NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2011; **18**:539–564
Published online 14 October 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/nla.741

Algebraic analysis of aggregation-based multigrid

Artem Napov*,† and Yvan Notay

Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165/84), 50, Av. F.D. Roosevelt,
B-1050 Brussels, Belgium

SUMMARY

A convergence analysis of two-grid methods based on coarsening by (unsmoothed) aggregation is presented. For diagonally dominant symmetric (M-)matrices, it is shown that the analysis can be conducted locally; that is, the convergence factor can be bounded above by computing separately for each aggregate a parameter, which in some sense measures its quality. The procedure is purely algebraic and can be used to control *a posteriori* the quality of automatic coarsening algorithms. Assuming the aggregation pattern is sufficiently regular, it is further shown that the resulting bound is asymptotically sharp for a large class of elliptic boundary value problems, including problems with variable and discontinuous coefficients. In particular, the analysis of typical examples shows that the convergence rate is insensitive to discontinuities under some reasonable assumptions on the aggregation scheme. Copyright © 2010 John Wiley & Sons, Ltd.

Received 4 December 2009; Revised 14 June 2010; Accepted 16 June 2010

KEY WORDS: multigrid; algebraic multigrid; two-grid cycle; aggregation; convergence analysis

1. INTRODUCTION

We consider multigrid methods [1–3] for solving large sparse $n\times n$ linear systems

$$A\mathbf{x}=\mathbf{b} \tag{1}$$

with symmetric positive-definite (SPD) system matrix A. Multigrid methods are based on the recursive use of a two-grid scheme. A basic two-grid method combines the action of a smoother, often a simple iterative method, and a coarse grid correction, which corresponds to the solution of the residual equations on a coarser grid. The convergence depends on the interplay between this two components and, when simple smoothers are used, it relies essentially on the *coarsening*; that is, on the way the fine grid equations are approximated by the coarse system.

Here, we consider coarsening by aggregation. In such schemes, the fine grid unknowns are grouped into disjoint sets, and each set is associated with a unique coarse grid unknown. Piecewise constant prolongation is then a common choice, which means that the solution of the residual equation computed on the coarse grid is transferred back to the fine grid by assigning the value of a given coarse variable to all fine grid variables associated with it. This makes the coarse grid matrix easy to compute and usually as sparse as the original fine grid matrix.

*Correspondence to: Artem Napov, Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165/84), 50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium.
†E-mail: anapov@ulb.ac.be

Contract/grant sponsor: The first author was supported by the Belgian FNRS (‘Aspirant’)
Contract/grant sponsor: The second author was supported by the Belgian FNRS (‘Directeur de recherches’)

Aggregation schemes are not new and trace back to [4, 5]. They did not receive much attention till recently because of the difficulty to obtain grid independent convergence on their basis [6, p. 522–524], see also [7, p. 663], where an accurate three grid analysis is presented for the model Poisson problem. This may be related to the fact that the piecewise constant prolongation does not correspond to an interpolation which is at least first-order accurate, as required by the standard multigrid theory [2, Sections 3.5 and 6.3.2].

That is why aggregation is often associated with *smoothed aggregation*, a procedure in which a tentative piecewise constant prolongation operator is smoothed [8, 9]. This allows to develop an appropriate convergence theory, but, at the same time, some of the attractive features of pure (unsmoothed) aggregation are lost. In particular, assuming the same aggregation pattern, the coarse grid matrices are less sparse and more costly to compute when using smoothed aggregation.

In this paper, we investigate such pure aggregation schemes based on the piecewise constant prolongation. They may indeed lead to two-grid methods with grid-independent convergence properties, as recently shown in [10] for model constant coefficient discrete partial differential equations (PDE) problems. There is no contradiction with the above quoted results, whose focus is on the convergence properties of two-grid methods used recursively in so-called V-cycle scheme [1]. Indeed, aggregation-based multigrid methods tend to scale poorly with the number of levels when using simple V- or even W-cycles, even though the two-grid scheme converges nicely [10, 11]. However, this may be cured using more sophisticated K-cycles, in which Krylov subspace acceleration is used at each level [12]. It is also possible to improve the scalability by increasing the number of smoothing steps on coarser levels [13].

Now, the (Fourier) analysis developed in [10] only addresses constant coefficient problems with artificial (periodic) boundary conditions. Although there are numerical evidences that aggregation-based methods can be robust in the presence of varying or discontinuous coefficients [11] (see also [14]), this yet remains to be proved. On the other hand, it is also lacking an analysis that would not only allow to assess a given aggregation scheme for a problem at hand, but could also serve as a guideline in the development of aggregation algorithms, in much the same way the coarsening strategies used in the classical AMG methods may be derived from the objective to keep reasonably bounded some convergence measure of the resulting two-grid scheme [6, 15–17].

In this paper, we fill these gaps by developing a convergence analysis that relates the global convergence to ‘local’ quantities associated with each aggregate. This analysis is based on a general algebraic result, which requires only the knowledge of a splitting of the system matrix A satisfying some given properties, and we show how this splitting can be constructed in a systematic way when the matrix is diagonally dominant. Furthermore, the needed local quantities are easy to compute solving an eigenvalue problem of the size of the aggregate. They can also be assessed analytically in a number of cases. This assessment reveals that the convergence is to a large extent insensitive to variations or discontinuities in PDE coefficients if one can introduce some reasonable assumptions on the aggregation scheme.

Moreover, as seen below, the bounds deduced in this way can often be shown asymptotically sharp provided that one assumes a simplified smoothing scheme with only one damped Jacobi pre- or post-smoothing step. Hence, we not only develop a qualitative analysis, but also a quantitative one, complementary to the Fourier analysis: this latter allows to assess the benefit of more smoothing steps or increasing smoother quality, but is restricted to constant coefficient problems on rectangular grids.

Returning to a qualitative viewpoint, it should be mentioned that, since the bound depends only on local quantities, it is independent of the global properties of the underlying PDE such as (full) elliptic regularity. For instance, estimates derived in Section 4 do not need the assumption that the underlying domain is convex, and, in fact, allows re-entering corners.

The presented results share some features with the analysis of element-based algebraic multigrid (AMGe) approaches, as developed in [18–21]. Convergence estimates presented there are also local and can be used to guide the coarsening process. The AMGe coarsening itself, however, differs substantially from aggregation. It applies only to finite element problems and requires the knowledge of element matrices, whereas the associated prolongation is denser than the basic piecewise constant prolongation considered here.

DOI: 10.1002/nla

The remainder of this paper is organized as follows. The general framework of aggregation-based two-grid methods is introduced in Section 2. The algebraic analysis is developed in Section 3, and illustrated in Sections 4 and 5 on PDE problems with, respectively, continuous and discontinuous coefficients. Concluding remarks are given in Section 6.

Notation

For any set Γ, $|\Gamma|$ is its size. For any matrix B, $\mathscr{R}(B)$ is its range and $\mathscr{N}(B)$ is its null space. For any square matrix C, $\sigma(C)$ is its spectrum and $\rho(C)$ is its spectral radius (that is, its largest eigenvalue in modulus). I stands for identity matrix.

2. AGGREGATION-BASED TWO-GRID SCHEMES

The coarsening procedure is based on the agglomeration of the unknowns of the system (1) into n_c non-empty disjoint sets called *aggregates*. The size of kth aggregate is denoted by $n^{(k)}>0$. Note that some aggregation procedures (e.g. [11]) leave part of the unknowns outside the coarsening process, for instance because the corresponding row is strongly dominated by its diagonal element. As will be seen below, our analysis gives a theoretical support to this approach. Therefore, besides the n_c regular aggregates we define the (pseudo) 0th aggregate as the (possibly empty) set of $n^{(0)}$ unknowns that are left outside the coarsening process. For the ease of presentation, and without loss of generality, we assume the ordering of the unknowns such that those belonging to $(k+1)$th aggregate have higher indices that those belonging to kth aggregate, $k=0,\ldots,n_c-1$.

The regular aggregates are the variables of the next (coarse) level in the multigrid hierarchy. Once they are defined, the $n\times n_c$ prolongation matrix is given by

$$(P)_{ij}=\begin{cases}1 & \text{if } i \text{ belongs to } j\text{th aggregate}, \quad j=1,\ldots,n_c,\\ 0 & \text{otherwise.}\end{cases} \tag{2}$$

Hence, setting $\mathbf{1}_m=(1,1,\ldots,1)^{\mathrm{T}}$, with m being the vector size, we have

$$P=\begin{pmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{1}_{n^{(1)}} & & & \\ & \mathbf{1}_{n^{(2)}} & & \\ & & \ddots & \\ & & & \mathbf{1}_{n^{(n_c)}} \end{pmatrix}. \tag{3}$$

In what follows, we assume a slightly more general form of (3)

$$P=\begin{pmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{p}^{(1)} & & & \\ & \mathbf{p}^{(2)} & & \\ & & \ddots & \\ & & & \mathbf{p}^{(n_c)} \end{pmatrix} \tag{4}$$

with $\mathbf{p}^{(k)}$ being a vector of size $n^{(k)}$. We shall see, however, that for the considered examples the choice $\mathbf{p}^{(k)}=\mathbf{1}_{n^{(k)}}$ is at the same time simple and natural.

Once the prolongation P is known, the $n_c\times n$ restriction matrix is set to its transpose and the $n_c\times n_c$ coarse grid matrix is given by the Galerkin formula $A_c=P^{\mathrm{T}}AP$. In order to complete the definition of a two-grid scheme, one also needs to specify the pre- and post-smoother matrices M_1, M_2, as well as the numbers ν_1 and ν_2 of pre- and post-smoothing steps, respectively. The iteration matrix E_{TG} of the two-grid cycle is then given by

$$E_{\mathrm{TG}}=(I-M_2^{-1}A)^{\nu_2}(I-P^{\mathrm{T}}A_c^{-1}PA)(I-M_1^{-1}A)^{\nu_1}. \tag{5}$$

The main objective of this paper is the analysis of its spectral radius $\rho(E_{\mathrm{TG}})$ (that is, its largest eigenvalue in modulus), which governs the convergence of the two-grid scheme.

It is often convenient to define a 'global' smoother X via the relation

$$I-X^{-1}A=(I-M_1^{-1}A)^{\nu_1}(I-M_2^{-1}A)^{\nu_2}. \tag{6}$$

X has the same effect in one iteration as ν_2 steps of post-smoothing followed by ν_1 steps of pre-smoothing. In what follows, we assume that X is SPD. In particular, this is the case when either

(i) $M_1=M_2=M$ is symmetric and satisfy $\rho(I-M^{-1}A)<1$ (as, e.g. in the case of weighted Jacobi); or
(ii) $M_1=M_2^{\mathrm{T}}=M$, $\nu_1=\nu_2$ and there holds $\rho((I-M^{-1}A)(I-M^{-\mathrm{T}}A))<1$ (as, e.g. in the case of symmetric Gauss–Seidel).

3. ALGEBRAIC ANALYSIS

The starting point of our analysis is a well-known identity for the two-grid convergence rate introduced in [22, Theorem 4.3] (see also [23, Theorem 3.19]). We recall it up to a slight generalization in Theorem 3.1 below. The generalization, that is based on the results in [24], allows to consider the above (i) smoothing scheme with $\nu_1\neq\nu_2$; in particular, $\nu_1=1$ and $\nu_2=0$ is allowed. This latter case is somehow important because the parameter μ_D for $D=\mathrm{diag}(A)$, which is investigated in the remainder of this paper, appears then directly connected to the convergence factor of a simplified two-grid scheme with only 1 pre- or post-smoothing step.

Theorem 3.1
Let A be an $n\times n$ SPD matrix and let P be an $n\times n_c$ matrix of rank $n_c<n$. Let M_1, ν_1 and M_2, ν_2 be such that X, defined by (6), is an $n\times n$ SPD matrix and let E_{TG} be the two-grid iteration matrix defined by (5).

Then, setting $\pi_X=P(P^{\mathrm{T}}XP)^{-1}P^{\mathrm{T}}X$, we have

$$\rho(E_{\mathrm{TG}})\leqslant\max\left(\lambda_{\max}(X^{-1}A)-1,1-\frac{1}{\mu_X}\right), \tag{7}$$

where

$$\mu_X=\max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}}\frac{\mathbf{v}^{\mathrm{T}}X(I-\pi_X)\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A\mathbf{v}}.$$

Moreover, for any $n\times n$ SPD matrix D, setting $\pi_D=P(P^{\mathrm{T}}DP)^{-1}P^{\mathrm{T}}D$ and

$$\mu_D=\max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}}\frac{\mathbf{v}^{\mathrm{T}}D(I-\pi_D)\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A\mathbf{v}}$$

there holds

$$\mu_X\leqslant\left(\max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}}\frac{\mathbf{v}^{\mathrm{T}}X\mathbf{v}}{\mathbf{v}^{\mathrm{T}}D\mathbf{v}}\right)\mu_D. \tag{8}$$

In particular, if $M_1=M_2=\omega^{-1}D$ with $\omega^{-1}\geqslant\lambda_{\max}(D^{-1}A)$, one has

$$\rho(E_{\mathrm{TG}})=1-\frac{1}{\mu_X} \tag{9}$$

with

$$\mu_X \leqslant \omega^{-1}\mu_D, \tag{10}$$

where an equality is reached when $\nu_1+\nu_2=1$.

Proof
The inequality (7) is a direct consequence of [24, Theorem 2.1 and Corollary 2.1], combined with the assumptions that A and X are SPD, which implies

$$\max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}}X(I-\pi_X)\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A\mathbf{v}} = \lambda_{\max}(A^{-1/2}X(I-\pi_X)A^{-1/2}) = \lambda_{\max}(A^{-1}X(I-\pi_X)).$$

The inequality (8) follows from Corollary 3.20 in [23] (or, alternatively, from Corollary 2.2 in [24], setting in this latter $Y=D$, $L_Y=D^{1/2}$ and $Q=\pi_D$).

To prove (9), observe that $\omega^{-1}\geqslant\lambda_{\max}(D^{-1}A)$ implies, together with (6), $\lambda_{\max}(X^{-1}A)\leqslant 1$. Hence, since it is known by [24, Theorem 2.1] that $\mu_X\geqslant 1$, the second term in (7) is larger than the first one. It follows then from the proof of [24, Corollary 2.2] that inequality (7) becomes an equality, hence (9).

The inequality (10) follows from (8) combined with

$$\begin{aligned}\max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}}X\mathbf{v}}{\mathbf{v}^{\mathrm{T}}D\mathbf{v}} &= \omega^{-1} \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}}\omega D^{-1}\mathbf{v}}{\mathbf{v}^{\mathrm{T}}X^{-1}\mathbf{v}}\\ &= \omega^{-1} \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}}\mathbf{v}-\mathbf{v}^{\mathrm{T}}(I-\omega A^{1/2}D^{-1}A^{1/2})\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\mathbf{v}-\mathbf{v}^{\mathrm{T}}(I-A^{1/2}X^{-1}A^{1/2})\mathbf{v}} \leqslant \omega^{-1},\end{aligned}$$

where the last inequality holds because $I-A^{1/2}X^{-1}A^{1/2}=(I-\omega A^{1/2}D^{-1}A^{1/2})^{\nu_1+\nu_2}$. Eventually, when $\nu_1+\nu_2=1$, one has $X=\omega^{-1}D$, which implies $X(I-\pi_X)=\omega^{-1}D(I-\pi_D)$, and, hence, that (10) is an equality. □

When D is chosen independently of P, the first factor in the right-hand side of (8) depends only on the smoothing scheme. If $M_1=M_2^{\mathrm{T}}=M$ and $\nu_1=\nu_2$, setting $S=I-M^{-1}A$, one has further

$$\frac{\mathbf{v}^{\mathrm{T}}X\mathbf{v}}{\mathbf{v}^{\mathrm{T}}D\mathbf{v}} \leqslant \sigma^{-1} \quad \forall\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\} \Longleftrightarrow \|S\mathbf{v}\|_A^2 \leqslant \|\mathbf{v}\|_A^2 - \sigma\|\mathbf{v}\|_{AD^{-1}A} \qquad \forall\mathbf{v}\in\mathbb{R}^n.$$

Hence, when $D=\mathrm{diag}(A)$ (the choice that is privileged in the rest of this work) the value of σ is nothing but of the *smoothing factor* in the Ruge–Stüben analysis [6]. On the other hand, the second factor in the right-hand side of (8) depends on P but not on X, and keeping it bounded amounts to satisfy an *approximation property*.

Now, our analysis is based on the splitting of A as

$$A = A_b + A_r, \tag{11}$$

where A_b and A_r are both symmetric non-negative definite and A_b is block diagonal

$$A_b = \begin{pmatrix} A^{(0)} & & & \\ & A^{(1)} & & \\ & & \ddots & \\ & & & A^{(n_c)} \end{pmatrix}, \tag{12}$$

where $A^{(k)}$, $k=0,\ldots,n_c$, is of size $n^{(k)}\times n^{(k)}$.

As an example, consider a symmetric diagonally dominant matrix A with positive diagonal entries (in particular, if all off-diagonal entries are non-positive, the matrix is an M-matrix).

The matrices $A^{(k)}$, $k=0,\dots,n_c$ can be constructed by restricting the matrix A to the unknowns belonging to the kth aggregate and then by subtracting the corresponding contribution $C^{(k)}=\operatorname{diag}(c_i)$ from its diagonal, in order to keep

$$A_r=\begin{pmatrix} C^{(0)} & * & \cdots & * \\ * & C^{(1)} & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & C^{(n_c)} \end{pmatrix} \tag{13}$$

diagonally dominant, and, hence, non-negative definite. Since A is diagonally dominant, any contribution subtracted from the diagonal of each $A^{(k)}$ that allows to satisfy

$$|(A)_{jj}|-\sum_{\substack{i=1\\ i\neq j}}^{n}|(A)_{ij}|\geqslant(A_b)_{jj}-\sum_{\substack{i=1\\ i\neq j}}^{n}|(A_b)_{ij}|\geqslant 0, \quad j=1,\dots,n \tag{14}$$

lead to a possible splitting. The case when the upper inequality becomes an equality; that is, when

$$(A_r)_{jj}-\sum_{\substack{i=1\\ i\neq j}}^{n}|(A_r)_{ij}|=0, \quad j=1,\dots,n \tag{15}$$

is of particular interest below.

Another splitting (11) of A can be constructed in the finite element context when each aggregate contains all and only the nodes belonging to one element or to the union of several elements. The matrices $A^{(k)}$, $k=0,\dots,n_c$, are then assembled from corresponding element matrices, whereas the assembly of remaining element matrices gives the matrix A_r. The non-negative definitness of A_b and A_r then follows from non-negative definiteness of the element matrices.

Once the splitting is known, the following theorem allows to estimate the 'global' approximation property constant μ_D by means of 'local' quantities $\mu_D^{(k)}$, $k=0,\dots,n_c$. Because each $\mu_D^{(k)}$ corresponds to a particular aggregate k, it may be seen as a measure of this aggregate's quality.

Theorem 3.2

Let $A=A_b+A_r$ be an $n\times n$ SPD matrix, with A_b and A_r symmetric non-negative definite and A_b having the block-diagonal form (12). Let P be an $n\times n_c$ matrix of rank $n_c<n$ and of the form (4). Let

$$D=\begin{pmatrix} D^{(0)} & & & \\ & D^{(1)} & & \\ & & \ddots & \\ & & & D^{(n_c)} \end{pmatrix} \tag{16}$$

be an $n\times n$ SPD matrix, let $\pi_D=P(P^{\mathrm{T}}DP)^{-1}P^{\mathrm{T}}D$ and set

$$\mu_D=\max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}}\frac{\mathbf{v}^{\mathrm{T}}D(I-\pi_D)\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A\mathbf{v}}. \tag{17}$$

Letting

$$\mu_D^{(0)}=\begin{cases} 0 & \text{if } n^{(0)}=0,\\ \sup\limits_{\mathbf{v}^{(0)}\in\mathbb{R}^{n^{(0)}}\backslash\mathcal{N}(A^{(0)})}\dfrac{\mathbf{v}^{(0)\mathrm{T}}D^{(0)}\mathbf{v}^{(0)}}{\mathbf{v}^{(0)\mathrm{T}}A^{(0)}\mathbf{v}^{(0)}} & \text{if } n^{(0)}>0\end{cases}$$

DOI: 10.1002/nla

and, for $k=1,\dots,n_c$,

$$\mu_D^{(k)}=\begin{cases}0 & \text{if } n^{(k)}=1,\\ \displaystyle\sup_{\mathbf{v}^{(\mathbf{k})}\in\mathbb{R}^{n^{(k)}}\backslash\mathcal{N}(A^{(k)})}\frac{\mathbf{v}^{(k)\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}^{(k)}}{\mathbf{v}^{(k)\mathrm{T}}A^{(k)}\mathbf{v}^{(k)}} & \text{if } n^{(k)}>1,\end{cases} \tag{18}$$

where

$$\pi_D^{(k)}=\mathbf{p}^{(k)}(\mathbf{p}^{(k)\mathrm{T}}D^{(k)}\mathbf{p}^{(k)})^{-1}\mathbf{p}^{(k)\mathrm{T}}D^{(k)} \tag{19}$$

there holds

$$\mu_D\leqslant\max_{k=0,\dots,n_c}\mu_D^{(k)}. \tag{20}$$

Moreover, $\mu_D^{(0)}<\infty$ if and only if $n^{(0)}=0$ or $A^{(0)}$ is SPD, and, for $k=1,\dots,n_c$, $\mu_D^{(k)}<\infty$ if and only if $\mathcal{N}(A^{(k)})\subset\operatorname{span}\{\mathbf{p}^{(k)}\}$, with, in the latter case,

$$\mu_D^{(k)}=\begin{cases}0 & \text{if } n^{(k)}=1,\\ \displaystyle\max_{\mathbf{v}^{(\mathbf{k})}\in\mathcal{R}(A^{(k)})\backslash\{\mathbf{0}\}}\frac{\mathbf{v}^{(k)\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}^{(k)}}{\mathbf{v}^{(k)\mathrm{T}}A^{(k)}\mathbf{v}^{(k)}} & \text{if } n^{(k)}>1.\end{cases} \tag{21}$$

Proof

We first prove the *if and only if* result for $k=1,\dots,n_c$, the case $k=0$ being trivial. The *if* statement assumes $\mathcal{N}(A^{(k)})\subset\operatorname{span}\{\mathbf{p}^{(\mathbf{k})}\}$ which means that either $\mathcal{N}(A^{(k)})=\{\mathbf{0}\}$ or $\mathcal{N}(A^{(k)})=\operatorname{span}\{\mathbf{p}^{(k)}\}$. In the former case the supremum in (18) becomes a maximum over $\mathbb{R}^{n^{(k)}}\backslash\{\mathbf{0}\}=\mathcal{R}(A^{(k)})\backslash\{\mathbf{0}\}$, hence, (21) and $\mu_D^{(k)}<\infty$. In the latter case, decomposing any vector that does not belong to $\mathcal{N}(A^{(k)})$ as $\mathbf{v}=\alpha\mathbf{p}^{(k)}+\mathbf{w}$, $\mathbf{w}\in\mathcal{R}(A^{(k)})\backslash\{\mathbf{0}\}$, and using $D^{(k)}(I-\pi_D^{(k)})\mathbf{p}^{(k)}=(D^{(k)}(I-\pi_D^{(k)}))^{\mathrm{T}}\mathbf{p}^{(k)}=\mathbf{0}$, we have

$$\mu_D^{(k)}=\sup_{\mathbf{v}\in\mathbb{R}^{n^{(k)}}\backslash\mathcal{N}(A^{(k)})}\frac{\mathbf{v}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A^{(k)}\mathbf{v}}=\max_{\mathbf{w}\in\mathcal{R}(A^{(k)})\backslash\{\mathbf{0}\}}\frac{\mathbf{w}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{w}}{\mathbf{w}^{\mathrm{T}}A^{(k)}\mathbf{w}}$$

leading to the same conclusions. The *only if* statement is proved assuming $\mathcal{N}(A^{(k)})\not\subseteq\operatorname{span}\{\mathbf{p}^{(k)}\}$ and showing that $\mu_D^{(k)}=\infty$. Indeed, taking $\mathbf{v}=\alpha\mathbf{u}+\mathbf{w}$ with $\mathbf{w}\in\mathcal{N}(A^{(k)})\backslash\operatorname{span}\{\mathbf{p}^{(k)}\}$ (exists by assumption) and $\mathbf{u}\in\mathcal{R}(A^{(k)})$ leads to

$$\mu_D^{(k)}=\sup_{\alpha\in\mathbb{R}\backslash\{0\}}\frac{|\mathbf{w}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{w}+2\alpha\mathbf{u}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{w}+\alpha^2\mathbf{u}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{u}|}{\alpha^2\mathbf{u}^{\mathrm{T}}A^{(k)}\mathbf{u}}.$$

Since $\mathbf{w}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{w}\neq 0$ by construction of $\mathbf{w}$, this last expression is unbounded for $\alpha\to 0$.

We now prove (20). Note that this inequality is obvious when $\mu_D^{(k)}=\infty$ for at least one k. Hence, without loss of generality we may assume $\mu_D^{(k)}$ finite for $k=0,\dots,n_c$. Moreover, since $n_c<n$, there holds $\mu_D>0$.

Now, observe that

$$D(I-\pi_D)=\begin{pmatrix}D^{(0)} & & & \\ & D^{(1)}(I-\pi_D^{(1)}) & & \\ & & \ddots & \\ & & & D^{(n_c)}(I-\pi_D^{(n_c)})\end{pmatrix} \tag{22}$$

DOI: 10.1002/nla

and, hence,

$$\begin{aligned}\mu_D &= \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}} D(I-\pi_D)\mathbf{v}}{\mathbf{v}^{\mathrm{T}} A\mathbf{v}} \\ &= \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}} D(I-\pi_D)\mathbf{v}}{\mathbf{v}^{\mathrm{T}} A_b\mathbf{v}+\mathbf{v}^{\mathrm{T}} A_r\mathbf{v}} \\ &= \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\sum_{k=1,\dots,n_c} \mathbf{v}^{(k)\mathrm{T}} D^{(k)}(I-\pi_D^{(k)})\mathbf{v}^{(k)}+\mathbf{v}^{(0)\mathrm{T}} D^{(0)}\mathbf{v}^{(0)}}{\sum_{k=0,\dots,n_c} \mathbf{v}^{(k)\mathrm{T}} A^{(k)}\mathbf{v}^{(k)}+\mathbf{v}^{\mathrm{T}} A_r\mathbf{v}}. \end{aligned} \tag{23}$$

Let $\mathbf{v}_* = (\mathbf{v}_*^{(0)\mathrm{T}}\mathbf{v}_*^{(1)\mathrm{T}}\dots\mathbf{v}_*^{(n_c)\mathrm{T}})^{\mathrm{T}}$ be the vector that realizes this maximum. Notice that $\sum_{k=0,\dots,n_c} \mathbf{v}_*^{(k)\mathrm{T}} A^{(k)}\mathbf{v}_*^{(k)}>0$. Indeed, because of the boundness of $\mu_D^{(k)}$, $k=0,\dots,n_c$, the equality $\sum_{k=0,\dots,n_c} \mathbf{v}_*^{(k)\mathrm{T}} A^{(k)}\mathbf{v}_*^{(k)}=0$ would imply a zero numerator in the right-hand side of (23), whereas, since A is SPD, $\mathbf{v}_*^{\mathrm{T}} A_r\mathbf{v}_*>0$, which would further lead to $\mu_D=0$. This latter contradicts our assumption $n_c<n$.

Next, since by assumption $\mu_D^{(k)}$, $k=1,\dots,n_c$, is finite, $\mathbf{v}_*^{(k)}\in\mathcal{N}(A^{(k)})$ implies

$$\mathbf{v}_*^{(k)\mathrm{T}} D^{(k)}(I-\pi_D^{(k)})\mathbf{v}_*^{(k)}=0.$$

Therefore, since $\pi_D^{(k)}=I$ when $n^{(k)}=1$ (entailing $D^{(k)}(I-\pi_D^{(k)})=0$)

$$\begin{aligned}\mu_D &= \frac{\sum_{k=1,\dots,n_c} \mathbf{v}_*^{(k)\mathrm{T}} D^{(k)}(I-\pi_D^{(k)})\mathbf{v}_*^{(k)}+\mathbf{v}_*^{(0)\mathrm{T}} D^{(0)}\mathbf{v}_*^{(0)}}{\sum_{k=0,\dots,n_c} \mathbf{v}_*^{(k)\mathrm{T}} A^{(k)}\mathbf{v}_*^{(k)}+\mathbf{v}_*^{\mathrm{T}} A_r\mathbf{v}_*} \\ &\leqslant \frac{\sum_{k=1,\dots,n_c} \mathbf{v}_*^{(k)\mathrm{T}} D^{(k)}(I-\pi_D^{(k)})\mathbf{v}_*^{(k)}+\mathbf{v}_*^{(0)\mathrm{T}} D^{(0)}\mathbf{v}_*^{(0)}}{\sum_{k=0,\dots,n_c} \mathbf{v}_*^{(k)\mathrm{T}} A^{(k)}\mathbf{v}_*^{(k)}} \\ &\leqslant \max_{\substack{k=0,\dots,n_c \\ \mathbf{v}_*^{(k)}\notin\mathcal{N}(A^{(k)})}} \frac{\mathbf{v}_*^{(k)\mathrm{T}} D^{(k)}(I-\pi_D^{(k)})\mathbf{v}_*^{(k)}}{\mathbf{v}_*^{(k)\mathrm{T}} A^{(k)}\mathbf{v}_*^{(k)}} \\ &\leqslant \max_{k=0,\dots,n_c} \mu_D^{(k)}. \end{aligned}$$

□

A practical consequence of this theorem is to show that nodes for which the corresponding row is strongly dominated by its diagonal element may be kept outside the aggregation process by putting them into the (pseudo) 0th aggregate. The proposition below presents a simple estimate of the pseudo aggregate's quality based on the diagonal dominance excess of corresponding rows.

Proposition 3.3
Assume that A is diagonally dominant, that the splitting $A=A_b+A_r$ satisfies (15) for $j=1,\dots,n^{(0)}$ and that $D^{(0)}=\mathrm{diag}\{(A)_{ii}\,|\,i=1,\dots,n^{(0)}\}$. If $n^{(0)}>0$, one has

$$\mu_D^{(0)} = \max_{\mathbf{v}\in\mathbb{R}^{n^{(0)}}} \frac{\mathbf{v}^{\mathrm{T}} D^{(0)}\mathbf{v}}{\mathbf{v}^{\mathrm{T}} A^{(0)}\mathbf{v}} \leqslant \max_{i=1,\dots,n^{(0)}} \frac{(A)_{ii}}{(A)_{ii}-\sum_{j=1,j\neq i}^{n} |(A)_{ij}|}. \tag{24}$$

Proof
Set $\eta_i=(A)_{ii}-\sum_{j=1,j\neq i}^{n} |(A)_{ij}|$ and note that if $\eta_i=0$ at least for one $i\leqslant n^{(0)}$, the inequality is trivially satisfied. Otherwise, observing that $A^{(0)}\geqslant\mathrm{diag}(\eta_i)$, the inequality (24) follows. □

DOI: 10.1002/nla

Regarding aggregates $1,\dots,n_c$, it is clear that the value of $\mu_D^{(k)}$ strongly depends on $\mathbf{p}^{(k)}$. In the theorem below we further indicate the scope of variation of the aggregate's quality measure if $A^{(k)}$ and $D^{(k)}$ are given, and determine the $\mathbf{p}^{(k)}$ that leads to the the best quality estimate.

Theorem 3.4
Let $A^{(k)}$ and $D^{(k)}$ be, respectively, an $n^{(k)}\times n^{(k)}$ non-zero symmetric non-negative definite matrix and an $n^{(k)}\times n^{(k)}$ SPD matrix, with $n^{(k)}>1$. Let $\mathbf{p}$ be a non-zero vector of size $n^{(k)}$. Let

$$\mu_D^{(k)}=\sup_{\mathbf{v}\in\mathbb{R}^{n^{(k)}}\backslash\mathcal{N}(A^{(k)})}\frac{\mathbf{v}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A^{(k)}\mathbf{v}},\tag{25}$$

where $\pi_D^{(k)}=\mathbf{p}(\mathbf{p}^{\mathrm{T}}D^{(k)}\mathbf{p})^{-1}\mathbf{p}^{\mathrm{T}}D^{(k)}$ and let $\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_{n^{(k)}}$ be the eigenvalues of $D^{(k)-1}A^{(k)}$. Then,

$$\lambda_2^{-1}\leqslant\mu_D^{(k)}\leqslant\lambda_1^{-1}.\tag{26}$$

Moreover, if

$$D^{(k)-1}A^{(k)}\mathbf{p}=\lambda_1\mathbf{p}\tag{27}$$

then

$$\mu_D^{(k)}=\frac{1}{\lambda_2}\tag{28}$$

and, assuming $\mu_D^{(k)}$ finite,

$$\mathbf{v}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}=\mu_D^{(k)}\mathbf{v}^{\mathrm{T}}A^{(k)}\mathbf{v}\quad\text{for some}\quad\mathbf{v}\in\mathcal{R}(A^{(k)})$$

if and only if

$$D^{(k)-1}A^{(k)}\mathbf{v}=\lambda_2\mathbf{v}\quad\text{with}\quad\mathbf{v}^{\mathrm{T}}D^{(k)}\mathbf{p}=0.\tag{29}$$

Proof
Note that the case $\mu_D^{(k)}=\infty$ implies non-empty $\mathcal{N}(A^{(k)})$ and, hence, $\lambda_1=0$. The inequalities (26) are then trivially satisfied. Moreover, according to Theorem 3.2 we have then $\mathcal{N}(A^{(k)})\not\subseteq\text{span}\{\mathbf{p}\}$. Hence, if (27) holds, $\dim(\mathcal{N}(A^{(k)}))\geqslant 2$, which in turn implies $\lambda_2=0$ and, therefore, (28).

Now, consider $\mu_D^{(k)}<\infty$ which, according to Theorem 3.2, implies $\mathcal{N}(A^{(k)})\subset\text{span}\{\mathbf{p}\}$, and, hence, $\lambda_2>0$. If $\mathcal{N}(A^{(k)})$ is non-empty, then $\mathcal{N}(A^{(k)})=\text{span}\{\mathbf{p}\}$ and $\lambda_1=0$, which in turn implies (27). Therefore, for all $\mathbf{v}\in\mathcal{R}(A^{(k)})$, $\mathbf{p}^{\mathrm{T}}D^{(k)}\mathbf{v}=0$, and, hence, $\pi_D^{(k)}\mathbf{v}=\mathbf{0}$. Then, according to Theorem 3.2, we further have

$$\mu_D^{(k)}=\max_{\mathbf{v}\in\mathcal{R}(A^{(k)})\backslash\{\mathbf{0}\}}\frac{\mathbf{v}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A^{(k)}\mathbf{v}}=\max_{\mathbf{v}\in\mathcal{R}(A^{(k)})\backslash\{\mathbf{0}\}}\frac{\mathbf{v}^{\mathrm{T}}D^{(k)}\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A^{(k)}\mathbf{v}}=\lambda_2^{-1}.$$

In addition, a vector $\mathbf{v}$ reaches the maximum if and only if (29) holds.

Finally, we treat the case where $\mathcal{N}(A^{(k)})$ is empty and, hence, $A^{(k)}$ is invertible. Let $\mathbf{x}_i$ be the eigenvector of $D^{(k)-1}A^{(k)}$ associated with the eigenvalue λ_i. To prove the left inequality (26), we set

$$\mathbf{v}=\begin{cases}\mathbf{x}_2 & \text{if } \mathbf{p}^{\mathrm{T}}D^{(k)}\mathbf{x}_2=0,\\ \mathbf{x}_1-\left(\dfrac{\mathbf{p}^{\mathrm{T}}D^{(k)}\mathbf{x}_1}{\mathbf{p}^{\mathrm{T}}D^{(k)}\mathbf{x}_2}\right)\mathbf{x}_2 & \text{otherwise}\end{cases}$$

and note that $\pi_D^{(k)}\mathbf{v}=\mathbf{0}$. Injecting such $\mathbf{v}\neq\mathbf{0}$ into (25) we find

$$\mu_D^{(k)}\geqslant\frac{\mathbf{v}^{\mathrm{T}}D^{(k)}\mathbf{v}}{\mathbf{v}^{\mathrm{T}}A^{(k)}\mathbf{v}}\geqslant\lambda_2^{-1}.$$

The right inequality (26) follows from

$$\mu_D^{(k)} = \max_{\mathbf{v}\in\mathbb{R}^{n^{(k)}}\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}} D^{(k)}(I-\pi_D^{(k)})\mathbf{v}}{\mathbf{v}^{\mathrm{T}} A^{(k)}\mathbf{v}} \leqslant \max_{\mathbf{v}\in\mathbb{R}^{n^{(k)}}\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}} D^{(k)}\mathbf{v}}{\mathbf{v}^{\mathrm{T}} A^{(k)}\mathbf{v}} = \lambda_1^{-1}.$$

Moreover, if $\mathbf{p}=\mathbf{x}_1$, then $\mathbf{x}_i$, $i=1,\ldots,n^{(k)}$ are also eigenvectors of $A^{(k)-1}D^{(k)}(I-\pi_D^{(k)})$ with corresponding eigenvalues $\widetilde{\lambda}_i$ such that $\widetilde{\lambda}_1=0$ and, for $i>1$, $\widetilde{\lambda}_i=\lambda_i^{-1}$. Since $\mu_D^{(k)}$ is the smallest eigenvalue of $A^{(k)-1}D^{(k)}(I-\pi_D^{(k)})$, (28) follows. Moreover, (29) holds if and only if $\mathbf{v}$ is an eigenvector $A^{(k)-1}D^{(k)}(I-\pi_D^{(k)})$ associated with $\lambda_2^{-1}=\mu_D^{(k)}$, which is in turn equivalent to $\mathbf{v}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}=\mu_D^{(k)}\mathbf{v}^{\mathrm{T}}A^{(k)}\mathbf{v}$. □

By the way of illustration, consider a symmetric diagonally dominant M-matrix and assume that the splitting $A=A_b+A_r$ is based on the lower bound of (14); that is, zero sum is required for every row of A_b. Then, each $A^{(k)}$ is singular with its null space equal to span$\{\mathbf{1}_{n^{(k)}}\}$. Theorem 3.2 then shows that one has to use $\mathbf{p}^{(k)}=\mathbf{1}_{n^{(k)}}$ to keep $\mu_D^{(k)}$ finite, in which case, by Theorem 3.4, $\mu_D^{(k)}=\lambda_2(D^{(k)-1}A^{(k)})^{-1}$. When the diagonal dominance is strict, the two-side inequality (14) indicates that there is some freedom in the choice of the diagonal entries of A_b, and one may wonder how to exploit it at the best. The following remarks give some clues in this respect.

Remark 3.1
When $A^{(k)}$ is irreducible and diagonally dominant with non-positive off-diagonal entries, and when $D^{(k)}$ is a diagonal matrix, $D^{(k)-1}A^{(k)}$ is an irreducible M-matrix and, hence, an eigenvector whose components are positive (e.g. $\mathbf{1}_{n^{(k)}}$ when the rows of $A^{(k)}$ have zero sum) is necessarily the eigenvector associated with the smallest eigenvalue, which is unique.

Remark 3.2
Consider a diagonally dominant M-matrix for which the splitting $A=A_b+A_r$ is based on (14). If the diagonal dominance is strict for some rows associated with aggregate k, assuming $\mathbf{p}^{(k)}=\mathbf{1}_{n^{(k)}}$, a nice way to quickly obtain a useful estimate consists in choosing diagonal entries of A_b as large as possible while satisfying (14) with the additional constraint that $\mathbf{1}_{n^{(k)}}$ is an eigenvector of $D^{(k)-1}A^{(k)}$, so that the condition ensuring (28) holds. In particular, when $D^{(k)}$ is a diagonal matrix, it amounts to using $A^{(k)}=A_0^{(k)}+\eta D^{(k)}$ with $A_0^{(k)}$ having all rows with zero sum and with η being the largest constant such that (14) still holds.

Note that, for the discrete PDE problems with constant or piecewise constant coefficients, the procedure in Remark 3.2 corresponds to the splitting (15), except possibly in the neighborhood of the boundary. Therefore, in the reminder of this paper we use the splitting (15) for the aggregates in the interior of the domain. In what follows, we also choose $D=\mathrm{diag}(A)$.

4. DISCRETE PDEs WITH CONSTANT AND SMOOTHLY VARYING COEFFICIENTS

4.1. Preliminaries

We start considering matrices associated with the 5-point stencil

$$\begin{bmatrix} & -\alpha_y & \\ -\alpha_x & \alpha_d & -\alpha_x \\ & -\alpha_y & \end{bmatrix} \text{ with } \alpha_x,\alpha_y>0 \quad\text{and}\quad \alpha_d\geqslant 2(\alpha_x+\alpha_y) \tag{30}$$

on a rectangular grid of arbitrary shape. For such matrices we want to assess boxwise aggregates with four nodes per aggregate (as on Figure 1(a)) and linewise aggregates with two, three and four nodes (as on Figure 1(b)). The prolongation vector is $\mathbf{p}^{(k)}=\mathbf{1}_{n^{(k)}}$, $k=1,\ldots,n_c$ and, as can

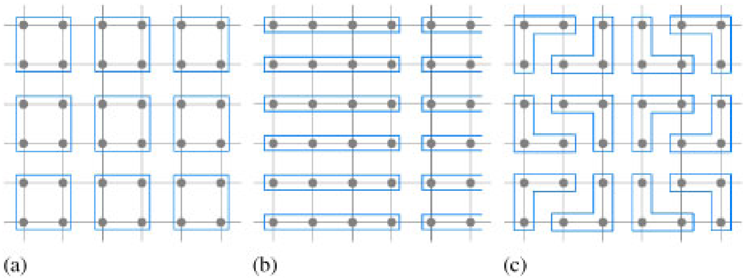

Figure 1. Examples of (a) boxwise; (b) linewise; and (c) L-shaped aggregation patterns.

be checked from (31) and (33) below, it is an eigenvector of $D^{(k)-1}A^{(k)}$ associated with the smallest eigenvalue $\delta_d\alpha_d^{-1}$, where $\delta_d = \alpha_d - 2(\alpha_x+\alpha_y) \geqslant 0$. Theorem 3.4 then implies that $\mu_D^{(k)} = \lambda_2(D^{(k)-1}A^{(k)})^{-1} = \alpha_d\lambda_2(A^{(k)})^{-1}$.

Considering more specifically boxwise aggregates, we have

$$A^{(k)} = \begin{pmatrix} \alpha_x+\alpha_y & -\alpha_x & -\alpha_y & 0 \\ -\alpha_x & \alpha_x+\alpha_y & 0 & -\alpha_y \\ -\alpha_y & 0 & \alpha_x+\alpha_y & -\alpha_x \\ 0 & -\alpha_y & -\alpha_x & \alpha_x+\alpha_y \end{pmatrix} + \delta_d I \tag{31}$$

and, hence,

$$\mu_D^{(k)} = \frac{2\alpha_x + 2\alpha_y + \delta_d}{2\min(\alpha_x,\alpha_y) + \delta_d}, \tag{32}$$

whereas for linewise aggregation of size m in the x direction

$$A^{(k)} = \begin{pmatrix} \alpha_x & -\alpha_x & & \\ -\alpha_x & 2\alpha_x & \ddots & \\ & \ddots & \ddots & -\alpha_x \\ & & -\alpha_x & \alpha_x \end{pmatrix} + \delta_d I \tag{33}$$

and, hence, the following formula holds for $m = 2, \ldots, 4$:

$$\mu_D^{(k)} = \frac{2\alpha_x + 2\alpha_y + \delta_d}{(2-\sqrt{m-2})\alpha_x + \delta_d}. \tag{34}$$

It follows that linewise aggregates of size 4 oriented in the direction of strong coupling become more attractive than boxwise aggregates whenever $\max(\alpha_x,\alpha_y) > (2+\sqrt{2})\min(\alpha_x,\alpha_y)$. Always choosing the best aggregate shape, we have then

$$\mu_D^{(k)} \leqslant 3+\sqrt{2}. \tag{35}$$

Since linewise aggregates of size 3 and 2 have better quality estimates than linewise aggregates of size 4, as can be concluded from (34), this upper bound holds for them as well.

4.2. *Constant coefficients*

We now discuss more specifically the five-point finite difference approximation of

$$\frac{\partial}{\partial x}\left(\alpha_x \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\alpha_y \frac{\partial u}{\partial y}\right) + \beta u = f \quad \text{on } \Omega \tag{36}$$

DOI: 10.1002/nla

(as can be obtained using the mesh box integration scheme [25]) with uniform mesh size h in both directions, where the boundary $\partial\Omega$ of the domain $\Omega \in \mathbb{R}^2$ is the union of segments parallel to the x or y axis and connecting the grid nodes. Note that Ω is possibly not convex and may contain holes.

If the PDE coefficients α_x, α_y and β are constant, the above results allow to assess aggregate's quality for some typical aggregate shapes. It is also easy to extend the reasoning to further aggregation schemes, leading to bound above $\mu_D^{(k)}$ by a modest constant if either coefficients are isotropic ($\alpha_x = \alpha_y$) or if one uses linewise aggregation along the strong coupling direction. For instance, if $\alpha_x = \alpha_y$, (34) with $m = 3$ also applies to L-shaped aggregates as illustrated on Figure 1(c).

Regarding Neumann boundary conditions, the quality of aggregates that contain boundary nodes cannot be directly deduced from the above analysis.[‡] Again, however, isotropic coefficients and linewise aggregates aligned with strong coupling yield bounds similar to (32) and (34). For instance, if $\alpha_x = \alpha_y$ and $\beta = 0$, boxwise aggregation near a Neumann boundary result in matrices $A^{(k)}$ and $D^{(k)}$ that have the form analyzed in Theorem 5.1 below, with $\alpha_1 = \alpha_2$ and $\alpha_3 = \alpha_4 = 0$ (boundary aligned with grid lines), $\alpha_2 = \alpha_3 = \alpha_4 = 0$ (resorting corners), or $\alpha_1 = \alpha_2 = \alpha_3$ and $\alpha_4 = 0$ (re-entering corners). As shown in this theorem, one has then $\mu_D^{(k)} \leqslant 2$ in the two former cases and $\mu_D^{(k)} \leqslant 2.23$ in the latter, compared to $\mu_D^{(k)} = 2$ away from the boundary.

Note that our analysis does not require all aggregates having the same shape, which in fact seldom occurs with practical aggregation algorithms (see [11] for an example). One should just take care that the global μ_D is not larger than desired because of a few irregular aggregates, which in practice can be prevented by breaking them into smaller pieces.

4.3. *Smoothly varying coefficients*

Consider now the same discrete PDE (36) but with smoothly varying coefficients. Because the matrices $A^{(k)}$ and $D^{(k)}$ are local to the aggregate at hand, they are equal, up to a $\mathcal{O}(h)$ perturbation, to the matrices $A_0^{(k)}$ and $D_0^{(k)}$ corresponding to PDE coefficients that are constant and equal to the mean value inside the aggregate. Furthermore, $\mathbf{1}_{n^{(k)}}$ remains the eigenvector of $D^{(k)-1}A^{(k)}$ associated with the smallest eigenvalue either because $\beta = 0$ and, hence, $\mathcal{N}(A^{(k)}) = \operatorname{span}\{\mathbf{1}_{n^{(k)}}\}$, or by using the trick suggested at the end of Section 3 in Remark 3.2 (see also Remark 3.1). Hence, as shown in Theorem 3.4, $\mu_D^{(k)}$ is the inverse of the second smallest eigenvalue of $D^{(k)-1}A^{(k)}$. Since the eigenvalues of a matrix are continuous function of its entries, it means that, asymptotically (for $h \to 0$), $\mu_D^{(k)}$ tends to the smallest eigenvalue of $D_0^{(k)-1}A_0^{(k)}$; that is, to the value obtained in the constant coefficient case. Therefore, the results of the previous subsection carry over the variable coefficient case, at least when the mesh size h is small enough.

4.4. *Numerical example*

We consider the linear system resulting from the 5-point finite difference discretization of (36) on $\Omega = [0, 1] \times [0, 1]$ with Dirichlet boundary conditions and constant coefficients α_x, α_y and $\beta = 0$. The discretization is performed on a uniform rectangular grid of mesh size $h = (N+1)^{-1}$ in both directions.

For the sake of simplicity, we let N be a multiple of 12, which allows that the whole domain is covered with aggregates of the same shape. Using the rule (15), the matrix $A^{(k)}$ is the same for all interior aggregates; the same $A^{(k)}$ is further considered for aggregates near the boundary. Since $D^{(k)}$ is also the same for all aggregates, so is the quality estimate $\mu_D^{(k)}$.

We consider first an isotropic situation ($\alpha_x = \alpha_y$). The columns from 2 to 7 of Table I then give the values of μ_D and of its upper bound $\mu_D^{(k)}$ for three types of aggregation pattern, presented

[‡] The considered discretization scheme implies that off diagonal entries along Neumann boundary are divided by a factor of two compared with entries connecting interior nodes. Hence, the matrices $A^{(k)}$ for aggregates that contain Neumann boundary nodes are not the same as those for aggregates in the interior of the domain.

DOI: 10.1002/nla

Table I. The value of μ_D and of its upper bound (20) for different grid sizes.

	$\alpha_x=\alpha_y,\ \delta_d=0$						$\alpha_x=10\alpha_y,\ \delta_d=0$					
	Pairwise		L-shaped		Boxwise		Linewise (size=3)		Linewise (size=4)		Boxwise	
N	$\mu_D^{(k)}$	μ_D	$\mu_D^{(k)}$	μ_D	$\mu_D^{(k)}$	μ_D	$\mu_D^{(k)}$	μ_D	$\mu_D^{(k)}$	μ_D	$\mu_D^{(k)}$	μ_D
12	2	1.940	4	2.315	2	1.959	2.2	2.184	3.756	3.638	11	8.431
24	2	1.984	4	2.377	2	1.989	2.2	2.196	3.756	3.744	11	10.185
48	2	1.996	4	2.394	2	1.997	2.2	2.199	3.756	3.753	11	10.778
96	2	1.999	4	2.399	2	1.999	2.2	2.200	3.756	3.755	11	10.943

in Figure 1. Observe that when nodes are added to an aggregate, its quality is not necessarily deteriorated, as can be seen comparing L-shaped and box aggregates. We next consider in columns 8 to 13 an anisotropic situation ($\alpha_x=10\alpha_y$). One sees that boxwise aggregation is not recommended in this case.

4.5. Sharpness of the estimate

The numerical results in Table I indicate that the bound (20) on μ_D can be asymptotically sharp for N large enough. Moreover, as shown in Theorem 3.1, if only one Jacobi smoothing iteration is performed, we further have $\rho(E_{\text{TG}})=1-\omega\mu_D^{-1}$. Hence, a sharp estimate of μ_D further leads to a sharp estimate of the two-grid convergence rate. The reader can wonder why and when this happens. This is what we investigate in the present subsection, starting with the first question for the particular case of boxwise aggregates.

Consider that the setting of the above example holds and assumes that both the global ordering of unknowns and of aggregates, and the local one, restricted to a particular aggregate, is lexicographic. Without loss of generality, we assume in addition that $\alpha_x\geqslant\alpha_y$. First, we recall that $D^{(k)-1}A^{(k)}\mathbf{p}^{(k)}=\lambda_1(D^{(k)-1}A^{(k)})\mathbf{p}^{(k)}$, and, hence, the vector $\mathbf{v}_b=(1,\ 1,\ -1,\ -1)^{\mathrm{T}}\in\mathscr{R}(A^{(k)})$ that can be checked to satisfy $D^{(k)-1}A^{(k)}\mathbf{v}_b=\lambda_2(D^{(k)-1}A^{(k)})\mathbf{v}_b$ reaches, according to Theorem 3.4, the supremum in definition (18) of $\mu_D^{(k)}$. Therefore, setting

$$\widetilde{\mathbf{v}}=(\gamma_1\mathbf{v}_b{}^{\mathrm{T}},\gamma_2\mathbf{v}_b{}^{\mathrm{T}},\dots,\gamma_{n_c}\mathbf{v}_b{}^{\mathrm{T}})^{\mathrm{T}}$$

we locally reproduce the maximizing vectors for every aggregate. Moreover, setting $\gamma_1=\gamma_2=\cdots=\gamma_{N/2}=-\gamma_{N/2+1}=\cdots=-\gamma_N=\gamma_{N+1}=\cdots=1$ we further make $\widetilde{\mathbf{v}}$ take the same value at every two connected nodes that belong to different aggregates. Hence, since A_r have the form (13) with diagonal blocks being diagonal matrices, there holds $(A_r)_{ij}((\widetilde{\mathbf{v}})_i-(\widetilde{\mathbf{v}})_j)=0$ for all i and j. Therefore, setting $\sigma_i=\sum_{j=1}^n(A_r)_{ij}$, and since $\sigma_i>0$ only for the unknowns near the boundary, there holds

$$\widetilde{\mathbf{v}}^{\mathrm{T}}A_r\widetilde{\mathbf{v}}=-\sum_{i,j=1}^{n}\frac{1}{2}(A_r)_{ij}((\widetilde{\mathbf{v}})_i-(\widetilde{\mathbf{v}})_j)^2+\sum_{i=1}^{n}\sigma_i(\widetilde{\mathbf{v}})_i^2 \tag{37}$$

$$=\sum_{i=1}^{n}\sigma_i(\widetilde{\mathbf{v}})_i^2$$

$$=2N(\alpha_x+\alpha_y)$$

$$=2N^{-1}(\alpha_x+\alpha_y)\alpha_d^{-1}\widetilde{\mathbf{v}}^{\mathrm{T}}D\widetilde{\mathbf{v}}. \tag{38}$$

On the other hand, note that $\mathbf{p}^{(k)\mathrm{T}} D^{(k)} \mathbf{v}_b = \mathbf{0}$ implies $\pi_D \widetilde{\mathbf{v}} = \mathbf{0}$, and, hence,

$$\begin{aligned}\mu_D &\geqslant \frac{\widetilde{\mathbf{v}}^{\mathrm{T}} D(I-\pi_D)\widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^{\mathrm{T}} A_b \widetilde{\mathbf{v}} + \widetilde{\mathbf{v}}^{\mathrm{T}} A_r \widetilde{\mathbf{v}}} \\ &= \frac{\widetilde{\mathbf{v}}^{\mathrm{T}} D \widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^{\mathrm{T}} A_b \widetilde{\mathbf{v}} + \widetilde{\mathbf{v}}^{\mathrm{T}} A_r \widetilde{\mathbf{v}}} \\ &= \frac{\widetilde{\mathbf{v}}^{\mathrm{T}} D \widetilde{\mathbf{v}}}{\mu_D^{(k)-1} \widetilde{\mathbf{v}}^{\mathrm{T}} D \widetilde{\mathbf{v}} + \widetilde{\mathbf{v}}^{\mathrm{T}} A_r \widetilde{\mathbf{v}}} \\ &= \frac{\mu_D^{(k)}}{1+\mu_D^{(k)} \frac{\widetilde{\mathbf{v}}^{\mathrm{T}} A_r \widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^{\mathrm{T}} D \widetilde{\mathbf{v}}}}. \end{aligned} \tag{39}$$

It then follows from (38) that $\mu_D \to \mu_D^{(k)}$ for $N \to \infty$.

The following theorem is useful in extending this analysis to a more general framework.

Theorem 4.1
Let $A = A_b + A_r$, P, D, μ_D, $\mu_D^{(k)}$ and $\pi^{(k)}$, $k=0,\ldots,n_c$, be defined as in Theorem 3.2. Assume $\mu_D^{(k)}$ finite for $k=0,\ldots,n_c$ and let, for $n^{(0)}>0$ and $n^{(k)}>1$, $k=1,\ldots,n_c$,

$$\begin{aligned} \widetilde{\mathbf{v}}_0 &\in \operatorname*{arg\,max}_{\mathbf{v^{(0)}} \in \mathbb{R}^{n^{(0)}} \backslash \{\mathbf{0}\}} \left(\frac{\mathbf{v}^{(0)\mathrm{T}} D^{(0)} \mathbf{v}^{(0)}}{\mathbf{v}^{(0)\mathrm{T}} A^{(0)} \mathbf{v}^{(0)}} \right), \\ \widetilde{\mathbf{v}}_k &\in \operatorname*{arg\,max}_{\mathbf{v^{(k)}} \in \mathscr{R}(A^{(k)}) \backslash \{\mathbf{0}\}} \left(\frac{\mathbf{v}^{(k)\mathrm{T}} D^{(k)} (I-\pi_D^{(k)}) \mathbf{v}^{(k)}}{\mathbf{v}^{(k)\mathrm{T}} A^{(k)} \mathbf{v}^{(k)}} \right) \end{aligned} \tag{40}$$

with $\widetilde{\mathbf{v}}^{(k)} = 1$ otherwise. Let γ_k, $k=0,\ldots,n_c$, be real parameters, and set

$$\widetilde{\mathbf{v}} = (\gamma_0 \theta_0^{-1} \widetilde{\mathbf{v}}_0{}^{\mathrm{T}}, \gamma_1 \theta_1^{-1} \widetilde{\mathbf{v}}_1^{\mathrm{T}}, \ldots, \gamma_{n_c} \theta_{n_c}^{-1} \widetilde{\mathbf{v}}_{n_c}{}^{\mathrm{T}})^{\mathrm{T}}, \tag{41}$$

where $\theta_k = (\widetilde{\mathbf{v}}_k^{\mathrm{T}} A^{(k)} \widetilde{\mathbf{v}}_k)^{1/2}$ if $n^{(k)}>1$ and $\theta_k = 1$ otherwise. Assume either that

$$\widetilde{\mathbf{v}}^{\mathrm{T}} A_r \widetilde{\mathbf{v}} \leqslant \varepsilon \widetilde{\mathbf{v}}^{\mathrm{T}} A_b \widetilde{\mathbf{v}} \tag{42}$$

or that $n^{(0)} = 0$, that $A^{(k)}$ is singular for $k=1,\ldots,n_c$ and that

$$(\mathbf{c}+\widetilde{\mathbf{v}})^{\mathrm{T}} A_r (\mathbf{c}+\widetilde{\mathbf{v}}) \leqslant \varepsilon \left(\max_{k=1,\ldots,n_c} \mu_D^{(k)} \right)^{-1} \widetilde{\mathbf{v}}^{\mathrm{T}} D \widetilde{\mathbf{v}} \tag{43}$$

for some vector $\mathbf{c} = (\xi_1 \mathbf{p}^{(1)\mathrm{T}}, \ldots, \xi_{n_c} \mathbf{p}^{(n_c)\mathrm{T}})^{\mathrm{T}}$.

Then

$$\mu_D \geqslant \frac{1}{1+\varepsilon} \frac{\sum_{k=0}^{n_c} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0}^{n_c} \gamma_k^2}. \tag{44}$$

Proof

We first prove the lower bound (44) based on the assumption (42). Starting with the equality (23) in the proof of Theorem 3.2 and setting $\mathbf{v}^{(k)}=\gamma_k\theta_k^{-1}\tilde{\mathbf{v}}_k$ together with $\mathbf{v}=\tilde{\mathbf{v}}$, we have

$$\begin{aligned}\mu_D &\geqslant \frac{\sum_{k=1,\ldots,n_c}\mathbf{v}^{(k)\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\mathbf{v}^{(k)}+\mathbf{v}^{(0)\mathrm{T}}D^{(0)}\mathbf{v}^{(0)}}{\sum_{k=0,\ldots,n_c}\mathbf{v}^{(k)\mathrm{T}}A^{(k)}\mathbf{v}^{(k)}+\mathbf{v}^{\mathrm{T}}A_r\mathbf{v}} \\ &\geqslant \frac{1}{1+\varepsilon}\frac{\sum_{k=1,\ldots,n_c}\gamma_k^2\theta_k^{-2}\tilde{\mathbf{v}}_k^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\tilde{\mathbf{v}}_k+\gamma_0^2\theta_0^{-2}\tilde{\mathbf{v}}_0{}^{\mathrm{T}}D^{(0)}\tilde{\mathbf{v}}_0}{\sum_{k=0,\ldots,n_c}\gamma_k^2\theta_k^{-2}\tilde{\mathbf{v}}_k^{\mathrm{T}}A^{(k)}\tilde{\mathbf{v}}_k} \\ &= \frac{1}{1+\varepsilon}\frac{\sum_{k=0,\ldots,n_c}\gamma_k^2\mu_D^{(k)}}{\sum_{k=0,\ldots,n_c}\gamma_k^2},\end{aligned} \tag{45}$$

where the last equality follows from $\theta_k^{-2}\tilde{\mathbf{v}}_k{}^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\tilde{\mathbf{v}}_k=\mu_D^{(k)}$.

Now, we prove the lower bound (44) based on the assumptions related to (43). We may then assume that

$$(\mathbf{c}+\tilde{\mathbf{v}})^{\mathrm{T}}A_r(\mathbf{c}+\tilde{\mathbf{v}})\leqslant\varepsilon\tilde{\mathbf{v}}^{\mathrm{T}}A_b\tilde{\mathbf{v}} \tag{46}$$

holds, this inequality being proved later. Since $\mu_D^{(k)}$ is finite, Theorem 3.2 implies $\mathcal{N}(A_r^{(k)})\subset\mathrm{span}\{\mathbf{p}^{(k)}\}$, $k=1,\ldots,n_c$. From the singularity of $A^{(k)}$, $k=1,\ldots,n_c$, we further conclude that $\mathcal{N}(A_r^{(k)})=\mathrm{span}\{\mathbf{p}^{(k)}\}$ and, hence,

$$(\xi_k\mathbf{p}^{(k)}+\tilde{\mathbf{v}}_k)^{\mathrm{T}}A^{(k)}(\xi_k\mathbf{p}^{(k)}+\tilde{\mathbf{v}}_k)=\tilde{\mathbf{v}}_k^{\mathrm{T}}A^{(k)}\tilde{\mathbf{v}}_k. \tag{47}$$

Moreover, using definition (19) of $\pi_D^{(k)}$, we also have

$$(\xi_k\mathbf{p}^{(k)}+\tilde{\mathbf{v}}_k)^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})(\xi_k\mathbf{p}^{(k)}+\tilde{\mathbf{v}}_k)=\tilde{\mathbf{v}}_k^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\tilde{\mathbf{v}}_k. \tag{48}$$

Therefore, injecting $\mathbf{v}=\mathbf{c}+\tilde{\mathbf{v}}$ and $\mathbf{v}^{(k)}=\xi_k\mathbf{p}^{(k)}+\gamma_k\theta_k^{-1}\tilde{\mathbf{v}}_k$ into (45) and using (47) and (48), the proof is finished as in the previous case.

We are thus left with the proof of (46). From Theorem 3.4 we conclude that $D^{(k)-1}A^{(k)}\tilde{\mathbf{v}}_k=\lambda_2(D^{(k)-1}A^{(k)})\tilde{\mathbf{v}}_k$ and $\tilde{\mathbf{p}}^{(k)T}D^{(k)}\tilde{\mathbf{v}}_k=0$. Therefore,

$$\tilde{\mathbf{v}}_k^{\mathrm{T}}D^{(k)}(I-\pi_D^{(k)})\tilde{\mathbf{v}}_k=\tilde{\mathbf{v}}_k^{\mathrm{T}}D^{(k)}\tilde{\mathbf{v}}_k,$$

which implies $\tilde{\mathbf{v}}_k^{\mathrm{T}}D^{(k)}\tilde{\mathbf{v}}_k=\mu_D^k\tilde{\mathbf{v}}_k{}^{\mathrm{T}}A^{(k)}\tilde{\mathbf{v}}_k$. Hence,

$$\tilde{\mathbf{v}}^{\mathrm{T}}D\tilde{\mathbf{v}}\leqslant\left(\max_{k=1,\ldots,n_c}\mu_D^{(k)}\right)\tilde{\mathbf{v}}^{\mathrm{T}}A_b\tilde{\mathbf{v}},$$

which, together with (43) implies (46). □

Now, we return to the previous example and prove the asymptotical sharpness for linewise aggregates of size $m\leqslant 4$. As in the boxwise case, the vector (40) is the second eigenvector of $\alpha_d^{-1}A^{(k)}$ given by

$$\mathbf{v}_b=\begin{cases}(1,\sqrt{2}-1,1-\sqrt{2},-1)^{\mathrm{T}} & \text{if } m=4,\\ (1,0,-1)^{\mathrm{T}} & \text{if } m=3,\\ (1,-1)^{\mathrm{T}} & \text{if } m=2.\end{cases}$$

Hence, choosing

$$\widetilde{\mathbf{v}} = (\gamma_1 \mathbf{v}_b^{\mathrm{T}}, \gamma_2 \mathbf{v}_b^{\mathrm{T}}, \ldots, \gamma_{n_c} \mathbf{v}_b^{\mathrm{T}})^{\mathrm{T}}$$

with $\gamma_1 = -\gamma_2 = \gamma_3 = \cdots = \gamma_{N/m} = -\gamma_{N/m+1} = \gamma_{N/m+2} = \cdots = 1$, we further make $\widetilde{\mathbf{v}}$ take the same value at every two connected nodes that belong to different aggregates. Next, using again (37) with first term in the right hand side vanishing, we have

$$\begin{aligned}\widetilde{\mathbf{v}} A_r \widetilde{\mathbf{v}} &= 2Nm^{-1}\alpha_y \|\mathbf{v}_b\|^2 + \alpha_x N(\|(\mathbf{v}_b)_1\|^2 + \|(\mathbf{v}_b)_m\|^2) \\ &\leqslant 2N(\alpha_x + \alpha_y)\|\mathbf{v}_b\|^2 \\ &= 2N^{-1}(\alpha_x + \alpha_y)\alpha_d^{-1} \mu_D^{(k)} \widetilde{\mathbf{v}}^{\mathrm{T}} A_b \widetilde{\mathbf{v}}, \end{aligned} \qquad (49)$$

and, hence, (44) holds with $\varepsilon = N^{-1}(\alpha_x + \alpha_y)\alpha_d^{-1}\mu_D^{(k)}$. The asymptotical sharpness follows then from Theorem 4.1.

Considering a general situation, we note that a lower bound close to the upper bound (20) can be proved via (44) if there exists a vector $\widetilde{\mathbf{v}}$ of the form (41), such that

(a) $\frac{\sum_{k=0}^{n_c} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0}^{n_c} \gamma_k^2}$ is close to $\max_{k=1,\ldots,n_c} \mu_D^{(k)}$;

(b) ε, defined via (42) or (43), is small compared to 1.

Now, the condition (a) can be satisfied by using large values of γ_k^2 where $\mu_D^{(k)}$ is large. When all $\mu_D^{(k)}$ are the same, we trivially have

$$\frac{\sum_{k=0}^{n_c} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0}^{n_c} \gamma_k^2} = \max_{k=1,\ldots,n_c} \mu_D^{(k)}$$

independently of the choice of γ_k. As illustrated in Section 5, the use of γ_k^2 with variable magnitude allows to prove the asymptotical sharpness in the case where the $\mu_D^{(k)}$s are not all the same.

Condition (b) is more difficult to check. One may start from relation (37) and look for a vector $\widetilde{\mathbf{v}}$ of the form (41) such that $(A_r)_{ij}((\widetilde{\mathbf{v}})_i - (\widetilde{\mathbf{v}})_j) = 0$ for all i and j. If such a vector exists, the first term in (37) is zero. Then, let $\Omega_h = \{1, \ldots, n\}$ be the set of all unknowns and set $\partial\Omega_h = \{i \,|\, \sigma_i = \sum_{j=1}^n (A_r)_{ij} \neq 0\}$. If as in the previous example σ_i is positive only for unknowns near the boundary, then $\partial\Omega_h$ is a set of 'boundary' unknowns. Assuming σ_i and $(\widetilde{\mathbf{v}})_i$, $i = 1, \ldots, n_c$ reasonably bounded, we have

$$\widetilde{\mathbf{v}}^{\mathrm{T}} A_r \widetilde{\mathbf{v}} = \sum_{i=0}^{n} \sigma_i (\widetilde{\mathbf{v}})_i^2 = \mathcal{O}(|\partial\Omega_h|),$$

whereas, assuming γ_k^2, $k = 1, \ldots, n_c$, bounded below,

$$\widetilde{\mathbf{v}}^{\mathrm{T}} A_b \widetilde{\mathbf{v}} = \sum_{k=0}^{n_c} \gamma_k^2 \theta_k^{-2} \widetilde{\mathbf{v}}_k^{\mathrm{T}} A^{(k)} \widetilde{\mathbf{v}}_k = \sum_{k=0}^{n_c} \gamma_k^2 = \mathcal{O}(|\Omega_h|),$$

In a discretized PDE context, the ratio $|\partial\Omega_h|/|\Omega_h|$ usually becomes arbitrary small as the mesh is refined.

Furthermore, the lower bound (44) can be obtained using only a (given) set of aggregates (numbered from 1 to $\bar{n}_c$ for convenience), setting

$$\widetilde{\mathbf{v}} = (\gamma_1 \theta_1^{-1} \widetilde{\mathbf{v}}_1^{\mathrm{T}}, \ldots, \gamma_{\bar{n}_c} \theta_{\bar{n}_c}^{-1} \widetilde{\mathbf{v}}_{\bar{n}_c}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}, \ldots, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}. \qquad (50)$$

Then, (37) becomes

$$\widetilde{\mathbf{v}}A_r\widetilde{\mathbf{v}} = -\sum_{i,j\in\bar{\Omega}_h}\frac{1}{2}(A_r)_{ij}((\widetilde{\mathbf{v}})_i-(\widetilde{\mathbf{v}})_j)^2+\sum_{i\in\bar{\Omega}_h}\bar{\sigma}_i(\widetilde{\mathbf{v}})_i^2,$$

where $\bar{\Omega}_h$ is the set of unknowns belonging to the first $\bar{n}_c$ aggregates and $\bar{\sigma}_i=\sum_{j\in\bar{\Omega}_h}(A_r)_{ij}$. Again, setting $\partial\bar{\Omega}_h=\{i\,|\,\bar{\sigma}_i\neq 0\}$ and repeating the steps described above, one obtains

$$\mu_D \geqslant \frac{1}{1+\bar{\varepsilon}}\,\frac{\sum_{k=1}^{\bar{n}_c}\gamma_k^2\mu_D^{(k)}}{\sum_{k=1}^{\bar{n}_c}\gamma_k^2}$$

with $\bar{\varepsilon}=\mathcal{O}(|\partial\bar{\Omega}_h|/|\bar{\Omega}_h|)$. In practice, it means that the upper bound (20) can also be asymptotically sharp when the $\mu_D^{(k)}$s are not all equal, providing that the aggregates for which $\mu_D^{(k)}$ is maximal cover a significant part of the domain.

As an example, consider a scalar PDE discretized on a grid from which we can extract a $\bar{\Omega}_h=\bar{N}\times\bar{N}$ square of nodes with every node corresponding to the same stencil of the form (30). Then, assuming that the whole square is covered with box aggregates as in the Figure 1(a), the relations (32), (39) and (38) can be used (with $\bar{N}$ instead of N) to show that

$$\mu_D \geqslant \frac{1}{1+\bar{\varepsilon}}\bar{\mu}_D \tag{51}$$

with $\bar{\mu}_D=\frac{2\alpha_x+2\alpha_y+\delta_D}{2\min(\alpha_x,\alpha_y)+\delta_d}$ and $\bar{\varepsilon}=2\bar{N}^{-1}(\alpha_x+\alpha_y)\alpha_d^{-1}\bar{\mu}_D$.

5. DISCRETE PDEs WITH DISCONTINUOUS COEFFICIENTS

5.1. Analysis

We consider the PDE (36) with piecewise constant isotropic coefficients ($\alpha_x(x,y)=\alpha_y(x,y)$) and $\beta=0$, and assume Dirichlet boundary conditions. As in the previous section, we consider the five-point finite difference approximation with uniform mesh size h in both the directions (mesh box integration scheme [25]), and assume that the boundary $\partial\Omega$ of $\Omega\subset\mathbb{R}^2$ is the union of segments parallel to the x or y axis and connecting the grid nodes. We aim at assessing boxwise aggregation as illustrated in Figure 1(a), which was shown relevant for isotropic coefficients in the previous section.

Here, we assume that the possible discontinuities match the grid lines. Hence, Ω is a union of non-overlapping subdomains Ω_i in which the coefficients are constant, and the boundary $\partial\Omega_i$ of each Ω_i is formed by segments aligned with grid lines and having grid nodes as end points. To exclude some uncommon situations, we assume that every two such end points are separated by a distance not less than $2h$ and that each box aggregate contains at least one point which is interior to one of the subdomains. In practice, this assumption is automatically met if the mesh size is small enough; in fact, it has to be not larger than $h_0/2$, where h_0 is the size of the coarsest mesh that still correctly reproduce the geometry of the problem.

The most general situation corresponding to this setting is then schematized in Figure 2(a), where the central aggregate has one node interior to Ω_1 and the opposite node at the intersection of four subdomains: Ω_1, Ω_2, Ω_3 and Ω_4. With the splitting satisfying (15), the corresponding aggregate's matrices $A^{(k)}$ and $D^{(k)}$ are given below by (52) and (53), respectively, with α_i, $i=1,\ldots,4$, being the PDE coefficient in the subdomain Ω_i. Because of the assumption (15) and of Theorem 3.4, aggregate's quality $\mu_D^{(k)}$ is the inverse of the second smallest eigenvalue of $D^{(k)-1}A^{(k)}$. The following theorem is helpful when assessing this latter. In order to alleviate the presentation, we give its proof in the Appendix.

DOI: 10.1002/nla

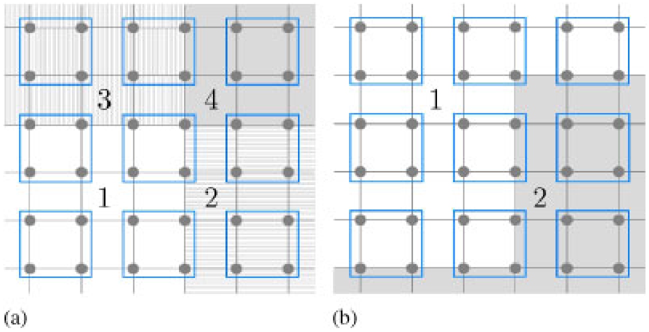

Figure 2. (a) general box aggregate situation with respect to discontinuities and (b) discontinuity nodes aggregated with nodes in white subdomain.

Theorem 5.1
Let

$$A_d=\frac{1}{2}\begin{pmatrix} 4\alpha_1 & -2\alpha_1 & -2\alpha_1 & \\ -2\alpha_1 & 3\alpha_1+\alpha_2 & & -\alpha_1-\alpha_2 \\ -2\alpha_1 & & 3\alpha_1+\alpha_3 & -\alpha_1-\alpha_3 \\ & -\alpha_1-\alpha_2 & -\alpha_1-\alpha_3 & 2\alpha_1+\alpha_2+\alpha_3 \end{pmatrix} \tag{52}$$

and

$$D_d=\operatorname{diag}(4\alpha_1 2(\alpha_1+\alpha_2)2(\alpha_1+\alpha_3)(\alpha_1+\alpha_2+\alpha_3+\alpha_4)), \tag{53}$$

where $\alpha_1>0$ and α_2, α_3, $\alpha_4>0$. A_d is positive semi-definite, and let $\lambda_2(D_d^{-1}A_d)$ be the smallest non-zero eigenvalue of $D_d^{-1}A_d$.

Then

$$\lambda_2(D_d^{-1}A_d)\geqslant\frac{5-\sqrt{17}}{8} \tag{54}$$

and, if $\alpha_1=\alpha_2$ and $\alpha_3=\alpha_4$, there holds

$$\lambda_2(D_d^{-1}A_d)=\min\left(\frac{1}{2},\frac{3\alpha_1+\alpha_3}{4(\alpha_1+\alpha_3)}\right). \tag{55}$$

Moreover, if $\alpha_1\geqslant\alpha_2$, α_3, α_4, one has

$$\lambda_2(D_d^{-1}A_d)\geqslant\beta \tag{56}$$

with $\beta=\lambda_2(D_d^{-1}A_d)(\approx 0.449)$ being evaluated for $\alpha_1=\alpha_2=\alpha_4=1$ and $\alpha_3=0$.

Furthermore,

$$\lambda_2(D_d^{-1}A_d)\geqslant\tfrac{1}{2}\quad\text{if}\quad\begin{cases}\alpha_1\geqslant\alpha_2=\alpha_3=\alpha_4,\\ \text{or } \alpha_1=\alpha_2\geqslant\alpha_3=\alpha_4,\\ \text{or } \alpha_1=\alpha_2=\alpha_3\geqslant\alpha_4.\end{cases} \tag{57}$$

Proof
See Appendix A.

This theorem enables us to draw the following conclusions:

- The approach is robust in all cases, since, by (54), $\mu_D^{(k)}$ is always bounded above independently of the relationship between the coefficients α_i.

DOI: 10.1002/nla

- Nevertheless, from a practical viewpoint, (54) allows a significant decrease of aggregate's quality compared with the constant coefficient case. However, according to (56), which implies $\mu_D^{(k)} \leqslant 2.23$ (compared with 2 in constant coefficient case), a major deterioration is avoided when $\alpha_1 \geqslant \alpha_2, \alpha_3, \alpha_4$. The latter condition is satisfied if nodes belonging to several subdomains Ω_i are always aggregated only with nodes that belong to Ω_i with largest PDE coefficient α_i. Roughly speaking, the rule may be summarized as 'aggregate discontinuity nodes with those of the strong coefficient region'.
- In many practical cases, no more than two subdomains are involved at a time for a single aggregate, and either $\alpha_1 = \alpha_2 = \alpha_3$, or $\alpha_1 = \alpha_2$ and $\alpha_3 = \alpha_4$, or $\alpha_2 = \alpha_3 = \alpha_4$ hold, as illustrated on Figure 2(b). Then, if the rule above is applied; that is, if α_1 is in addition the largest coefficient, (57) applies and shows that there is no deterioration at all compared with the constant coefficient case.

5.2. *Numerical example*

Consider the PDE (36) on a square domain $\Omega = [0,1] \times [0,1]$ with $\beta = 0$,

$$\alpha_x(x,y) = \alpha_y(x,y) = \begin{cases} 1 & \text{if } x \leqslant 1/2, \\ d(>1) & \text{if } x > 1/2 \end{cases}$$

and with Dirichlet boundary conditions. Consider the linear system (1) resulting from its five-point finite difference discretization (mesh box integration scheme [25]) on the regular grid of mesh size $h = N^{-1}$. Since the discontinuity needs to be aligned with grid lines, N has to be even. For simplicity of presentation, we further assume that it is a multiple of 4. The number of unknowns being $(N-1) \times (N-1)$ (there is no unknown for Dirichlet nodes), the grid cannot be covered with box aggregates only and the coarsening is completed by pair and singleton aggregates. Then, the domain may be covered with box aggregates starting from the left bottom corner (as on Figure 3(a)) or from the right bottom corner (as on Figure 3(b)). The splitting (15) is used for interior aggregates and zero sum is imposed for the rows of $A^{(i)}$ on the boundary.

Note that the quality of aggregates outside discontinuity is at most 2, as can be concluded in the isotropic case ($\alpha_x = \alpha_y$) from (32) (for box aggregates) or from (34) with $m = 2$ (for pair aggregates). The bound is therefore determined by the quality of aggregates containing nodes on the discontinuity, which are given for $d = 10$ in Table II. Observe that for the second strategy the convergence estimate is exactly the same as in the constant coefficient case. For box aggregate, this follows from the analysis in the previous subsection: the aggregates then obeys the 'strong coefficient' rule stated above. Regarding the first aggregation strategy, note that for box aggregates one has

$$\mu_D^{(k)} = \lambda_2(D^{(k)-1} A^{(k)})^{-1} = \frac{4(1+d)}{3+d} \tag{58}$$

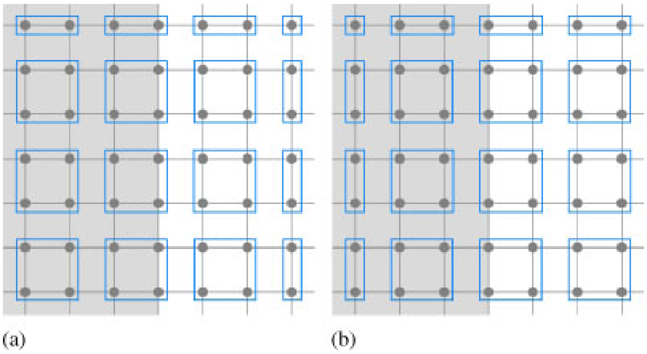

Figure 3. Two potential aggregation strategies for the numerical example.

DOI: 10.1002/nla

Table II. The value of μ_D and of its upper bound (20) for different aggregation strategies and for $d=10$.

	Strategy (a)		Strategy (b)	
N	$\max_{k=0,\ldots,n_c} \mu_D^{(k)}$	μ_D	$\max_{k=0,\ldots,n_c} \mu_D^{(k)}$	μ_D
32	3.385	3.181	2	1.993
64	3.385	3.286	2	1.998
128	3.385	3.336	2	2.000
256	3.385	3.361	2	2.000

using (55) with $\alpha_1=\alpha_2=1$ and $\alpha_3=\alpha_4=d$. This is also true in the pairwise case, since then

$$A^{(k)}=\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad D^{(k)}=\begin{pmatrix} 4 & \\ & 2(d+1) \end{pmatrix}.$$

Note that (58) implies $\mu_D^{(k)}=3.38$ for $d=10$ and $\mu_D^{(k)}\to 4$ for $d\to\infty$.

5.3. Sharpness of the estimate

Table II indicates that, once again, the upper bound (20) is seemingly asymptotically exact. In fact, the reason developed at the end of Section 4 shows that, asymptotically, μ_D cannot be smaller than 2 for an isotropic ($\alpha_x=\alpha_y$) PDE (36) with $\beta=0$ and a regular covering by box aggregates in at least one subdomain in which the PDE coefficients are constant. Hence, our analysis is accurate when discontinuity nodes are aggregated with nodes in the strong coefficient region, since then $\mu_D^{(k)}\leqslant 2.23$. If, in addition, $\mu_D^{(k)}\leqslant 2$, like in the numerical example above, then the bound is asymptotically sharp.

It is more challenging to show the sharpness when $\mu_D^{(k)}$ is significantly larger than 2 for some aggregates along discontinuity, essentially because the proportion of such aggregates is $\mathcal{O}(h)$ or less. Nevertheless, it is interesting to confirm that, as seen in Table II, such a limited amount of low quality aggregates is sufficient to affect the global convergence, and hence that the rule ‘aggregate discontinuity nodes with those of the strong coefficient region’ has some practical relevance.

In this view, we prove the sharpness of our estimate for the above numerical example with the first aggregation strategy (depicted in Figure 3(a)), which does not follow the ‘strong coefficient’ rule. Note that, using the same trick as explained at the end of Section 4, a similar lower bound on μ_D can be obtained in more complicated examples whose domain would contain a rectangular region with two subdomains separated by a line in the middle and covered similarly with box aggregates.

To apply Theorem 4.1, we need to construct two vectors $\widetilde{\mathbf{v}}$ and $\widetilde{\mathbf{c}}$ such that

$$\frac{\sum_{k=0}^{n_c} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0}^{n_c} \gamma_k^2} \to \max_{k=1,\ldots,n_c} \mu_D^{(k)} \quad \text{for } N\to\infty, \tag{59}$$

whereas ε, defined by (43), goes to 0 as N becomes large. In the example under investigation, there are some pair and singleton aggregates (see Figure 3), but we limit the support of both vectors to the $\bar{n}_c=(2\ell+1)\times(2\ell+1)$ box aggregates, where $\ell=N/4-1$. We identify each such aggregate k with a couple $(i_x^{(k)}, i_y^{(k)})$ of indices, $1\leqslant i_x^{(k)}, i_y^{(k)}\leqslant 2\ell+1$, such that $(i_x^{(k)}+1, i_y^{(k)})$, $(i_x^{(k)}-1, i_y^{(k)})$, $(i_x^{(k)}, i_y^{(k)}+1)$ and $(i_x^{(k)}, i_y^{(k)}-1)$ are, respectively, its right, left, top and bottom neighboring aggregates. Note that the center of the domain is a node belonging to aggregate $(\ell+1,\ell+1)$ and that discontinuity aggregates satisfy $i_x^{(k)}=\ell+1$.

Since $\mathbf{p}^{(k)}=\mathbf{1}_{n^{(k)}}$, the vector $\widetilde{\mathbf{v}}_k$ from Theorem 4.1 is given by the eigenvector of $D^{(k)-1}A^{(k)}$, associated with the second smallest eigenvalue $\lambda_2(D^{(k)-1}A^{(k)})$; that is, by $\mathbf{v}_d=(\tau,1,\tau,1)^{\mathrm{T}}$ for discontinuity aggregate, with $\tau=-(d+1)/2$, and by $\mathbf{v}_o=(-1,1,-1,1)^{\mathrm{T}}$ for the ordinary ones.

DOI: 10.1002/nla

The corresponding local energy (semi-) norms are given by $\theta_d^2=\mathbf{v}_d^{\mathrm{T}}A^{(k)}\mathbf{v}_d=(3+d)^2/2$ for discontinuity aggregates, by $\theta_o^2=\mathbf{v}_o^{\mathrm{T}}A^{(k)}\mathbf{v}_o=8$ for the aggregates on the left of the discontinuity line and by $d\theta_o^2$ for those on the right of it.

Then, the vector $\widetilde{\mathbf{v}}$ is defined by $(\widetilde{\mathbf{v}}^{(1)\mathrm{T}},\widetilde{\mathbf{v}}^{(2)\mathrm{T}},\ldots,\widetilde{\mathbf{v}}^{(\bar{n}_c)\mathrm{T}})^T$ with

$$\widetilde{\mathbf{v}}^{(k)}=\ell^{-1}(\ell-|\ell+1-i_y^{(k)}|)\times\begin{cases}\dfrac{\tau}{2}\ell^{-1}\mathbf{v}_o & \text{if } 1\leqslant i_x^{(k)}<\ell+1,\\ \mathbf{v}_d & \text{if } i_x^{(k)}=\ell+1,\\ \dfrac{1}{2}\ell^{-1}\mathbf{v}_o & \text{if } \ell+1<i_x^{(k)}\leqslant 2\ell+1,\\ \mathbf{0} & \text{otherwise}\end{cases}\tag{60}$$

and the vector $\widetilde{\mathbf{c}}$ corresponds to $(\widetilde{\mathbf{c}}^{(1)\mathrm{T}},\widetilde{\mathbf{c}}^{(2)\mathrm{T}},\ldots,\widetilde{\mathbf{c}}^{(\bar{n}_c)\mathrm{T}})^T$ with

$$\widetilde{\mathbf{c}}^{(k)}=\left(\ell-|\ell+1-i_x^{(k)}|+\tfrac{1}{2}\right)(\ell-|\ell+1-i_y^{(k)}|)\ell^{-2}\times\begin{cases}\tau\mathbf{1}_4 & \text{if } 1\leqslant i_x^{(k)}<\ell+1,\\ \mathbf{0} & \text{if } i_x^{(k)}=\ell+1,\\ \mathbf{1}_4 & \text{if } \ell+1<i_x^{(k)}\leqslant 2\ell+1,\\ \mathbf{0} & \text{otherwise.}\end{cases}$$

From (60) we conclude that

$$\gamma_k^2=\ell^{-2}(\ell-|\ell+1-i_y^{(k)}|)^2\times\begin{cases}\dfrac{\tau^2}{4}\ell^{-2}\theta_o^2 & \text{if } 1\leqslant i_x^{(k)}<\ell+1,\\ \theta_d^2 & \text{if } i_x^{(k)}=\ell+1,\\ \dfrac{1}{4}\ell^{-2}d\theta_o^2 & \text{if } \ell+1<i_x^{(k)}\leqslant 2\ell+1,\\ \mathbf{0} & \text{otherwise,}\end{cases}\tag{61}$$

and, setting

$$\mathrm{s}(\ell)=\sum_{i=1}^{2\ell-1}(\ell-|\ell-i|)^2=\sum_{i=1}^{\ell}(i^2+(i-1)^2)=\ell(\ell+1)(2\ell+1)/3-\ell^2$$

there holds

$$\sum_{k:i_x^{(k)}=\ell+1}\gamma_k^2=\theta_d^2\sum_{1<i_y^{(k)}<2\ell+1}\ell^{-2}(\ell-|\ell+1-i_y^{(k)}|)^2=\theta_d^2\ell^{-2}\mathrm{s}(\ell),$$

$$\sum_{k:i_x^{(k)}\neq\ell+1}\gamma_k^2=\theta_o^2\frac{d+\tau^2}{4}\sum_{\substack{1<i_y^{(k)}<2\ell+1\\ 1\leqslant i_x^{(k)}<\ell+1}}\ell^{-4}(\ell-|\ell+1-i_y^{(k)}|)^2=\theta_o^2\frac{d+\tau^2}{4}\ell^{-3}\mathrm{s}(\ell).$$

Hence, $\sum_k\gamma_k^2\mu_D^{(k)}=(1+\mathcal{O}(\ell^{-1}))\sum_{k:i_x^{(k)}=\ell+1}\gamma_k^2\mu_D^{(k)}$, entailing (59) since $\mu_D^{(k)}$ is maximal for $i_x^{(k)}=\ell+1$.

On the other hand, observe that $\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}}$ takes the same value at any two connected nodes belonging to aggregates $(i_x^{(k)},i_y^{(k)})$ and $(i_x^{(k)}+1,i_y^{(k)})$. Moreover, $\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}}$ vanishes on the boundary of the region delimited by box aggregates. Hence, the only contribution to $(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})^{\mathrm{T}}A_r(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})$ as expressed by (37) comes from connections between $(i_x^{(k)},i_y^{(k)})$ and $(i_x^{(k)},i_y^{(k)}+1)$. In this latter case, let j_1 and j_2 be two connected nodes belonging to aggregates $(i_x^{(k)},i_y^{(k)})$ and $(i_x^{(k)},i_y^{(k)}+1)$, respectively, with

$i_y^{(k)}\leqslant 2\ell$. For every box aggregate k, let k_+ (resp. k_-) be the set of two nodes belonging to this aggregate with larger (resp. smaller) abscise. One then has

$$|(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})_{j_1}-(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})_{j_2}|=\begin{cases}\tau\ell^{-2}(\ell-|\ell+1-i_x^{(k)}|) & \text{if } 1\leqslant i_x^{(k)}<\ell+1\text{and } j_1\in k_-,\\ \tau\ell^{-2}(\ell-|\ell+1-i_x^{(k)}|+1) & \text{if } 1\leqslant i_x^{(k)}<\ell+1\text{and } j_1\in k_+,\\ \tau\ell^{-1} & \text{if } i_x^{(k)}=\ell+1\text{and } j_1\in k_-,\\ \ell^{-1} & \text{if } i_x^{(k)}=\ell+1\text{and } j_1\in k_+,\\ \ell^{-2}(\ell-|\ell+1-i_x^{(k)}|+1) & \text{if } \ell+1<i_x^{(k)}\leqslant 2\ell+1\text{and } j_1\in k_-,\\ \ell^{-2}(\ell-|\ell+1-i_x^{(k)}|) & \text{if } \ell+1<i_x^{(k)}\leqslant 2\ell+1\text{and } j_1\in k_+.\end{cases}$$

Therefore, using (37) with, this time, the first term being non-zero and the second one vanishing because of the limited scope of $\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}}$, we have

$$\begin{aligned}(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})^{\mathrm{T}}A_r(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}}) &= \left(\tau^2+\frac{1+d}{2}\right)\sum_{\substack{i_x^{(k)}=\ell+1\\ 1\leqslant i_y^{(k)}\leqslant 2\ell}}\ell^{-2}\\ &\quad +(\tau^2+d)\sum_{\substack{1\leqslant i_x^{(k)}<\ell+1\\ 1\leqslant i_y^{(k)}\leqslant 2\ell}}\ell^{-4}((\ell-|\ell+1-i_x^{(k)}|+1)^2+(\ell-|\ell+1-i_x^{(k)}|)^2)\\ &= (2\tau^2+1+d)\ell^{-1}+2(\tau^2+d)\ell^{-3}\sum_{1\leqslant i_x^{(k)}<\ell+1}(i_x^{(k)2}+(i_x^{(k)}-1)^2)\\ &= (2\tau^2+1+d)\ell^{-1}+2(\tau^2+d)\ell^{-3}\mathrm{s}(\ell),\end{aligned}$$

whereas

$$\begin{aligned}\widetilde{\mathbf{v}}^{\mathrm{T}}D\widetilde{\mathbf{v}} &= \mathbf{v}_d^{\mathrm{T}}\mathbf{v}_d(2d+2)\sum_{\substack{1<i_y^{(k)}<2\ell+1\\ i_x^{(k)}=\ell+1}}\ell^{-2}(\ell-|\ell+1-i_y^{(k)}|)^2+\mathbf{v}_o^{\mathrm{T}}\mathbf{v}_o(d+\tau^2)\sum_{\substack{1<i_y^{(k)}<2\ell+1\\ 1\leqslant i_x^{(k)}<\ell+1}}\ell^{-4}(\ell-|\ell+1-i_y^{(k)}|)^2\\ &= 4((\tau^2+1)(d+1)+(d+\tau^2)\ell^{-1})\ell^{-2}\mathrm{s}(\ell).\end{aligned}$$

Hence, $\widetilde{\mathbf{v}}^{\mathrm{T}}D\widetilde{\mathbf{v}}=\mathscr{O}(\ell)$ whereas $(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})^{\mathrm{T}}A_{rest}(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})=\mathscr{O}(1)$, showing with (54) that (43) holds with $\varepsilon=\mathscr{O}(\ell^{-1})$, and therefore, together with (59), proving the asymptotical sharpness of the estimate.

6. CONCLUSION

We have developed an analysis of aggregation-based two-grid method for SPD linear systems. When the system matrix is diagonally dominant, an upper bound on the convergence factor can be obtained in a purely algebraic way, assessing locally and independently the quality of each aggregate by solving an eigenvalue problem of the size of the aggregate. Our analysis also shows that nodes for which the corresponding row is strongly dominated by its diagonal element can be safely kept outside the coarsening process (see Proposition 3.3).

We have applied our bound to scalar elliptic PDE problems in two dimensions, showing that aggregation-based two-grid methods are robust if

- in the presence of anisotropy, one uses linewise aggregates aligned with the direction of strong coupling;
- in the presence of discontinuities, one avoids mixing inside an aggregate nodes belonging to a strong coefficient region or its boundary with nodes interior to a weak coefficient region.

DOI: 10.1002/nla

Furthermore, we have shown that the bound is asymptotically sharp when a significant part of the domain is regularly covered by box or line aggregates of the same shape.

Note that we have conducted the analysis in two dimensions for the sake of simplicity. The same type of analysis can be developed for three dimensional problems, leading to similar conclusions.

Our results may also have an impact on practical aggregation schemes. Because of the above mentioned sharpness, it is indeed sensible to expect that aggregation methods can be improved by improving aggregates' quality. And because aggregates' quality is cheap to assess, this parameter can effectively be taken into account in the design of aggregation algorithms. For instance, one may *a posteriori* check aggregates' quality and break low quality aggregates into smaller pieces. It is also possible, in a greedy-like approach, to decide wether a node (or a group of nodes) should be added to an aggregate according its impact on the aggregate's quality and/or select the neighboring (sets of) nodes that are the most favorable in this respect. These practical aspects are subject to further research.

ACKNOWLEDGEMENTS

The authors thank all the referees for their valuable comments that helped to improve the presentation of the results.

APPENDIX A

Proof of Theorem 5.1

We first prove (54). Since the diagonal entries of D_d are non-decreasing functions of α_4 and A_d does not depend on this latter, $\lambda_2(D_d^{-1}A_d)$ does not increase with increasing α_4. Hence, setting $C_d = \lim_{\alpha_4\to\infty} D_d^{-1}A_d$, we have

$$\lambda_2(D_d^{-1}A_d) \geqslant \lim_{\alpha_4\to\infty} \lambda_2(D_d^{-1}A_d) = \lambda_2(C_d), \tag{A1}$$

where

$$C_d = \begin{pmatrix} D_r^{-1}A_r & * \\ \mathbf{0}^{\mathrm{T}} & 0 \end{pmatrix}$$

with

$$A_r = \frac{1}{2}\begin{pmatrix} 4\alpha_1 & -2\alpha_1 & -2\alpha_1 \\ -2\alpha_1 & 3\alpha_1+\alpha_2 & \\ -2\alpha_1 & & 3\alpha_1+\alpha_3 \end{pmatrix} \quad\text{and}\quad D_r = 2\begin{pmatrix} 2\alpha_1 & & \\ & \alpha_1+\alpha_2 & \\ & & \alpha_1+\alpha_3 \end{pmatrix}.$$

Hence, $\lambda_2(C_d)$ is the smallest eigenvalue of $D_r^{-1}A_r$. Now, assume without loss of generality that $\alpha_3 \geqslant \alpha_2$ (one may see that they play a symmetric role in the definition of A_d and D_d). Then, setting

$$\widetilde{A}_r = \frac{1}{2}\begin{pmatrix} 4\alpha_1 & -2\alpha_1 & -2\alpha_1 \\ -2\alpha_1 & 3\alpha_1+\alpha_2 & \\ -2\alpha_1 & & 3\alpha_1+\alpha_2 \end{pmatrix} \quad\text{and}\quad \widetilde{D}_r = 2\begin{pmatrix} 2\alpha_1 & & \\ & \alpha_1+\alpha_2 & \\ & & \alpha_1+\alpha_2 \end{pmatrix}$$

we have,

$$\begin{aligned}\lambda_{\min}(D_r^{-1}A_r) &= \min_{\mathbf{v}\in\mathbb{R}^3\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}}A_r\mathbf{v}}{\mathbf{v}^{\mathrm{T}}D_r\mathbf{v}} = \min_{\mathbf{v}\in\mathbb{R}^3\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^{\mathrm{T}}\widetilde{A}_r\mathbf{v}+\frac{1}{2}(\alpha_3-\alpha_2)(\mathbf{v})_3^2}{\mathbf{v}^{\mathrm{T}}\widetilde{D}_r\mathbf{v}+2(\alpha_3-\alpha_2)(\mathbf{v})_3^2} \\ &\geqslant \min(\lambda_{\min}\left(\widetilde{D}_r^{-1}\widetilde{A}_r), \tfrac{1}{4}\right).\end{aligned} \tag{A2}$$

DOI: 10.1002/nla

One may check that the set of eigenvalues of $\widetilde{D}_r^{-1}\widetilde{A}_r$ is

$$\left\{\frac{3\alpha_1+\alpha_2}{4(\alpha_1+\alpha_2)},\frac{3}{8}+\frac{2\alpha_1\pm\sqrt{17\alpha_1^2+14\alpha_1\alpha_2+\alpha_2^2}}{8(\alpha_1+\alpha_2)}\right\}$$

(e.g. by assessing the determinant of $\widetilde{A}_r-\widetilde{\lambda}\widetilde{D}_r$ for all $\widetilde{\lambda}$ belonging to the set[§]). Since $2\alpha_1-\sqrt{17\alpha_1^2+14\alpha_1\alpha_2+\alpha_2^2}\geqslant(2-\sqrt{17})(\alpha_1+\alpha_2)$ holds for α_1, $\alpha_2>0$, the inequality (54) follows.

On the other hand, if $\alpha_1=\alpha_2$ and $\alpha_3=\alpha_4$, the set of eigenvalues of $D_d^{-1}A_d$ is given by

$$\left\{0,\frac{1}{2},\frac{3\alpha_1+\alpha_3}{4(\alpha_1+\alpha_3)},\frac{1}{2}+\frac{3\alpha_1+\alpha_3}{4(\alpha_1+\alpha_3)}\right\},$$

which leads to (55).

To prove (56) we note that, as previously observed, $\lambda_2(D_d^{-1}A_d)$ does not increase with increasing α_4. Since α_1 is the largest coefficient by assumption, setting $\alpha_4=\alpha_1$ gives a worst case estimate. Next, we assume without loss of generality that $\alpha_2\geqslant\alpha_3$ (again, they play a symmetric role). Let then $\widetilde{A}_{0,0}$, $\widetilde{A}_{1,0}$ and $\widetilde{A}_{1,1}$ be the matrices defined via (52) with $\alpha_1=\alpha_4=1$ and the couple (α_2,α_3) given by, respectively, $(0,0)$, $(1,0)$ and $(1,1)$; that is

$$\widetilde{A}_{0,0}=\begin{pmatrix}2&-1&-1&\\-1&\frac{3}{2}&&-\frac{1}{2}\\-1&&\frac{3}{2}&-\frac{1}{2}\\&-\frac{1}{2}&-\frac{1}{2}&1\end{pmatrix},\quad\widetilde{A}_{1,0}=\begin{pmatrix}2&-1&-1&\\-1&2&&-1\\-1&&\frac{3}{2}&-\frac{1}{2}\\&-1&-\frac{1}{2}&\frac{3}{2}\end{pmatrix},$$

$$\widetilde{A}_{1,1}=\begin{pmatrix}2&-1&-1&\\-1&2&&-1\\-1&&2&-1\\&-1&-1&2\end{pmatrix}.$$

Similarly, let $\widetilde{D}_{0,0}$, $\widetilde{D}_{1,0}$ and $\widetilde{D}_{1,1}$ be the matrices defined via (53) with $\alpha_1=\alpha_4=1$ and (α_2,α_3) being, respectively, $(0,0)$, $(1,0)$ and $(1,1)$; that is, $\widetilde{D}_{0,0}=\mathrm{diag}(4\ 2\ 2\ 2)$, $\widetilde{D}_{1,0}=\mathrm{diag}(4\ 4\ 2\ 3)$ and $\widetilde{D}_{1,1}=\mathrm{diag}(4\ 4\ 4\ 4)$. Then,

$$\begin{aligned}A_d&=(\alpha_1-\alpha_2)\widetilde{A}_{0,0}+(\alpha_2-\alpha_3)\widetilde{A}_{1,0}+\alpha_3\widetilde{A}_{1,1},\\D_d&=(\alpha_1-\alpha_2)\widetilde{D}_{0,0}+(\alpha_2-\alpha_3)\widetilde{D}_{1,0}+\alpha_3\widetilde{D}_{1,1}.\end{aligned}$$

Next, using the min-max theorem (e.g. [26, Lemma 3.13]), we have

$$\begin{aligned}\lambda_2(D_d^{-1}A_d)&=\max_{\mathbf{v}\in\mathbb{R}^4\backslash\{\mathbf{0}\}}\min_{\mathbf{w}\perp\mathbf{v}}\frac{\mathbf{w}^{\mathrm{T}}D_d^{-1/2}A_dD_d^{-1/2}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{w}}\\&=\max_{\mathbf{v}\in\mathbb{R}^4\backslash\{\mathbf{0}\}}\min_{\mathbf{z}\perp\mathbf{v}}\frac{\mathbf{z}^{\mathrm{T}}A_d\mathbf{z}}{\mathbf{z}^{\mathrm{T}}D_d\mathbf{z}}\end{aligned}$$

[§] All eigenvalues explicitly given in this proof have been checked with computer algebra.

DOI: 10.1002/nla

$$= \max_{\mathbf{v}\in\mathbb{R}^4\setminus\{\mathbf{0}\}} \min_{\mathbf{z}\perp\mathbf{v}} \frac{(\alpha_1-\alpha_2)\mathbf{z}^{\mathrm{T}}\widetilde{A}_{0,0}\mathbf{z}+(\alpha_2-\alpha_3)\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,0}\mathbf{z}+\alpha_3\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,1}\mathbf{z}}{(\alpha_1-\alpha_2)\mathbf{z}^{\mathrm{T}}\widetilde{D}_{0,0}\mathbf{z}+(\alpha_2-\alpha_3)\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,0}+\alpha_3\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,1}\mathbf{z}}$$

$$\geqslant \max_{\mathbf{v}\in\mathbb{R}^4\setminus\{\mathbf{0}\}} \min\left(\min_{\mathbf{z}\perp\mathbf{v}}\frac{\mathbf{z}^{\mathrm{T}}\widetilde{A}_{0,0}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\widetilde{D}_{0,0}\mathbf{z}},\ \min_{\mathbf{z}\perp\mathbf{v}}\frac{\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,0}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,0}\mathbf{z}},\ \min_{\mathbf{z}\perp\mathbf{v}}\frac{\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,1}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,1}\mathbf{z}}\right).$$

Hence,

$$\lambda_2(D_d^{-1}A_d)\geqslant\min\left(\min_{\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4}\frac{\mathbf{z}^{\mathrm{T}}\widetilde{A}_{0,0}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\widetilde{D}_{0,0}\mathbf{z}},\ \min_{\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4}\frac{\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,0}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,0}\mathbf{z}},\ \min_{\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4}\frac{\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,1}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,1}\mathbf{z}}\right), \tag{A3}$$

where the second term in the minimum further becomes, since $\widetilde{D}_{1,0}^{1/2}\mathbf{1}_4\in\mathcal{N}(\widetilde{D}_{1,0}^{-1/2}\widetilde{A}_{1,0}\widetilde{D}_{1,0}^{-1/2})$,

$$\min_{\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4}\frac{\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,0}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,0}\mathbf{z}}=\lambda_2(\widetilde{D}_{1,0}^{-1}\widetilde{A}_{1,0})=\beta.$$

Therefore, the proof of (56) is done if we show that the second term in (A3) is the smallest. For this, we note that the vector $z=(64\ -34\ 33\ -62)^{\mathrm{T}}$ is orthogonal to $\widetilde{D}_{1,0}\mathbf{1}_4=(4\ 4\ 2\ 3)^{\mathrm{T}}$ and that $\mathbf{z}^{\mathrm{T}}\widetilde{A}_{1,0}\mathbf{z}=15861.5$ with $\mathbf{z}^{\mathrm{T}}\widetilde{D}_{1,0}\mathbf{z}=34718$. Hence, the second term is smaller than 0.46. Furthermore, the first term in (A3) is larger than 0.46, as can be concluded from positive definiteness of

$$\widetilde{A}_{0,0}+\widetilde{D}_{1,0}\mathbf{1}_4(\widetilde{D}_{1,0}\mathbf{1}_4)^{\mathrm{T}}-0.46\widetilde{D}_{0,0}=\begin{pmatrix}16.16 & 15 & 7 & 12\\ 15 & 16.58 & 8 & 11.5\\ 7 & 8 & 4.58 & 5.5\\ 12 & 11.5 & 5.5 & 9.08\end{pmatrix},$$

which implies $\mathbf{z}^{\mathrm{T}}\widetilde{A}_{0,0}\mathbf{z}-0.46\mathbf{z}^{\mathrm{T}}\widetilde{D}_{0,0}\mathbf{z}\geqslant0$ for any $\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4$. Similarly, the third term in (A3) is larger than 0.46 since

$$\widetilde{A}_{1,1}+\widetilde{D}_{1,0}\mathbf{1}_4(\widetilde{D}_{1,0}\mathbf{1}_4)^{\mathrm{T}}-0.46\widetilde{D}_{1,1}=\begin{pmatrix}16.16 & 15 & 7 & 12\\ 15 & 16.16 & 8 & 11\\ 7 & 8 & 4.16 & 5\\ 12 & 11 & 5 & 9.16\end{pmatrix}$$

is also positive definite. The positive definiteness can be proved, for instance, checking that the determinants of upper left 1×1, 2×2, 3×3 and 4×4 blocks are positive.

Eventually, we prove (57). If $\alpha_1=\alpha_2$ and $\alpha_3=\alpha_4$, the inequality is already proved in (55). If $\alpha_2=\alpha_3=\alpha_4$, taking $\widetilde{D}_d=\mathrm{diag}(4\alpha_1\ 2(\alpha_1+\alpha_2)\ 2(\alpha_1+\alpha_2)\ 2(\alpha_1+\alpha_2))$ we have

$$\lambda_k(D_d^{-1}A_d)\geqslant\lambda_k(\widetilde{D}_d^{-1}A_d)\in\left\{0,\frac{1}{2},\frac{3\alpha_1+\alpha_2}{4(\alpha_1+\alpha_2)},\frac{1}{2}+\frac{3\alpha_1+\alpha_2}{4(\alpha_1+\alpha_2)}\right\},$$

which leads to the same conclusions. If $\alpha_1=\alpha_2=\alpha_3$, the set of eigenvalues of $D_d^{-1}A_d$ is given by

$$\left\{0,\frac{1}{2},\frac{1}{2}+\frac{4\alpha_1\pm\sqrt{10\alpha_1^2+4\alpha_1\alpha_4+2\alpha_4^2}}{4(3\alpha_1+\alpha_4)}\right\}$$

and, since $\alpha_1\geqslant\alpha_4$ implies $10\alpha_1^2+4\alpha_1\alpha_4+2\alpha_4^2\leqslant16\alpha_1^2$, the inequality (57) follows. □

REFERENCES

1. Trottenberg U, Oosterlee CW, Schüller A. *Multigrid*. Academic Press: London, 2001.
2. Hackbusch W. *Multi-grid Methods and Applications*. Springer: Berlin, 1985.
3. Wesseling P. *An Introduction to Multigrid Methods*. Wiley: Chichester, 1992.
4. Braess D. Towards algebraic multigrid for elliptic problems of second order. *Computing* 1995; **55**:379–393.
5. Bulgakov VE. Multi-level iterative technique and aggregation concept with semi-analytical preconditioning for solving boundary-value problems. *Communications in Numerical Methods in Engineering* 1993; **9**:649–657.
6. Stüben K. An introduction to algebraic multigrid. In *Multigrid*, Trottenberg U, Oosterlee CW, Schüller A (eds). Academic Press: London, 2001. Appendix A.
7. Wienands R, Oosterlee CW. On three-grid Fourier analysis for multigrid. *SIAM Journal on Scientific Computing* 2001; **23**:651–671.
8. Vaněk P, Mandel J, Brezina M. Algebraic multigrid based on smoothed aggregation for second and fourth order problems. *Computing* 1996; **56**:179–196.
9. Vaněk P, Brezina M, Tezaur R. Two-grid method for linear elasticity on unstructured meshes. *SIAM Journal on Scientific Computing* 1999; **21**:900–923.
10. Muresan AC, Notay Y. Analysis of aggregation-based multigrid. *SIAM Journal on Scientific Computing* 2008; **30**:1082–1103.
11. Notay Y. An aggregation-based algebraic multigrid method. *Electronic Transactions on Numerical Analysis* 2010; **37**:123–146.
12. Notay Y, Vassilevski PS. Recursive Krylov-based multigrid cycles. *Numerical Linear Algebra with Applications* 2008; **15**:473–487.
13. Kim H, Xu J, Zikatanov L. A multigrid method based on graph matching for convection–diffusion equations. *Numerical Linear Algebra with Applications* 2003; **10**:181–195.
14. Emans M. Performance of parallel AMG-preconditioners in CFD-codes for weakly compressible flow. *Parallel Computing* 2010; **36**:326–338.
15. Brandt A, McCormick SF, Ruge JW. Algebraic multigrid (amg) for sparse matrix equations. In *Sparsity and its Application*, Evans DJ (ed.). Cambridge University Press: Cambridge, 1984; 257–284.
16. Ruge JW, Stüben K. Algebraic multigrid (AMG). In *Multigrid Methods*, McCormick SF (ed.). Frontiers in Applied Mathematics, vol. 3. SIAM: Philadelphia, PA, 1987; 73–130.
17. Stüben K. Algebraic multigrid (AMG): experiences and comparisons. *Applied Mathematics and Computation* 1983; **13**:419–452.
18. Brezina M, Cleary AJ, Falgout RD, Henson VE, Jones JE, Manteuffel TA, McCormick SF, Ruge JW. Algebraic multigrid based on element interpolation (AMGe). *SIAM Journal on Scientific Computing* 2000; **22**:1570–1592.
19. Chartier T, Falgout RD, Henson VE, Jones J, Manteuffel T, McCormick S, Ruge J, Vassilevski PS. Spectral AMGe (ρAMGe). *SIAM Journal on Scientific Computing* 2004; **25**:1–26.
20. Jones JE, Vassilevski PS. AMGE based on element agglomeration. *SIAM Journal on Scientific Computing* 2001; **23**:109–133.
21. Kolev TV, Vassilevski PS. AMG by element agglomeration and constrained energy minimisation interpolation. *Numerical Linear Algebra with Applications* 2006; **13**:771–788.
22. Falgout RD, Vassilevski PS, Zikatanov LT. On two-grid convergence estimates. *Numerical Linear Algebra with Applications* 2005; **12**:471–494.
23. Vassilevski PS. *Multilevel Block Factorization Preconditioners*. Springer: New York, 2008.
24. Notay Y. Algebraic analysis of two-grid methods: the nonsymmetric case. *Numerical Linear Algebra with Applications* 2010; **17**:73–97.
25. Nakamura S. *Computational Methods in Engineering and Science*. Wiley: New York, 1977.
26. Axelsson O. *Iterative Solution Methods*. Cambridge University Press: Cambridge, 1994.

DOI: 10.1002/nla