tetrahedral faces).

We make no assumption on the relationship between the subdomains Ω'_j and the regions Λ_k introduced in the previous section. Let the distance between $\partial\Omega_j$ and $\partial\Omega'_j$ be $\delta_j > 0$. For each j, define $\mathcal{N}(\Omega'_j) = \mathcal{N} \cap \Omega'_j$ and let $R_j^T : [\mathcal{N}(\Omega'_j)] \to [\mathcal{N}]$ be the prolongation operator which takes $\mathbf{x}_j \in [\mathcal{N}(\Omega'_j)]$ to the vector in $\mathcal{N}$ with value $\mathbf{x_j}$ on $\mathcal{N}(\Omega'_j)$ and 0 elsewhere. We can identify R_j^T with its matrix representation using standard bases in $[\mathcal{N}(\Omega'_j)]$ and $[\mathcal{N}]$ and, with this convention, the operator $R_j : [\mathcal{N}] \to [\mathcal{N}(\Omega'_j)]$ is defined to be the transpose of R_j^T and is simply the restriction from $[\mathcal{N}]$ to $[\mathcal{N}(\Omega'_j)]$. The action of $K(\mathbf{a})$ on $\mathcal{N}(\Omega'_j)$ is then defined as

$$K(\mathbf{a})_j = R_j K(\mathbf{a}) R_j^T \ . \tag{2.1}$$

In addition we introduce a coarse mesh of tetrahedra or triangles which we call $\mathcal{T}_0$. In the simplest case this is just a coarser subdivision of Ω than $\mathcal{T}$. However, the theory of [6], [5] also allows the case when $\mathcal{T}_0$ covers a region different from Ω such as may arise when coarsening techniques are applied to the original fine mesh. In this case the coarse mesh must also satisfy some technical assumptions which basically require that the region it covers should not be too much different from Ω (see (**A1**), (**A2**), (**A3**) of §2 in [5]). If $\mathcal{N}_0$ is the set of nodes of $\mathcal{T}_0$ then we introduce a linear operator $R_0^T : [\mathcal{N}_0] \to [\mathcal{N}]$ by defining $R_0^T \mathbf{x}_0$ to be the piecewise linear interpolant on $\mathcal{T}_0$ to $\mathbf{x}_0 \in \mathcal{N}_0$ evaluated at the nodes of $\mathcal{N}$. Then we set $R_0 = (R_0^T)^T$ and define

$$K(\mathbf{a})_0 = R_0 K(\mathbf{a}) R_0^T \ . \tag{2.2}$$

Remarks. If $\mathcal{T}_0$ does not cover all of Ω then R_0^T has to be modified slightly to make it well-defined - see [5]. There the authors also propose several variants of (2.2). For example, under certain technical assumptions, (2.2) may be replaced by the stiffness matrix of (1.1), (1.2) directly computed on $\mathcal{T}_0$. In addition the the interpolation process used to define R_0^T may be replaced by other projections. We mention explicitly only (2.2) because of its inherent simplicity: the matrix representation of the operator R_0^T is constructed simply by evaluating the coarse grid basis functions at the fine grid points. However the results of the present paper remain true for any of the coarse grid representations allowable by the theory in [6], [5].

Each of the matrices (2.1), (2.2) is symmetric and positive definite (since $K(\mathbf{a})$ is), and the classical additive Schwarz preconditioner for $K(\mathbf{a})$ is then defined to be

$$M(\mathbf{a})^{-1} = \sum_{j=0}^{s} R_j^T K(\mathbf{a})_j^{-1} R_j \ . \tag{2.3}$$

As is well-known, the preconditioned matrix $M(\mathbf{a})^{-1}K(\mathbf{a})$ has the representation

$$M(\mathbf{a})^{-1} K(\mathbf{a}) = \sum_{j=0}^{s} P_j(\mathbf{a}) =: P(\mathbf{a}) \tag{2.4}$$

where, for $j = 0, \ldots, s$, $P_j(\mathbf{a})$ is the orthogonal projection of $[\mathcal{N}]$ onto $\mathrm{Im}\, R_j^T \subset [\mathcal{N}]$ with respect to the energy inner product

$$(\mathbf{u}, \mathbf{v})_{K(\mathbf{a})} = \mathbf{u}^T K(\mathbf{a}) \mathbf{v} \ . \tag{2.5}$$

Using this formalism and considerably non-trivial extensions of the classical theory to cope with the unstructured meshes, Chan, Smith and Zou [5] prove the following theorem.

THEOREM 2.1. *Suppose the maximum sizes of the triangles (or tetrahedra) in $\mathcal{T}_0$ and $\mathcal{T}$ are denoted H and h respectively. Then if $\mathcal{T}$ and $\mathcal{T}_0$ are refined in a shape-regular (but not necessarily quasiuniform) way, we have a constant $C(\mathbf{a})$ independent of h and H but depending on $\mathbf{a}$, such that*

$$\kappa(M(\mathbf{a})^{-1}K(\mathbf{a})) \leq C(\mathbf{a})\left(1+\frac{H}{\delta}\right)^2$$

as $h,\ H \to 0$, where $\delta = \min_j\{\delta_j\}$ and κ denotes the condition number with respect to the spectral norm. $\square$

Thus provided the overlap of the subdomains and the diameter of the coarse mesh are kept proportional we expect no asymptotic degradation of the condition number of the preconditioned matrix as the mesh is refined.

An important special case of (2.3) arises when, for $p \in \mathcal{N}$, we let Ω_p' denote the *interior* of the set

$$\cup\{\bar{T} \in \mathcal{T} : p \in \bar{T}\}\ .$$

Then p is the only node of $\mathcal{T}$ in Ω_p' and for all $\mathbf{x} \in \mathcal{N}$, $R_p\mathbf{x} = x_p$. Hence $K(\mathbf{a})_p$ (defined by (2.1)) is simply the entry of the diagonal of $K(\mathbf{a})$ corresponding to node p. If the sum in (2.3) is taken over all nodes $p \in \mathcal{N}$ and if the coarse grid solve is suppressed then we simply have

$$M(\mathbf{a})^{-1} = (\mathrm{diag}\ K(\mathbf{a}))^{-1} \tag{2.6}$$

where diag $K(\mathbf{a})$ is the matrix which has its main diagonal coinciding with that of $K(\mathbf{a})$ and has zeros elsewhere, i.e. (2.3) reduces to classical diagonal scaling. As in (2.4) we also have

$$(\mathrm{diag}\,K(\mathbf{a}))^{-1}K(\mathbf{a}) = \sum_{p\in\mathcal{N}} Q_p(\mathbf{a}) =: Q(\mathbf{a})\ , \tag{2.7}$$

where $Q_p(\mathbf{a})$ is the orthogonal projection of $[\mathcal{N}]$ onto $\langle \mathbf{e}_p \rangle$ with respect to (2.5) (where $\mathbf{e}_p \in [\mathcal{N}]$ takes value 1 at node p and zero elsewhere).

For the reasons explained above, a detailed study of the dependence of (2.7) on a is the key method in this paper. First we give an example which points to the kind of results we should expect to be able to prove.

3. Motivating Example. Consider the problem (1.1) in the unit square $\Omega = (0,1)\times(0,1)$, with homogeneous Dirichlet boundary conditions on the bottom and top edges $\{(x,0) : x \in [0,1]\} \cup \{(x,1) : x \in [0,1]\}$ and homogeneous Neumann boundary conditions elsewhere. Let a have constant values a_1, a_2 and a_3 on the three regions $\Lambda_1, \Lambda_2, \Lambda_3$ defined as follows:

$$\Lambda_1 = (0,1/3)\times(1/3,2/3),\ \ \Lambda_2 = (2/3,1)\times(1/3,2/3),\ \ \Lambda_3 = \Omega\backslash\{\bar{\Lambda}_1 \cup \bar{\Lambda}_2\}. \tag{3.1}$$

The vector of the three values of a on these subdomains is $\mathbf{a} = (a_1,a_2,a_3)^T$. We discretise this problem with a uniform mesh with nodes

$$\mathcal{N} = \{(i/12, j/12),\ \ i = 0,...,12,\ \ j = 1,...,11\},$$

which resolves the discontinuity in a.

Four subdomains for this problem are constructed using the graph partitioning code *Jostle* [25]. This divides the graph defined by the mesh up into four roughly equal connected pieces while trying to minimise communication between the pieces. The algorithm uses only information about the fine mesh $\mathcal{N}$ and does not try in any way to construct partitions which resolve the regions Λ_k. We then extend these disjoint subdomains to overlapping ones by adding one layer of triangles in every direction (except for nodes adjoining the boundary).

In this simple case we could of course choose a coarse grid which resolves the discontinuity in a but here we want to investigate what might happen in an unstructured context. Accordingly we follow the scenario which might arise from employing some coarsening strategy and choose a coarse grid with 9 degrees of freedom at the nodes

$$\mathcal{N}_c = \{(-1/4 + 3i/4, -1/4 + 3j/4), \ \ i, j = 0, ..., 2\},$$

which covers a slighty larger domain than Ω. The coarse grid stiffness matrix is constructed as in (2.2) of §2.

We consider two additive Schwartz preconditioners with local solves on the subdomains described above and either without or with the coarse grid solve. These are denoted M_{AS} and M_{ASC} respectively. To test the resilience of the solver to discontinuities in the coefficients, we consider a sequence of coefficients $\mathbf{a} = \mathbf{a}^{(m)}$ depending on an integer parameter m given by $a_1^{(m)} = m$, $a_2^{(m)} = m$ and $a_3^{(m)} = 1$. The results in Table 1 give the condition numbers of the stiffness matrix K and the two preconditioned stiffness matrices as m varies. Clearly all three of these quantities grow linearly in m as $m \to \infty$. These results show that if the domain decomposition does not represent the discontinuity adequately then the corresponding preconditioner may have little effect on the asymptotic behaviour of the condition number as the discontinuity increases .

m	$\kappa(K)$	$\kappa(M_{\mathrm{AS}}^{-1}K)$	$\kappa(M_{\mathrm{ASC}}^{-1}K)$
1.0	1.26(2)	1.30(1)	5.55
1.0(1)	1.09(3)	4.98(1)	1.41(1)
1.0(2)	1.08(4)	4.19(2)	8.55(1)
1.0(3)	1.08(5)	4.11(3)	7.70(2)
1.0(5)	1.08(7)	4.10(5)	7.59(4)
1.0(10)	1.08(12)	4.10(10)	7.59(9)

Table 1

Condition number of the stiffness matrix and the two preconditioned stiffness matrices.

m	K	$M_{\mathrm{AS}}^{-1}K$	$M_{\mathrm{ASC}}^{-1}K$
1.0	38	14	12
1.0(1)	65	16	15
1.0(2)	106	19	17
1.0(3)	138	21	19
1.0(5)	>150	24	21
1.0(10)	>150	36	34

Table 2

Number of conjugate gradient iterations

However in Table 2 we show the result of solving the system (1.4) (with fixed randomly chosen right hand side $\mathbf{b}$) by the conjugate gradient method, first without preconditioner and then using each of the two precondioners introduced above. The starting guess for the conjugate gradient method was $\mathbf{0}$ and the iteration was terminated when the energy norm was reduced from its original value by a factor of 10^{-5}. We notice the very poor performance the unpreconditioned method but the rather good performance of each of the preconditioned methods (with or without the coarse grid solve) for fixed mesh,as m varies. This strongly suggests that both preconditioners (almost) eliminate the effect of the discontinuity. However it is clear from Table 1 that we will not be able to explain this phenomenon simply by bounding condition numbers.

4. Relation between additive Schwarz and diagonal scaling. The purpose of this section is to prove Theorem 4.4, which shows that the eigenvalues of the additive Schwarz preconditioned stiffness matrix can be bounded below in terms of the eigenvalues of the diagonally scaled stiffness matrix. To do this we need three simple results. The first two concern the eigenvalues of symmetric matrices. If A is any $n \times n$ symmetric we denote its eigenvalues by

$$\lambda_{\min}(A) =: \lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_n(A) := \lambda_{\max}(A) \ .$$

LEMMA 4.1. *If A and E are symmetric $n \times n$ matrices then*

$$\lambda_k(A) + \lambda_{\min}(E) \leq \lambda_k(A+E) \leq \lambda_k(A) + \lambda_{\max}(E), \quad k = 1, \ldots, n.$$

This is a corollary of the Courant-Fischer Minimax Theorem — see ([11, Corollary 8.1.3]). It has the following useful consequence.

COROLLARY 4.2. *Suppose A and B are symmetric $n \times n$ matrices and for some inner product $(\cdot,\cdot)$ on $\mathbb{R}^n$, suppose we have $(A\mathbf{x},\mathbf{x}) \geq (B\mathbf{x},\mathbf{x}),\ \mathbf{x} \in \mathbf{R}^n$. Then, $\lambda_k(A) \geq \lambda_k(B),\ \ k = 1, \ldots n$*

Proof. Immediate from the inequalities $\lambda_k(A) \geq \lambda_k(B) + \lambda_{\min}(A - B) \geq \lambda_k(B)$. □

Now, recalling the connection between preconditioned matrices and orthogonal projections ((2.4), (2.7)) motivates the following lemma.

LEMMA 4.3. *Let $(\cdot,\cdot)$ be an inner product on a finite dimensional vector space H. Let T and U be subspaces of H with $U \subseteq T \subseteq H$. Let P_T (respectively P_U) denote the corresponding orthogonal projection of H onto T* (respectively U). *Then*

$$(P_U\mathbf{x},\mathbf{x}) \leq (P_T\mathbf{x},\mathbf{x}), \quad \mathbf{x} \in H.$$

Proof. If $\mathbf{x} \in H$ and $\mathbf{u} \in U \subseteq T$, we have, by definition, $(P_U P_T\mathbf{x},\mathbf{u}) = (P_T\mathbf{x},\mathbf{u}) = (\mathbf{x},\mathbf{u})$. Thus $P_U P_T = P_U$. Hence, writing

$$P_T\mathbf{x} = P_U P_T\mathbf{x} + (I - P_U)P_T\mathbf{x}$$

, we have

$$\begin{aligned}(P_T\mathbf{x},\mathbf{x}) = (P_T\mathbf{x},P_T\mathbf{x}) &= (P_U P_T\mathbf{x}, P_U P_T\mathbf{x}) + ((I-P_U)P_T\mathbf{x},(I-P_U)P_T\mathbf{x}) \\ &\geq (P_U P_T\mathbf{x}, P_U P_T\mathbf{x}) = (P_U\mathbf{x},P_U\mathbf{x}) = (P_U\mathbf{x},\mathbf{x}) \ ,\end{aligned}$$

as required. □

These results lead to the following theorem.

THEOREM 4.4. *There exists constants $C_1 \geq 1$ and $C_2 > 0$ which depend only on the connectivity of the mesh such that, for all $k = 1, \ldots, n$ and all $\mathbf{a} \in \mathbb{R}^d_+$,*

$$\lambda_k\{(\mathrm{diag}K(\mathbf{a}))^{-1}K(\mathbf{a})\} \leq C_1 \lambda_k\{M(\mathbf{a})^{-1}K(\mathbf{a})\} \leq C_2. \tag{4.1}$$

Proof. By a "colouring argument" (see, e.g., [4, p.92-93]), for each $j = 1, \ldots, s$ the nodes in $\mathcal{N}(\Omega'_j) = \mathcal{N} \cap \Omega'_j$ may be partitioned into *disjoint* subsets

$$\mathcal{N}(\Omega'_j) = \mathcal{N}_j^{(1)} \cup \mathcal{N}_j^{(2)} \cup \ldots \cup \mathcal{N}_j^{(r_j)} \tag{4.2}$$

with the property that for each $r = 1, \ldots, r_j$,

$$(\mathbf{e}_p, \mathbf{e}_q)_{K(\mathbf{a})} = 0 \quad \text{when} \quad p, q \in \mathcal{N}_j^{(r)} \ .$$

Moreover,

$$r_j \leq C_1 \ , \tag{4.3}$$

where C_1 depends only on the connectivity of the mesh and not on h. (In the two-dimensional case the Four Colour Theorem tells us that C_1 can always be taken to be 4.) Then, for each $r = 1, \ldots, r_j$ and using the notation of (2.7), define the operator:

$$Q^{(r)}(\mathbf{a}) = \sum_{p \in \mathcal{N}_j^{(r)}} Q_p(\mathbf{a}) \ .$$

This is the orothogonal projection (with respect to the inner product (2.5)) onto the direct sum $\oplus_{p \in \mathcal{N}_j^{(r)}} \langle \mathbf{e}_p \rangle$, which is a subsapce of $\mathrm{Im} R_j^T$. Hence, recalling the definition of $P_j(\mathbf{a})$ in §2, and using Lemma 4.3 we have

$$(Q^{(r)}(\mathbf{a})\mathbf{x}, \mathbf{x})_{K(\mathbf{a})} \leq (P_j(\mathbf{a})\mathbf{x}, \mathbf{x})_{K(\mathbf{a})}, \qquad r = 1, \ldots, r_j \ , \ \ j = 1, \ldots, s.$$

Then it follows on using (4.3) that, for each $j = 1, \ldots s$,

$$\left(\sum_{r=1}^{r_j} Q^{(r)}(\mathbf{a})\mathbf{x}, \mathbf{x} \right)_{K(\mathbf{a})} \leq C_1 (P_j(\mathbf{a})\mathbf{x}, \mathbf{x})_{K(\mathbf{a})} \ . \tag{4.4}$$

Thus, recalling (2.4) and using the fact that $P_0(\mathbf{a})$ is positive semidefinite,

$$C_1 (P(\mathbf{a})\mathbf{x}, \mathbf{x})_{K(\mathbf{a})} = C_1 \left(\sum_{j=0}^{s} P_j(\mathbf{a})\mathbf{x}, \mathbf{x} \right)_{K(\mathbf{a})} \geq \left(\sum_{j=1}^{s} \sum_{r=1}^{r_j} \ Q^{(r)}(\mathbf{a})\mathbf{x}, \mathbf{x} \right)_{K(\mathbf{a})} \ . \tag{4.5}$$

Also, (4.2) implies that

$$\sum_{r=1}^{r_j} \ Q^{(r)}(\mathbf{a}) = \sum_{p \in \mathcal{N}(\Omega'_j)} Q_p(\mathbf{a}) \ . \tag{4.6}$$

Then, since the subdomains Ω'_j cover Ω and since orthogonal projections are positive semidefinite, (4.5) and (4.6) yield

$$\begin{aligned} C_1 (P(\mathbf{a})\mathbf{x}, \mathbf{x})_{K(\mathbf{a})} \geq \left(\sum_{j=1}^{s} \ \sum_{p \in \mathcal{N}(\Omega'_j)} \ Q_p(\mathbf{a})\mathbf{x}, \mathbf{x} \right)_{K(\mathbf{a})} \quad &\geq \quad \left(\sum_{p \in \mathcal{N}} \ Q_p(\mathbf{a})\mathbf{x}, \mathbf{x} \right)_{K(\mathbf{a})} \\ &= \quad (Q(\mathbf{a})\mathbf{x}, \mathbf{x})_{K(\mathbf{a})} \ . \end{aligned}$$

Then identifying $[\mathcal{N}]$ with $\mathbb{R}^n$ in the obvious way, and using Corollary 4.2 together with (2.4) and (2.7) yields the left-hand side of the required estimate (4.1). The right-hand side of (4.1) comes from (2.4) and a colouring argument on the subdomains and is a standard result in the domain decomposition literature, e.g. [4, Theorem 12]. $\square$

5. Diagonal Scaling. Motivated by Theorem 4.4, we study in this section the behaviour of the spectrum of the diagonally scaled matrix $(\operatorname{diag} K(\mathbf{a}))^{-1}K(\mathbf{a})$, or equivalently the spectrum of the symmetrically scaled matrix

$$S(\mathbf{a}) := (\operatorname{diag} K(\mathbf{a}))^{-\frac{1}{2}} K(\mathbf{a}) (\operatorname{diag} K(\mathbf{a}))^{-\frac{1}{2}} \tag{5.1}$$

as $\mathbf{a}$ varies in $\mathbb{R}^d_+$. To do this we need to introduce some notation. Let $D = \{1, \ldots, d\}$. Let $\triangle$ denote the Laplacian. For each $k \in D$ let N_k denote the stiffness matrix of $-\triangle$ arising from linear finite element approximation with respect to $\mathcal{T}$ on Λ_k, subject to Dirichlet boundary condition $\partial\Lambda_k \cap \partial\Omega_D$ and natural (homogeneous Neumann) condition elsewhere. Using the same notation as in §2, the degrees of freedom of N_k are

$$\mathcal{N}(\bar{\Lambda}_k) := \mathcal{N} \cap \bar{\Lambda}_k. \tag{5.2}$$

The matrices N_k can be trivially extended by zeros to matrices defined on all of $[\mathcal{N}]$. With a slight abuse of notation we also denote these by N_k and so we can write

$$K(\mathbf{a}) = \sum_{k \in D} a_k N_k \ . \tag{5.3}$$

Note that the N_k are only needed for theoretical purposes and have no connection with the local problems solved in (2.3). To examine the dependence of the (p, q) th entry of $K(\mathbf{a})$ on $\mathbf{a}$ it is convenient to introduce, for each $p \in \mathcal{N}$,

$$\mathcal{D}(p) = \{\ell : p \in \bar{\Lambda}_\ell\}. \tag{5.4}$$

Then for all $p, q \in \mathcal{N}$, using the fact that $a_l = a|_{\Lambda_l}$ is constant, we have the usual additive form of the stiffness matrix:

$$K(\mathbf{a})_{pq} = \sum_{\ell \in \mathcal{D}(p) \cap \mathcal{D}(q)} a_\ell (N_\ell)_{pq} \ . \tag{5.5}$$

This leads to a convenient representation of $S(\mathbf{a})_{pq}$:

$$\begin{aligned} S(\mathbf{a})_{pq} &= K(\mathbf{a})_{pp}^{-\frac{1}{2}} K(\mathbf{a})_{pq} K(\mathbf{a})_{qq}^{-\frac{1}{2}} \\ &= \left(\sum_{\ell \in \mathcal{D}(p)} a_\ell (N_\ell)_{pp} \right)^{-\frac{1}{2}} \left(\sum_{\ell \in \mathcal{D}(p) \cap \mathcal{D}(q)} a_k (N_\ell)_{pq} \right) \left(\sum_{\ell \in \mathcal{D}(q)} a_\ell (N_\ell)_{qq} \right)^{-\frac{1}{2}} \end{aligned} \tag{5.6}$$

Trivially, $S(\mathbf{a})_{pq} = 0$ unless p, q lie in a common triangle (tetrahedron) of $\mathcal{T}$. Moreover,

$$S(\mathbf{a})_{pp} = 1 \ , \quad p \in \mathcal{N} \ . \tag{5.7}$$

Recalling that the Λ_k are assumed open it follows that if $p, q \in \Lambda_k$ for any k, then $\mathcal{D}(p) = \mathcal{D}(q) = \{k\}$ and the expression (5.6) is then independent of $\mathbf{a}$, in fact

$$S(\mathbf{a})_{pq} = S(\mathbf{1})_{pq} \ . \tag{5.8}$$

Thus $S(\mathbf{a})_{pq}$ depends on $\mathbf{a}$ if and only if

$$(5.9)\qquad \begin{cases} p \neq q \text{ and } p, q \text{ lie in the same triangle (or tetrahedron) of } \mathcal{T} \\ \text{and} \\ \textit{either } p \in \partial\Lambda_k \textit{ or } q \in \partial\Lambda_k \text{ for some } k \ . \end{cases}$$

To analyse the dependence of $S(\mathbf{a})$ on $\mathbf{a}$ it is very useful to note that for any constant $c \in \mathbb{R}_+$ (5.6) implies

$$(5.10)\qquad S(\mathbf{a})_{pq} = S(\mathbf{a}/c)_{pq}, \quad p, q \in \mathcal{N}.$$

Choosing $c = a_k$ with k as in (5.9) and recalling (5.6) we see that $S(\mathbf{a})_{pq}$ only depends on $\mathbf{a}$ through the ratios of a_ℓ/a_k for $k, \ell \in \mathcal{D}(p) \cup \mathcal{D}(q)$. This turns out to be a key point in the later analysis. Our first result is the following.

LEMMA 5.1.
(i) $K(\mathbf{a})$, $S(\mathbf{a})$ *and their inverses are continuous in* $\mathbf{a} \in \mathbb{R}^d_+$.
(ii) For each compact subset B *of* $\mathbb{R}^d_+$ *there exists a positive constant* C_1 *independent of* $\mathbf{a}$ *but depending on* B *such that*

$$\lambda_{\min}(S(\mathbf{a})) \geq C_1 \ , \quad \mathbf{a} \in B \ .$$

(iii) There exists a positive constant C_2 *independent of* $\mathbf{a}$ *such that*

$$(5.11)\qquad \lambda_{\max}(S(\mathbf{a})) \leq C_2 \ , \quad \mathbf{a} \in \mathbb{R}^d_+ \ .$$

(iv) If $\partial\Lambda_k \cap \partial\Omega_D \neq \emptyset$ *for all* $k = 1, \dots, d$ *then there exists a positive constant* C_3 *independent of* $\mathbf{a}$ *such that*

$$(5.12)\qquad \lambda_{\min}(S(\mathbf{a})) \geq C_3 \ , \quad \mathbf{a} \in \mathbb{R}^d_+ \ .$$

The constants C_1, C_2, C_3 *in (ii), (iii), (iv) can depend on the mesh* $\mathcal{T}$.

Proof. (i) By (5.3), each element of $K(\mathbf{a})$ is continuous in $\mathbf{a} \in \mathbb{R}^d_+$ and $K(\mathbf{a})$ has a positive diagonal. Hence $\operatorname{diag} K(\mathbf{a})$, $(\operatorname{diag} K(\mathbf{a}))^{\frac{1}{2}}$ and $(\operatorname{diag} K(\mathbf{a}))^{-\frac{1}{2}}$ are all continuous in $\mathbf{a} \in \mathbb{R}^d_+$. Since $K(\mathbf{a})^{-1}$ exists for all $\mathbf{a} \in \mathbb{R}^d_+$, standard perturbation arguments show that (with any matrix norm $\|\cdot\|$),

$$\|K(\mathbf{a}')^{-1}\| \leq 2\|K(\mathbf{a})^{-1}\|$$

for all $\mathbf{a}' \in \mathbb{R}^d_+$ in some suitably small neighbourhood of $\mathbf{a} \in \mathbb{R}^d_+$. If $\mathbf{a}'$ is in this neighbourhood then

$$\begin{aligned} \|K(\mathbf{a})^{-1} - K(\mathbf{a}')^{-1}\| &= \|K(\mathbf{a})^{-1}(K(\mathbf{a}') - K(\mathbf{a}))K(\mathbf{a}')^{-1}\| \\ &\leq 2\|K(\mathbf{a})^{-1}\|^2 \|K(\mathbf{a}') - K(\mathbf{a})\| \to 0 \quad \text{as } \mathbf{a}' \to \mathbf{a} \end{aligned}$$

which shows that $K(\mathbf{a})^{-1}$ is also continuous in $\mathbf{a} \in \mathbb{R}^d_+$. This proves part (i).

(ii) It follows from part (i) that $S(\mathbf{a})^{-1}$ is bounded as a function of $\mathbf{a}$ on any compact subset $B \subset \mathbb{R}^d_+$, i.e. $\|S(\mathbf{a})^{-1}\| \leq C_1$ and (ii) follows from this on choosing the spectral norm.

(iii) Here we use a different argument. Let $\operatorname{diag} N_k$ denote the diagonal matrix containing the diagonal of N_k. Since there are a finite number of regions Λ_k there exists a constant C_2 independent of k (but depending on $\mathcal{T}$) such that

$$\lambda_{\max}((\operatorname{diag} N_k)^{-\frac{1}{2}} N_k (\operatorname{diag} N_k)^{-\frac{1}{2}}) \leq C_2 \ .$$

Hence if $\mathbf{x} \in [\mathcal{N}]$ and $\mathbf{x}_k$ denotes its restriction to $[\mathcal{N}_k]$ then

$$\mathbf{x}_k^T N_k \mathbf{x}_k \leq C_2 \mathbf{x}_k^T \ (\mathrm{diag}\, N_k)\mathbf{x}_k \ .$$

Multiplying by a_k and summing over k yields

$$\sum_k \mathbf{x}_k^T (a_k N_k)\mathbf{x}_k \leq C_2 \sum_k \mathbf{x}_k^T \ (\mathrm{diag}\, a_k N_k)\mathbf{x}_k \ ,$$

which is identical to

$$\mathbf{x}^T K(\mathbf{a})\mathbf{x} \ \leq \ C_2 \mathbf{x}^T \ (\mathrm{diag}\, K(\mathbf{a}))\mathbf{x} \ .$$

This is true for all $\mathbf{x} \in [\mathcal{N}]$ and implies (5.11).

(iv) If $\partial\Lambda_k \cap \partial\Omega_D \neq \emptyset$ for all $k = 1, \ldots d$, then each N_k is positive definite and

$$\mathbf{x}_k^T N_k \mathbf{x}_k^T \geq C_3 \mathbf{x}_k^T (\mathrm{diag} N_k)\mathbf{x}_k^T ,$$

with C_3 independent of k. Then (5.12) follows analogously to (5.11). □

Remarks.

1. The proofs of (iii) and (iv) are like the "element-by-element" preconditioning analysis (e.g. [26]).
2. It follows from (iii), (iv) that if each region Λ_k on which the coefficient a in (1.1) is constant touches the Dirichlet boundary $\partial\Omega_D$ then the condition number of the diagonally scaled stiffness matrix is bounded independently of $\mathbf{a}$. The same result is also true for any additive Schwarz preconditioned matrix by Theorem 4.4. This trivial result surprisingly does not appear to be in the literature. Unfortunately this result does not remain true for more general geometries as the following example shows.

EXAMPLE 5.1. Consider the problem (1.1), (1.2) in $\Omega = (0,1) \times (0,1) \times (0,1) \subset \mathbb{R}^3$ with homogeneous Dirichlet boundary condition on $\partial\Omega_D = \partial\Omega$, and suppose a is constant on the interior region

$$\Lambda_1 = (1/3, 2/3) \times (1/3, 2/3) \times (1/3, 2/3)$$

and on its complement $\Lambda_2 = \Omega \backslash \Lambda_1$. We discretise this problem with a uniform mesh of tetrahedra, with 125 interior degrees of freedom at the nodes

$$\mathcal{N} = \{(i/6, j/6, k/6) : i, j, k = 1, \ldots, 5\} \ .$$

Table 3 shows results for the coefficients $\mathbf{a} = \mathbf{a}^{(m)}$ given by $a_1^{(m)} = m$ and $a_2^{(m)} = 1$ as m increases. Clearly both the condition numbers $\kappa(S(\mathbf{a}))$ and $\kappa(K(\mathbf{a}))$ grow proportionately to m and hence diagonal scaling has had no effect on the asymptotic behaviour of the condition number. However, note that the spectrum of $S(\mathbf{a})$ is considerably better behaved than that of $K(\mathbf{a})$. $K(\mathbf{a})$ has 26 eigenvalues which approach ∞ and 99 which remain bounded as $m \to \infty$. In contrast $S(\mathbf{a})$ has only one eigenvalue which approaches 0 as $m \to \infty$, the rest remain bounded above and below.

□

An elementary explanation for the behaviour observed in this example can be obtained by noting that in this case (5.3) becomes simply

$$K(\mathbf{a}) = mN_1 + N_2 \ . \tag{5.13}$$

m	$\kappa(K(\mathbf{a}))$	$\lambda_1(K(\mathbf{a}))$	$\lambda_{99}(K(\mathbf{a}))$	$\lambda_{100}(K(\mathbf{a}))$	$\kappa(S(\mathbf{a}))$	$\lambda_1(S(\mathbf{a}))$	$\lambda_2(S(\mathbf{a}))$
1	4.91	1.22(-1)	4.96(-1)	4.96(-1)	4.91	2.75(-1)	5.02(-1)
10	3.79(1)	1.25(-1)	5.50(-1)	8.41(-1)	1.46(1)	9.93(-2)	5.37(-1)
10^2	3.76(2)	1.25(-1)	5.51(-1)	5.41	1.12(2)	1.34(-2)	5.05(-1)
10^3	3.76(3)	1.25(-1)	5.51(-1)	5.13(1)	1.08(3)	1.39(-3)	5.01(-1)
10^6	3.92(6)	1.20(-1)	5.51(-1)	5.09(4)	1.13(6)	1.32(-6)	5.00(-1)

TABLE 3

Condition numbers and eigenvalues of the stiffness matrix and diagonally scaled stiffness matrix.

Thus if m is large, $K(\mathbf{a})$ will have an approximate block structure corresponding to coefficient matrix mN_1 on $\bar{\Lambda}_1$, and elements of $0(1)$ elsewhere. Since $\bar{\Lambda}_1$ contains 27 degrees of freedom and N_1 has a single zero eigenvalue, we expect that 26 eigenvalues growing with m, as observed.

To explain the behaviour of $S(\mathbf{a})$, recall (5.9). Consider $S(\mathbf{a})_{pq}$ when $p \neq q$ and p, q lie in the same tetrahedron with $p \in \overline{\Lambda}_1 \cap \overline{\Lambda}_2$, i.e $\mathcal{D}(p) = \{1, 2\}$.

For q there are three cases to consider. First if $q \in \Lambda_2$ (i.e.$\mathcal{D}(q) = \{2\}$), then (5.6) yields

$$\begin{aligned} S(\mathbf{a})_{pq} &= (m(N_1)_{pp} + (N_2)_{pp})^{-\frac{1}{2}}(N_2)_{pq}(N_2)_{qq}^{-\frac{1}{2}} \\ &= O(\frac{1}{m^{\frac{1}{2}}}) \quad \text{as } m \to \infty \ . \end{aligned} \tag{5.14}$$

Secondly, if $q \in \overline{\Lambda}_1 \cap \overline{\Lambda}_2$ (i.e. $\mathcal{D}(q) = \{1, 2\}$), then we have

$$\begin{aligned} S(\mathbf{a})_{pq} &= (m(N_1)_{pp} + (N_2)_{pp})^{-\frac{1}{2}}(m(N_1)_{pq} + (N_2)_{pq})(m(N_1)_{qq} + (N_2)_{qq})^{-\frac{1}{2}} \\ &= (N_1)_{pp}^{-\frac{1}{2}}(N_1)_{pq}(N_1)_{qq}^{-\frac{1}{2}} \ + \ O\left(\frac{1}{m}\right) \ , \quad \text{as } m \to \infty \ . \end{aligned} \tag{5.15}$$

Finally if $q \in \Lambda_1$ then

$$\begin{aligned} S(\mathbf{a})_{pq} &= (m(N_1)_{pp} + (N_2)_{pp})^{-\frac{1}{2}} m^{\frac{1}{2}}(N_1)_{pq}(N_1)_{qq}^{-\frac{1}{2}} \\ &= (N_1)_{pp}^{-\frac{1}{2}}(N_1)_{pq}(N_1)_{qq}^{-\frac{1}{2}} \ + \ O\left(\frac{1}{m}\right) \ , \quad \text{as } m \to \infty \ . \end{aligned} \tag{5.16}$$

To write these calculations more succinctly, introduce the matrix $\tilde{N}_2$ obtained from N_2 by removing all rows and columns corresponding to nodes on $\overline{\Lambda}_1 \cap \overline{\Lambda}_2$. Then consider the "reduced" version of (5.16):

$$\tilde{K}(\mathbf{a}) = mN_1 + \tilde{N}_2, \tag{5.17}$$

which is block diagonal and has a single zero eigenvalue corresponding to the zero eigenvalue of N_1 (the matrix corresponding to a *pure Neumann problem* on Λ_1). If $\tilde{S}(\mathbf{a})$ denotes the diagonally scaled version of (5.17) then $\tilde{S}(\mathbf{a})$ also has a single zero eigenvalue (in fact $\tilde{S}(\mathbf{a}) = \tilde{S}(\mathbf{1})$ which is even independent of m).

We see from the above limiting arguments that $S(\mathbf{a}) - \tilde{S}(\mathbf{a})$ approaches the zero matrix as $m \to 0$. The asymptotic behaviour of the spectrum of $S(\mathbf{a})$ as $m \to \infty$ is determined by the spectrum of $\tilde{S}(\mathbf{a})$. In fact, using Lemma 4.1 together with (5.9) and (5.14) – (5.16), we have

$$\begin{aligned} 0 < \lambda_{\min}(S(\mathbf{a})) &\leq \lambda_{\min}(\tilde{S}(\mathbf{a})) + \lambda_{\max}(S(\mathbf{a}) - \tilde{S}(\mathbf{a})) \\ &= \lambda_{\max}(S(\mathbf{a}) - \tilde{S}(\mathbf{a})) = O\left(\frac{1}{m^{\frac{1}{2}}}\right) \ , \quad \text{as } m \to \infty \ . \end{aligned}$$

Moreover, there exists $C > 0$ independent of m such that, for all $k > 1$,

$$\lambda_k(S(\mathbf{a})) \geq \lambda_k(\tilde{S}(\mathbf{a})) + \lambda_{\min}(S(\mathbf{a}) - \tilde{S}(\mathbf{a})) = \lambda_k(\tilde{S}(\mathbf{a})) - O(m^{-\frac{1}{2}}) = \lambda_k(\tilde{S}(\mathbf{1})) - O(m^{-\frac{1}{2}}) \geq C$$

as $m \to \infty$. So $S(\mathbf{a})$ has a single "bad" eigenvalue which approaches 0 as $m \to \infty$, as observed in Table 3. The results above give an upper bound on $\lambda_1(S(\mathbf{a})) = \lambda_{\min}(S(\mathbf{a}))$ but they are not sharp since Table 3 shows that in fact $\lambda_1(S(\mathbf{a})) = O(m^{-1})$ as $m \to \infty$. In general for the analysis of the conjugate gradient method we need a *lower* bound on $\lambda_1(M(\mathbf{a})^{-1}K(\mathbf{a}))$. We expect to obtain this by using proving a lower bound on $\lambda_1(S(\mathbf{a}))$ and then using Theorem 4.4. The following simple argument yields lower bounds (and will be applied to the general case later). In the case (5.13) if $\mathbf{x} \in [\mathcal{N}]$ and $\mathbf{x}_k := \mathbf{x}|_{\mathcal{N}_k}$, we have

$$\mathbf{x}^T K(\mathbf{a})\mathbf{x} = m\mathbf{x}_1^T N_1 \mathbf{x}_1 + \mathbf{x}_2^T N_2 \mathbf{x}_2 \ .$$

Since each N_k is positive semi-definite, we have, for $m \geq 1$,

$$\mathbf{x}^T K(\mathbf{a})\mathbf{x} \ \geq \ \mathbf{x}^T K(\mathbf{1})\mathbf{x} \ \geq \ \frac{1}{m}\mathbf{x}^T K(\mathbf{a})\mathbf{x}$$

and similarly

$$\mathbf{x}^T \text{ diag } K(\mathbf{a})\mathbf{x} \ \geq \ \mathbf{x}^T \text{ diag } K(\mathbf{1})\mathbf{x} \ \geq \ \frac{1}{m}\mathbf{x}^T \text{ diag } K(\mathbf{a})\mathbf{x} \ .$$

Also $C_1 := \lambda_1(S(\mathbf{1})) > 0$ (independently of m), so

$$\mathbf{x}^T K(\mathbf{a})\mathbf{x} \ \geq \ \mathbf{x}^T K(\mathbf{1})\mathbf{x} \ \geq \ C_1\mathbf{x}^T \text{ diag } K(\mathbf{1})\mathbf{x} \ \geq \ (C_1/m)\mathbf{x}^T \text{ diag } K(\mathbf{a})\mathbf{x} \ , \tag{5.18}$$

which shows $\lambda_1(S(\mathbf{a})) \geq Cm^{-1}$. This bound is sharp with respect to m as can be observed in Table 3.

This example shows that the condition number of $S(\mathbf{a})$ can not in general be bounded independently of $\mathbf{a} \in \mathbb{R}_+^d$ and further that, depending on the geometry of the regions Λ_k, $\lambda_{\min}(S(\mathbf{a}))$ may be made arbitrarily close to 0 by choosing an "appropriately bad" coefficient sequence $\{\mathbf{a}^{(m)}\} \subset \mathbb{R}_+^d$. The geometrical condition on the Λ_k which allows such bad behaviour can be summarised in the following result.

Lemma 5.2.

$$\inf_{\mathbf{a} \in \mathbb{R}_+^d} \lambda_{\min}(S(\mathbf{a})) = 0 \tag{5.19}$$

if and only if $\partial\Lambda_k \cap \partial\Omega_D = \emptyset$ *for some* k.

Proof. If there exists Λ_k with $\partial\Lambda_k \cap \partial\Omega_D = \emptyset$ then by choosing $a_k = m$ and $a_\ell = 1$, for $\ell \neq k$ the analysis of Example 5.1 generalises trivially and shows that

$$C_1 m^{-1} \ \leq \ \lambda_{\min}(S(\mathbf{a})) \ \leq \ C_2 m^{-\frac{1}{2}} \to 0 \text{ as } m \to \infty \ .$$

The converse follows from Lemma 5.1 (iv). □

We shall show in Theorem 5.5 below that, under quite general conditions, the number of eigenvalues of $S(\mathbf{a})$ which may approach 0 as $\mathbf{a}$ varies corresponds to the largest number of disjoint interior "islands" which can be constructed from unions of those Λ_k which do not intersect $\partial\Omega_D$. A precise definition of this number will be given in (5.26) and (5.27) below. In fact Example 5.1 gives us a useful prototype for the general result which we shall obtain in Theorem 5.5. Rather than consider the behaviour of $S(\mathbf{a})$ for a completely general variable $\mathbf{a} \in \mathbb{R}_+^d$, we shall instead study $S(\mathbf{a}^{(m)})$ for any coefficient sequence $\{\mathbf{a}^{(m)}\}_{m=1}^{\infty} \subset \mathbb{R}_+^d$ which satisfies the following mild hypotheses. These assumptions are introduced to suppress uninteresting pathologies.

First we require $\{\mathbf{a}^{(m)}\}$ to satisfy an ordering independently of m, i.e. for all $k, \ell \in D$

$$a_k^{(1)} \geq a_\ell^{(1)} \text{ implies } a_k^{(m)} \geq a_\ell^{(m)} \text{ for all } m = 1, 2 \ldots . \tag{5.20}$$

Secondly, if

$$E := \{(k,\ell) : k \neq l \text{ and } \partial\Lambda_k \cap \partial\Lambda_\ell \neq \emptyset\} ,$$

we assume that for each $(k,\ell) \in E$, $a_k^{(m)}/a_\ell^{(m)}$ either approaches 0, or remains bounded away from 0 and ∞ or approaches ∞. Thus we introduce the sets

$$\begin{aligned}
\alpha\{\mathbf{a}^{(m)}\} &= \{(k,\ell) \in E : a_k^{(m)}/a_\ell^{(m)} \to 0 \text{ as } m \to \infty\} , \\
\beta\{\mathbf{a}^{(m)}\} &= \{(k,\ell) \in E : a_k^{(m)}/a_\ell^{(m)} \text{ lies in a compact subset of } \mathbb{R}_+\} , \\
\omega\{\mathbf{a}^{(m)}\} &= \{(k,\ell) \in E : a_k^{(m)}/a_\ell^{(m)} \to \infty \text{ as } m \to \infty\} ,
\end{aligned}$$

which we shall assume partition E. Obviously $(k,l) \in \alpha\{\mathbf{a}^{(m)}\}$ if and only if $(l,k) \in \omega\{\mathbf{a}^{(m)}\}$. Hence the spectrum of $S(\mathbf{a})$ will remain bounded above and below (by Lemma 5.1) unless both $\alpha\{\mathbf{a}^{(m)}\}$ and $\omega\{\mathbf{a}^{(m)}\}$ are non empty. Thus our precise assumption is:

$$\alpha\{\mathbf{a}^{(m)}\} \cup \beta\{\mathbf{a}^{(m)}\} \cup \omega\{\mathbf{a}^{(m)}\} = E \text{ and } \alpha\{\mathbf{a}^{(m)}\} \neq \emptyset \ . \tag{5.21}$$

The assumptions (5.20), (5.21) basically mean that we only consider coefficients which, in relation to their neighbours, can become small, large or remain bounded, and we are not interested in oscillatory patterns for the coefficients. Since (1.1), (1.2) typically model potential theory in a region Ω, with material properties characterised by $\mathbf{a}$, (5.20) and (5.21) are quite reasonable assumptions.

Definition. *The class of sequences $\{\mathbf{a}^{(m)}\}$ satisfying (5.20), (5.21) is denoted $\mathcal{C}$.* We now study the behaviour of the spectrum of $S(\mathbf{a}^{(m)})$ for $\{\mathbf{a}^{(m)}\} \in \mathcal{C}$. To do this we generalise the argument used above to explain the results observed in Example 5.1. There the coefficient has only two values in the domain and any ratio of these values either approaches 0 or ∞ as $m \to \infty$. In the general case, under the assumptions (5.20) and (5.21), more complicated behaviour will arise but the procedure of proof is analogous to that used for Example 5.1. For any $\{\mathbf{a}^{(m)}\} \in \mathcal{C}$ we define, for $k \in D$, the set

$$\tilde{\mathcal{N}}_k = \{p \in \mathcal{N}(\bar{\Lambda}_k) : (\ell,k) \notin \omega\{\mathbf{a}^{(m)}\} \text{ , for all } \ell \in \mathcal{D}(p)\backslash\{k\}\} . \tag{5.22}$$

Thus $\tilde{\mathcal{N}}_k$ contains all interior nodes on Λ_k as well as nodes on $\partial\Lambda_k$, provided they do not also lie on another $\partial\Lambda_\ell$ with $a_\ell^{(m)}/a_k^{(m)} \to \infty$ as $m \to \infty$. We then define $\tilde{N}_k$ to be the minor of N_k corresponding to rows and columns in $\tilde{\mathcal{N}}_k$. An example is in (5.17), where $\tilde{\mathcal{N}}_2$ would contain only the nodes in the *interior* of Λ_2, since $a_1^{(m)}/a_2^{(m)} \to \infty$. $\tilde{\mathcal{N}}_2$ is then decoupled from $\mathcal{N}_1$. Note that in general $\tilde{N}_k$ depends on $\{\mathbf{a}^{(m)}\}$ only through the *limiting values* of $a_\ell^{(m)}/a_k^{(m)}$ as $m \to \infty$.

As is well-known, each $\tilde{N}_k$ is positive semidefinite on $[\tilde{\mathcal{N}}_k]$. If $\partial\Omega_D \cap \partial\Lambda_k \neq \emptyset$ or if $\tilde{\mathcal{N}}_k \neq \mathcal{N}_k$ then $\tilde{N}_k$ is positive definite. Otherwise $\ker(\tilde{N}_k) = \ker(N_k) = \langle \mathbf{1}_k \rangle$ where $\mathbf{1}_k$ is the vector of ones in $[\mathcal{N}_k]$ (a nontrivial solution to the discrete homogeneous Neumann problem on $\bar{\Lambda}_k$).

Now, generalising (5.17), we can consider the "reduced" stiffness matrix

$$\tilde{K}(\mathbf{a}^{(m)}) = \sum_{k=1}^{d} a_k^{(m)} \tilde{N}_k \ . \tag{5.23}$$

Since (5.20) ensures that each $p \in \mathcal{N}$ is in at least one $\tilde{\mathcal{N}}_k$ and since the diagonal of each $\tilde{N}_k$ is strictly positive it follows that

$$\tilde{S}(\mathbf{a}^{(m)}) := (\text{diag } \tilde{K}(\mathbf{a}^{(m)}))^{-\frac{1}{2}} \tilde{K}(\mathbf{a}^{(m)}) (\text{diag } \tilde{K}(\mathbf{a}^{(m)}))^{-\frac{1}{2}} \tag{5.24}$$

is well-defined and clearly satisfies the analogue of (5.10):

$$\tilde{S}(\mathbf{a}^{(m)}) = \tilde{S}(\mathbf{a}^{(m)}/c) \ \text{, for any} \quad c \in \mathbb{R}_+. \tag{5.25}$$

In contrast to the particular case in Example 5.1, $\tilde{S}(\mathbf{a}^{(m)})$ is not necessarily independent of $\mathbf{a}^{(m)}$ because all the blocks may not decouple. But there *is* decoupling along any interface where the ratio of coefficient values becomes infinite and these are the interfaces which are of interest in the spectral analysis. We shall show in Lemma 5.4 that $S(\mathbf{a}^{(m)})$ becomes arbitrarily close to $\tilde{S}(\mathbf{a}^{(m)})$ for m sufficiently large. We expect from Example 5.1 that $\tilde{S}(\mathbf{a}^{(m)})$ may have a zero eigenvalue. Here we show this and investigate the multiplicity of this zero eigenvalue. Define the set

$$\mathcal{L} = \{k : \ker(\tilde{N}_k) = \langle \mathbf{1}_k \rangle\} \ , \tag{5.26}$$

and the integer $L \geq 0$:

$$L = \ \text{the number of components in } \cup \{\bar{\Lambda}_k : k \in \mathcal{L}\}. \tag{5.27}$$

(Here the components are the disconnected pieces of the set. There is a finite number of these which is always less than d.)

LEMMA 5.3. *For each m, $\tilde{K}(\mathbf{a}^{(m)})$ and $\tilde{S}(\mathbf{a}^{(m)})$ are symmetric positive semidefinite with a zero eigenvalue of multiplicity L. The corresponding eigenvectors of $\tilde{K}(\mathbf{a}^{(m)})$ take constant values on one or more of the components of $\cup\{\bar{\Lambda}_k : k \in \mathcal{L}\}$ and zero elsewhere.*

Proof. By (5.23), if $\mathbf{x} \in [\mathcal{N}]$ then

$$\mathbf{x}^T \tilde{K}(\mathbf{a}^{(m)}) \mathbf{x} = \sum_{k \in D} a_k^{(m)} \mathbf{x}_k^T \tilde{N}_k \mathbf{x}_k \ \geq \ 0 \ , \tag{5.28}$$

with equality if and only if

$$\mathbf{x}_k \ \begin{cases} \in \ \langle \mathbf{1}_k \rangle & , \ k \in \mathcal{L} \\ = \ \mathbf{0} & , \ k \notin \mathcal{L} \ . \end{cases} \tag{5.29}$$

Thus $\tilde{K}(\mathbf{a}^{(m)})\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x}$ satisfies (5.29). The number of linearly independent vectors satisfying (5.29) is L. This proves the result for $\tilde{K}(\mathbf{a}^{(m)})$. The statement about the eigenvalues of $\tilde{S}(\mathbf{a}^{(m)})$ follows directly from (5.24). □

Now we show that $S(\mathbf{a}^{(m)})$ and $\tilde{S}(\mathbf{a}^{(m)})$ are near each other when m becomes large.

LEMMA 5.4. *Let $\{\mathbf{a}^{(m)}\} \in \mathcal{C}$ and define*

$$\mu^{(m)} = \max\{a_k^{(m)}/a_\ell^{(m)} : (k, \ell) \in \alpha\{\mathbf{a}^{(m)}\}\} \ .$$

Then

$$\|S(\mathbf{a}^{(m)}) - \tilde{S}(\mathbf{a}^{(m)})\| = O(\mu^{(m)})^{\frac{1}{2}} \ , \quad \text{as } m \to \infty \ .$$

(The norm may be any matrix norm; in particular the spectral norm.)

Proof. The method of proof generalises the argument used to generate the estimates (5.14), (5.15) and (5.16) obtained in the particular case of Example 5.1.

Note first that $(S(\mathbf{a}^{(m)}) - \tilde{S}(\mathbf{a}^{(m)}))_{pq}$ can differ from 0 only when p and q satisfy (5.9). There are two cases to consider:

- Case A : $p, q \in \cap_{k=1}^{k'} \partial\Lambda_k$ with $k' \geq 2$.
- Case B : $p \in \partial\Lambda_k, \quad q \in \Lambda_k$.

We shall discuss only Case A; Case B is analogous but simpler. Thus we can assume that

$$\mathcal{D}(p) \cap \mathcal{D}(q) = \{1, \ldots, k'\} \tag{5.30}$$

and, by (5.6) and (5.24),

$$S(\mathbf{a}^{(m)})_{pq} = \left(\sum_{\ell \in \mathcal{D}(p)} a_\ell^{(m)} (N_\ell)_{pp}\right)^{-\frac{1}{2}} \left(\sum_{k=1}^{k'} a_k^{(m)} (N_k)_{pq}\right) \left(\sum_{\ell \in \mathcal{D}(q)} a_\ell^{(m)} (N_\ell)_{qq}\right)^{-\frac{1}{2}} , \tag{5.31}$$

$$\tilde{S}(\mathbf{a}^{(m)}) = \left(\sum_{\ell \in \mathcal{D}(p)} a_\ell^{(m)} (\tilde{N}_\ell)_{pp}\right)^{-\frac{1}{2}} \left(\sum_{k=1}^{k'} a_k^{(m)} (\tilde{N}_k)_{pq}\right) \left(\sum_{\ell \in \mathcal{D}(q)} a_\ell^{(m)} (\tilde{N}_\ell)_{qq}\right)^{-\frac{1}{2}} . \tag{5.32}$$

Note that $\mathcal{D}(p)$ and $\mathcal{D}(q)$ are supersets of $\{\mathbf{1}, \ldots, k'\}$.

First we shall prove the result in the special case when

$$(k_1, k_2) \in \beta\{\mathbf{a}^{(m)}\} \ , \quad \text{for all } k_1, k_2 \in \{1, \ldots, k'\} \ . \tag{5.33}$$

To do this we shall first write (5.32) in a different way. Suppose there exists $\ell \in \mathcal{D}(p)$ with the property that $(\ell, k) \in \alpha\{\mathbf{a}^{(m)}\}$ for some $k \in \{1, \ldots, k'\}$. Then $(k, \ell) \in \omega\{\mathbf{a}^{(m)}\}$ and so by definition (5.22), $p \notin \tilde{\mathcal{N}}_l$ which means that by definition, $(\tilde{N}_\ell)_{pp} = 0$. Thus the sums over $\ell \in \mathcal{D}(p)$ in (5.32) can be restricted to those ℓ with $(\ell, k) \notin \alpha\{\mathbf{a}^{(m)}\}$ for any $k = 1, \ldots, k'$. The sums over $\ell \in \mathcal{D}(q)$ in (5.32) can be analogously restricted.

On the other hand if $\ell \in \mathcal{D}(p) \cup \mathcal{D}(q)$ and $(\ell, k) \in \omega\{\mathbf{a}^{(m)}\}$ then by (5.33), $\ell \notin \{1, \ldots, k'\}$ and so $(\ell, k) \in \omega\{\mathbf{a}^{(m)}\}$ *for all* $k \in \{1, \ldots, k'\}$. In this case $(\tilde{N}_k)_{pq} = 0$ for all $k \in \{1, \ldots, k'\}$ and (5.32) vanishes. This motivates the introduction of the set

$$\tilde{\mathcal{D}}(p) = \{\ell \in \mathcal{D}(p) \ : (\ell, k) \in \beta\{\mathbf{a}^{(m)}\}\}$$

for any $k \in \{1, \ldots, k'\}$. Note that $\tilde{\mathcal{D}}(p)$ contains the same elements no matter what k is used on account of (5.33). Analogously introduce the set $\tilde{\mathcal{D}}(q)$. Then the above arguments show that (5.32) may be rewritten

$$\tilde{S}(\mathbf{a}^{(m)})_{pq} = \left(\sum_{\ell \in \tilde{\mathcal{D}}(p)} a_\ell^{(m)} (\tilde{N}_\ell)_{pp}\right)^{-\frac{1}{2}} \left(\sum_{k=1}^{k'} a_k^{(m)} (\tilde{N}_k)_{pq}\right) \left(\sum_{\ell \in \tilde{\mathcal{D}}(q)} a_\ell^{(m)} (\tilde{N}_\ell)_{qq}\right)^{-\frac{1}{2}} . \tag{5.34}$$

We now find an analogous way to rewrite (5.31).

First if $(\ell, k) \in \omega\{\mathbf{a}^{(m)}\}$ for any $\ell \in \mathcal{D}(p) \cup \mathcal{D}(q)$ and $k \in \{1, \ldots, k'\}$ then by (5.33) $\ell \notin \{1, \ldots k'\}$. We can use (5.10) with $c = a_k^{(m)}$ to show that

$$S(\mathbf{a}^{(m)})_{pq} = O(\mu^{(m)})^{\frac{1}{2}} = \tilde{S}(\mathbf{a}^{(m)})_{pq} + O(\mu^{(m)})^{\frac{1}{2}}, \quad \text{as } m \to \infty \ .$$

So we can assume that $(\ell, k) \notin \omega\{\mathbf{a}^{(m)}\}$ for all $\ell \in \mathcal{D}(p) \cup \mathcal{D}(q)$ and $k \in \{1, \ldots, k'\}$. By considering $(\ell, k) \in \alpha\{\mathbf{a}^{(m)}\}$ for any $\ell \in \mathcal{D}(p) \cup \mathcal{D}(q)$ and using (5.10) again with $c = a_k^{(m)}$ shows that

$$S(\mathbf{a}^{(m)})_{pq} = \left(\sum_{\ell \in \tilde{\mathcal{D}}(p)} a_\ell^{(m)} (N_\ell)_{pp} \right)^{-\frac{1}{2}} \left(\sum_{k=1}^{k'} a_k^{(m)} (N_k)_{pq} \right) \left(\sum_{\ell \in \tilde{\mathcal{D}}(q)} a_\ell^{(m)} (N_\ell)_{qq} \right)^{-\frac{1}{2}} + O(\mu^{(m)}) \ ,$$

as $m \to \infty$. Recalling the definition of $\tilde{N}_k$, we see that the first term on the right-hand side of this last expression is exactly $\tilde{S}(\mathbf{a}^{(m)})_{pq}$. This proves the result in the case when (5.33) holds.

If (5.33) is not true then there must exist at least one pair (k_1, k_2) such that

$$k_1, k_2 \in \{1, \ldots, k'\} \text{ with } (k_1, k_2) \in \omega\{\mathbf{a}^{(m)}\} \ . \tag{5.35}$$

Then by definition $(\tilde{N}_{k_2})_{pq} = 0$. Using (5.10) with $c = a_{k_1}^{(m)}$ shows that the term corresponding to $k = k_2$ can be omitted from (5.31) with an error of $O(\mu^{(m)})$. Doing this for all pairs (k_1, k_2) satisfying (5.35) brings us back to the case (5.33) which is dealt with by the procedure above. □

In the following theorem we characterise the eigenvalues of $S(\mathbf{a}^{(m)})$.

Theorem 5.5. *Let $\{\mathbf{a}^{(m)}\} \in \mathcal{C}$, let L be as in (5.27), and set*

$$\mathcal{J}^{(m)} = \max_{k, \ell \in \mathcal{D}} \{a_k^{(m)} / a_\ell^{(m)}\}.$$

(i.e. $\mathcal{J}^{(m)}$ is the maximum ratio of values of the coefficient function a). Then
(i) There exist constants C_1, C_2 independent of m such that

$$(\mathcal{J}^{(m)})^{-1} \lambda_1(S(\mathbf{1})) \leq \lambda_k \left(S(\mathbf{a}^{(m)}) \right), \quad k = 1, ..., L, \quad m \in \mathbb{N} \tag{5.36}$$

and

$$C_1 \leq \lambda_k \left(S(\mathbf{a}^{(m)}) \right) \leq C_2, \quad k = L+1, ...n, \quad m \in \mathbb{N}. \tag{5.37}$$

(ii) The estimates in (i) remain true when $S(\mathbf{a}^{(m)})$ is replaced by $M(\mathbf{a}^{(m)})^{-1} K(\mathbf{a}^{(m)})$, where $M(\mathbf{a}^{(m)})^{-1}$ is the additive Schwarz preconditioner (2.3).

Proof. (i) Analogously to the argument leading to (5.18), we have, for $\mathbf{x} \in [\mathcal{N}]$,

$$\frac{1}{\min_k a_k^{(m)}} \, \mathbf{x}^T K(\mathbf{a}^{(m)}) \mathbf{x} \ \geq \ \mathbf{x}^T K(\mathbf{1}) \mathbf{x} \ \geq \ \frac{1}{\max_k a_k^{(m)}} \, \mathbf{x}^T K(\mathbf{a}^{(m)}) \mathbf{x}$$

and the same inequality with K replaced by diag K. So

$$\mathbf{x}^T K(\mathbf{a}^{(m)}) \mathbf{x} \geq \lambda_1(S(\mathbf{1})) \frac{\min_k a_k^{(m)}}{\max_k a_k^{(m)}} \mathbf{x}^T \text{diag } K(\mathbf{a}^{(m)}) \mathbf{x},$$

which yields (5.36).

To obtain (5.37) observe that when p, q satisfy (5.9) then (5.34) and (5.25) show that $\tilde{S}(\mathbf{a}^{(m)})$ depends on $\{\mathbf{a}^{(m)}\}$ only as a continuous (even analytic) function of the parameters

$$a_k^{(m)}/a_l^{(m)}, \quad (k,l) \in \beta\{\mathbf{a}^{(m)}\}. \tag{5.38}$$

(In (5.34) this fact was established for the special case (5.33), but the remarks at the end of the proof of Lemma 5.4 show it to be true in any case.) For p, q not satisfying (5.9), $\tilde{S}(\mathbf{a}^{(m)})$ is independent of $\{\mathbf{a}^{(m)}\}$. So by standard perturbation theorems (e.g. [18, p.116]), the eigenvalues of $\tilde{S}(\mathbf{a}^{(m)})$ are continuous functions of the parameters (5.38), which, by definition of $\beta\{\mathbf{a}^{(m)}\}$, lie in a compact subset of a finite dimensional Euclidean space. Since, for each m, Lemma 5.3 shows that $\lambda_{L+1}(\tilde{S}(\mathbf{a}^{(m)})) > 0$, and a continuous function on a compact set achieves its infimum, we must also have $\lambda_{L+1}(\tilde{S}(\mathbf{a}^{(m)})) \geq C > 0$ for all m. By a similar argument $\lambda_n(\tilde{S}(\mathbf{a}^{(m)}))$ must also be bounded above independently of m. Then (5.37) follows from Lemma 4.1 and Lemma 5.4.
(ii) The right-hand side of (5.37) with $S(\mathbf{a}^{(m)})$ replaced by $M(\mathbf{a}^{(m)})^{-1}K(\mathbf{a}^{(m)})$ was proved in Theorem 4.4. In addition, the analogues of both (5.36) and the left-hand side of (5.37) are also obtained from Theorem 4.4 and part (i) of the present theorem (recalling that $S(\mathbf{a}^{(m)})$ is similar to $(\mathrm{diag}K(\mathbf{a}^{(m)}))^{-1}K(\mathbf{a}^{(m)})$). □

6. Convergence of the Conjugate Gradient method. Consider now the preconditioned conjugate gradient method for (1.4) with the additive Schwarz preconditioner $M(\mathbf{a})$ defined in (2.3). From now on $\{\mathbf{a}^{(m)}\}$ denotes any coefficient sequence from the class $\mathcal{C}$ introduced in §5, $\| \mathbf{x} \|_m = \{\mathbf{x}^T K(\mathbf{a}^{(m)})\mathbf{x}\}^{\frac{1}{2}}$ denotes the energy norm induced by $K(\mathbf{a}^{(m)})$ and we denote the eigenvalues of the preconditioned matrix by $\lambda_k^{(m)} = \lambda_k(M(\mathbf{a}^{(m)})^{-1}K(\mathbf{a}^{(m)}))$, for $k = 1, \ldots, n$.

Then it is well-known (see, for example, [13, §9.4.3]) that the conjugate gradient iterates $\mathbf{x}^j$, $j \geq 0$ satisfy

$$\| \mathbf{x} - \mathbf{x}^j \|_m \leq \| P_j(M(\mathbf{a}^{(m)})^{-1}K(\mathbf{a}^{(m)}))(\mathbf{x} - \mathbf{x}^0) \|_m,$$

for any $P_j \in \mathcal{P}_j$, the space of polynomials p of degree $\leq j$ with $p(0) = 1$. Hence, with σ denoting the spectrum of a matrix we have the bound

$$\frac{\| \mathbf{x} - \mathbf{x}^j \|_m}{\| \mathbf{x} - \mathbf{x}^0 \|_m} \leq \max\{|P_j(\lambda)| : \lambda \in \sigma(M(\mathbf{a}^{(m)})^{-1}K(\mathbf{a}^{(m)}))\}, \quad P_j \in \mathcal{P}_j \ , \tag{6.1}$$

and a bound on the error reduction is obtained by choosing P_j in such a way that the right-hand side is as small as possible. This is usually done with the help of the following standard theorem (the proof of which follows the lines of [13, Lemma 7.3.4], for example).

THEOREM. *Let T_j denote the Chebyshev polynomial of the first kind of degree j and let $0 < \underline{\lambda} < \overline{\lambda}$. The problem of minimising*

$$\max\{|q_j(\xi)| : \xi \in [\underline{\lambda}, \overline{\lambda}]\}$$

over all polynomials $q_j \in \mathcal{P}_j$ has the unique solution

$$q_j(\xi) = \frac{1}{C_j} T_j\left(\frac{\overline{\lambda} + \underline{\lambda} - 2\xi}{\overline{\lambda} - \underline{\lambda}}\right), \text{ with } C_j = T_j\left(\frac{\overline{\lambda} + \underline{\lambda}}{\overline{\lambda} - \underline{\lambda}}\right). \tag{6.2}$$

Moreover the constant C_j can be written

$$\frac{1}{C_j} = \frac{2c^j}{1 + c^{2j}}, \text{ with } c = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \text{ and } \kappa = \frac{\overline{\lambda}}{\underline{\lambda}},$$

from which it follows that (for $\kappa \geq 1+\sqrt{2}$ and $\epsilon < 2$)

$$\frac{1}{C_j} \leq \epsilon \quad \text{provided} \quad j \geq \sqrt{\kappa} \log \frac{2}{\epsilon}. \tag{6.3}$$

The usual procedure for bounding (6.1) is to apply the above theorem on the interval $[\lambda_1^{(m)}, \lambda_n^{(m)}]$ and to choose P_j in (6.1) as the resulting q_j. Since $|T_j(x)| \leq 1$ for $x \in [-1,1]$, this yields the result that

$$\frac{\| \mathbf{x} - \mathbf{x}^j \|_m}{\| \mathbf{x} - \mathbf{x}^0 \|_m} \leq \epsilon \quad \text{provided} \quad j \geq \sqrt{\kappa^{(m)}} \log \frac{2}{\epsilon}, \tag{6.4}$$

where $\kappa^{(m)} := \lambda_n^{(m)} / \lambda_1^{(m)}$. Here we know that $\kappa^{(m)}$ in general grows linearly with $\mathcal{J}^{(m)}$ and so the bound (6.4) does not explain the convergence observed in the example of §3.

A better bound can be obtained by choosing P_j in (6.1) to be a product of a polynomial of degree L which vanishes at the "bad" eigenvalues $\lambda_1^{(m)}, \ldots \lambda_L^{(m)}$ and a normalised Chebyshev polynomial on the interval $[\lambda_{L+1}^{(m)}, \lambda_n^{(m)}]$ containing the "good" eigenvalues. To state this bound, introduce the "reduced" condition number

$$\kappa_{L+1}^{(m)} = \frac{\lambda_n^{(m)}}{\lambda_{L+1}^{(m)}}.$$

THEOREM 6.1. *Let the conditions of Theorem 5.5 hold. Then there is a constant C independent of m, h, H and δ such that for each $\epsilon > 0$,*

$$\frac{\| \mathbf{x} - \mathbf{x}^j \|_m}{\| \mathbf{x} - \mathbf{x}^0 \|_m} \leq \epsilon \quad \text{provided} \quad j \geq L + \sqrt{\kappa_{L+1}^{(m)}} \left\{ \log \frac{2}{\epsilon} + L \log \frac{C}{\lambda_1^{(m)}} \right\} . \tag{6.5}$$

Remark. For fixed m, the choice $L = 0$ yields the usual bound for the number of iterates, the behavior of which, as h, H and δ vary, can be determined by Theorem 2.1. On the other hand if the mesh remains fixed and the coefficient values are allowed to vary then with L determined as in (5.27) the bounds in Theorem 5.5 show that, after an initial L iterations the number of iterations grows at worst logarithmically in $\mathcal{J}^{(m)}$.

Proof. Consider the polynomial of degree j given by

$$P_j(\xi) = \prod_{j=1}^{L} \left(\frac{\lambda_j^{(m)} - \xi}{\lambda_j^{(m)}} \right) q_{j-L}(\xi), \tag{6.6}$$

with q_{j-L} obtained by applying the previous theorem with j replaced by $j - L$ and the interval $[\underline{\lambda},\ \overline{\lambda}]$ taken to be $[\lambda_{L+1}^{(m)},\ \lambda_n^{(m)}]$. Then $P_j(0) = 1$ and $P_j(\xi)$ vanishes at $\xi = \lambda_{j'}^{(m)},\ \ j' = 1, ..., L$. Moreover, for $j' = L+1, .., n$ we have

$$|P_j(\lambda_{j'}^{(m)})| \leq \prod_{j=1}^{L} \left\{ \frac{\lambda_{j'}^{(m)} - \lambda_j^{(m)}}{\lambda_j^{(m)}} \right\} \frac{1}{C_{j-L}^{(m)}} \tag{6.7}$$

with $C_{j-L}^{(m)} := T_{j-L}((\lambda_n^{(m)} + \lambda_{j-L}^{(m)})/(\lambda_n^{(m)} - \lambda_{j-L}^{(m)}))$. Since $\lambda_n^{(m)}$ is bounded independently of h, H, δ and m, we have

$$|P_j(\lambda_{j'}^{(m)})| \leq \prod_{j=1}^{L} \left(\frac{\lambda_{j'}^{(m)}}{\lambda_j^{(m)}} \right) \frac{1}{C_{j-L}^{(m)}} \leq \left(\frac{C}{\lambda_1^{(m)}} \right)^L \frac{1}{C_{j-L}^{(m)}} \leq \epsilon \tag{6.8}$$

provided

$$\frac{1}{C_{j-L}^{(m)}} \le \epsilon \left(\frac{\lambda_1^{(m)}}{C}\right)^L \ , \tag{6.9}$$

which, using (6.3), yields the required result. $\square$

This paper has been concerned with the iterative solution of (1.1) with coefficients $a^{(m)}$ as m varies. For large m the energy norm $\|\cdot\|_m$ may not be considered the best place in which to measure the error: If $a_k^{(m)}$ becomes very small on some region Λ_k then the error $\mathbf{x}-\mathbf{x}^j$ could be large there without violating (6.5). A better norm to measure the error would be the Euclidean norm $\|\cdot\|_2$. The following simple corollary shows that we expect the same qualitative behaviour in this setting as has been shown for the energy norm.

COROLLARY 6.2. *With the conditions and notation of Theorem 6.1,*

$$\frac{\|\mathbf{x}-\mathbf{x}^j\|_m}{\|\mathbf{x}-\mathbf{x}^0\|_m} \le \epsilon \ \text{ provided } \ j \ge L + \sqrt{\kappa_R^{(m)}}\left\{\log\frac{2}{\epsilon} + L\log\frac{C}{\lambda_1^{(m)}} + \frac{1}{2}\log\kappa + \frac{1}{2}\log\mathcal{J}^{(m)}\right\} \ ,$$

where κ is the condition number of the stiffness matrix corresponding to the discretisation of (1.1) with $a = 1$ on the mesh $\mathcal{T}$.

Proof. Note that the number of iterates j in the elementary estimate (6.4) for the conjugate gradient method grows logarithmically with $2/\epsilon$. Suppose we are interested in a different norm $\|\cdot\|_*$ for which the equivalence estimates $\underline{\gamma}\|\mathbf{x}\|_* \le \|\mathbf{x}\|_m \le \overline{\gamma}\|\mathbf{x}\|_*$ hold, for all $\mathbf{x}$. Then the number of iterates to ensure a error reduction of ϵ in the norm $\|\cdot\|_*$ is easily seen to grow logarithmically in $(2\overline{\gamma})/(\epsilon\underline{\gamma})$, ie the change of norm has a weak effect on the number of iterations, once the convergence is guaranteed in the energy norm.

The proof of this corollary uses the same principle applied to (6.5). It is easily verified that

$$(\underline{\lambda}\min_k\{a_k^{(m)}\})^{\frac{1}{2}}\,\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_m \le \ (\overline{\lambda}\max_k\{a_k^{(m)}\})^{\frac{1}{2}}\,\|\mathbf{x}\|_2 \ .$$

where $\kappa = \underline{\lambda}/\overline{\lambda}$. and the result follows from Theorem 6.1 on using the principle described above. $\square$

EXAMPLE 6.1. As a final example to illustrate the last two results in the case of an unstructured grid in three dimensions we consider (1.1), (1.2) in $\Omega = (0,1)\times(0,1)\times(0,1)$. $\partial\Omega_D$ is chosen to be the intersection of $\partial\Omega$ with the planes $z=0$ and $z=1$ and

$$g(x,y,z) := \begin{cases} 1 \text{ on } z=1 \\ 0 \text{ on } z=0 \end{cases}$$

while $\partial\Omega_N = \partial\Omega\backslash\partial\Omega_D$, with $\tilde{g} = 0$. We take a to have values a_1, a_2 on each of the two regions $\Lambda_1 = (1/3,2/3)\times(1/3,2/3)\times(1/3,2/3)$ and $\Lambda_2 = \Omega\backslash\Lambda_1$. We construct a mesh for this problem which resolves the interface between $\partial\Omega_1$ and $\partial\Omega_2$ and is refined near the interface (to reflect expected irregularity in the solution there) but is otherwise unstructured. This yields a tetrahedral mesh with 4335 degrees of freedom. Figure 1 shows a two- dimensional slice through the grid, slices in the other two directions are the same.

We construct two preconditioners - a diagonal preconditioner D, and an additive Schwarz preconditioner with no coarse grid, which we denote M_{AS}. This latter preconditioner is based on a partitioning of the nodes into 6 subdomains, overlapping by one node,

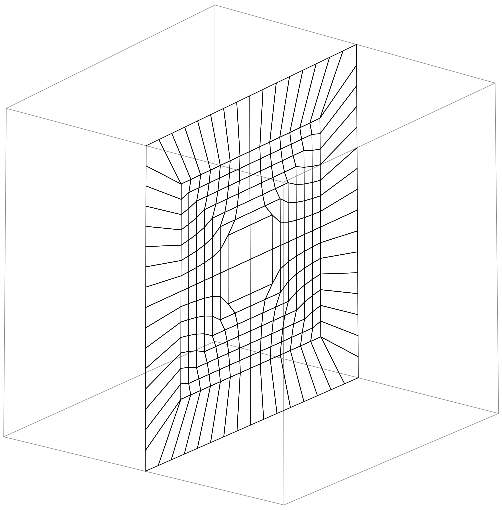

FIG. 1. *2D slice of the grid used in Example 6.1.*

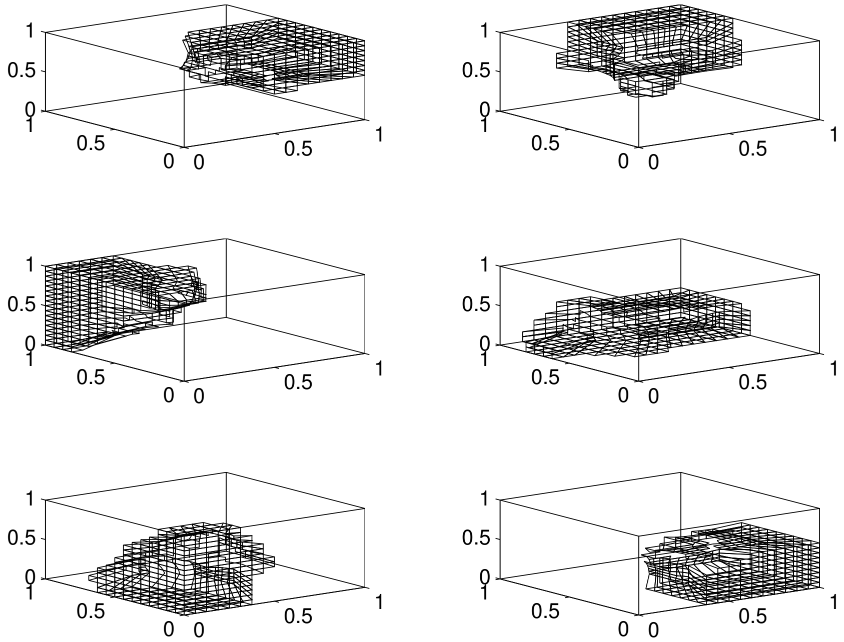

FIG. 2. *Mesh partitions produced by Jostle from the grid used in Example 6.1.*

analogously to the example in §3. These have no dependence on the position of Λ_1 and Λ_2 - see Figure 2.

Table 4 illustrates the convergence of the conjugate gradient method, first without preconditioning and then preconditioning with D and M_{AS} in the case when the coefficients are given by $a_1^{(m)} = m$ and $a_2^{(m)} = 1$. The initial guess was taken to be $\mathbf{x}^0 = \mathbf{0}$ and the iteration was continued until $\| \mathbf{x} - \mathbf{x}_k \|_m / \| \mathbf{x} \|_m < 10^{-5}$ was satisfied, where $\mathbf{x}$ denotes the true solution which was computed using an LU factorisation. In Figure 3 we show, for the two preconditioned methods, a graph of the number of iterations to convergence against the log of the coefficient. We see the linear dependence on $\log m$ and the improvement obtained by using the additive Schwarz method rather than diagonal scaling. The addition of an unstructured coarse grid solve would further improve the performance, but will not necessarily completely remove the effect of the coefficient. However this effect is now so weak as to be relatively unimportant, as the above theoretical results show. We also ran this experiment again using as stopping criterion: $\| \mathbf{x} - \mathbf{x}_k \|_2 / \| \mathbf{x} \|_2 < 10^{-5}$. We found that the numbers of iterations required differed only by one or two from those reported in Table 4. These results clearly illustrate the results of Theorem 6.1 and Corollary 6.2.

m	K	$D^{-1}K$	$M_{\mathrm{AS}}^{-1}K$
1.0	53	44	22
1.0(1)	102	48	23
1.0(2)	249	53	26
1.0(3)	618	59	30
1.0(5)	>1000	69	36
1.0(10)	>1000	94	49

TABLE 4

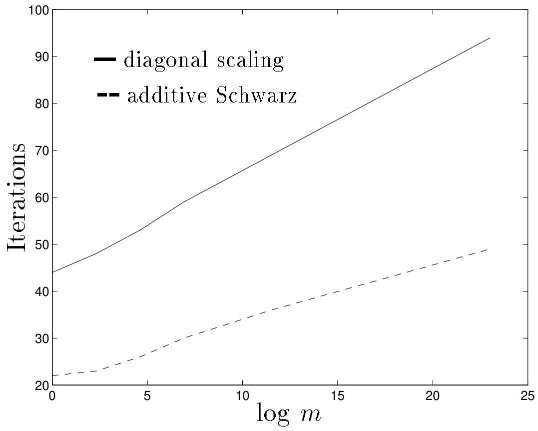

FIG. 3. .

$\square$

Acknowledgements. We thank Jens Burmeister and Wolfgang Hackbusch for helpful comments and Tarek Mathew for a useful discussion of some of the results in [4].

REFERENCES

[1] R. E. ALCOUFFE, A. BRANDT, J. E. DENDY, JR, AND J. W. PAINTER, *The multi-grid method for the diffusion equation with strongly discontinuous coeffients*, SIAM J. on Sci. Stat. Comp., 2 (1981), pp. 430–454.
[2] S. F. ASHBY, R. D. FALGOUT, S. G. SMITH, AND A. F. B. TOMPSON, *Modeling groundwater flow on MPPS*, in Proceedings of the Scalable Parallel Libraries Conference, 1993.
[3] J. H. BRAMBLE, J. E. PASCIAK, AND A. L. SCHATZ, *The construction of preconditioners for elliptic problems by substructuring*, Math. Comp., 47 (1986), pp. 103–134.
[4] T. F. CHAN AND T. P. MATHEW, *Domain decomposition algorithms*, Acta Numerica, (1994), pp. 61–143.
[5] T. F. CHAN, B. SMITH, AND J. ZOU, *Overlapping Schwarz methods on unstructured meshes using non-matching coarse grids*, Numer. Math., 73 (1996), pp. 149–167.

[6] T. F. Chan and J. Zou, *Additive Schwarz domain decomposition methods for elliptic problems on unstructured meshes*, Numerical Algorithms, 8 (1994), pp. 329–346.

[7] R. K. Coomer and I. G. Graham, *Domain decomposition methods for device modelling*, in Domain Decomposition Methods in Science and Engineering, D. Keyes and J. Xu, eds., Providence, 1995, American Mathematical Society.

[8] ———, *Massively parallel methods for semiconductor device modelling*, Computing, 56 (1996), pp. 1–28.

[9] M. Dryja, M. Sarkis, and O. Widlund, *Multilevel schwarz methods for elliptic problems with discontinuous coefficients in three dimensions*, Numer. Math., 72 (1996), pp. 313–348.

[10] M. Dryja and O. Widlund, *Some domain decomposition methods for elliptic problems*, in Iterative Methods for Large Linear Systems, Academic Press, 1989.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations (Second Edition)*, Johns Hopkins University Press, Baltimore, 1989.

[12] A. Greenbaum, C. Li, and H. Z. Chao, *Parallising PCG algorithms*, Comp. Phys. Comm., 53 (1989), pp. 295–309.

[13] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, Berlin, 1994.

[14] M. J. Hagger, *Iterative solution of large, sparse systems of equations, arising in groundwater flow models*, PhD thesis, University of Bath, 1995.

[15] B. Heise, *Parallel solvers for fem-bem equations with applications to non-linear magnetic field problems*, in Numerical Treatment of Coupled Systems, Notes on Numerical Fluid Dynamics Vol 51, W. Hackbusch and G. Wittum, eds., Braunschweig, 1995, Vieweg.

[16] B. Heise and M. Kuhn, *Parallel solves for linear and non-linear exterior magnetic field problems based on upon coupled FE/BE formulations*, tech. report, Institutsbericht Nr. 486, Johannes-Kepler-Universtat, Linz, 1995.

[17] M. C. Hill, *Solving groundwater flow problems by conjugate-gradient methods and the strongly implicit procedure*, Water Resources Research, 26 (1990), pp. 1961–1969.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1976.

[19] B. N. Khoromskij, G. E. Mazurkevich, and G. Wittum, *Robust iterative methods with highly varying coefficients in thin substructures. Applications in skin modelling*, in Proceedings of the first European Conference on Numerical Mathematics and Advanced Applications, Paris, France, 1995, GAMNI/SMAI.

[20] P. LeTallec, Y.-H. Roeck, and M.Vidrascu, *Domain decomposition methods for large linearly elliptic three-dimensional problems*, J. Comput. Appl. Math., 34 (1991).

[21] J. Mandel, *Balancing domain decomposition*, Commun. Numer. Meth. Engrg., 9 (1992), pp. 233–241.

[22] S. Nepomnyaschikh, *Domain decomposition and Schawrz methods in a subspace for the approximate solution of elliptic boundary value problems*, Computing Center of the Siberian Branch of the USSR Academy of Sciences, Novosibirsk, (1996).

[23] A. Ramage and A. J. Wathen, *On preconditioning for finite-element equations on irregular grids*, SIAM J. on Matrix Anal. and Appl., 15 (1994), pp. 909–921.

[24] B. Smith, *A domain decomposition method for elliptic problems in three dimensions*, Numer. Math., 60 (1991), pp. 219–234.

[25] C. Walshaw, M. Cross, and M. G. Everett, *A parallelisable algorithm for optimising unstructured mesh partitions*, tech. report, University of Greenwich, 1995.

[26] A. J. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.