Article:
Chen, M-H and Greenbaum, A (2015) Analysis of an aggregation-based algebraic two-grid method for a rotated anisotropic diffusion problem. Numerical Linear Algebra with Applications, 22 (4). pp. 681-701. ISSN 1070-5325

https://doi.org/10.1002/nla.1980

Analysis of an Aggregation-Based Algebraic Two-grid Method for a Rotated Anisotropic Diffusion Problem

Meng-Huo Chen* Anne Greenbaum†

January 2, 2015

Abstract

A two-grid convergence analysis based on the paper [*Algebraic analysis of aggregation-based multigrid*, by A. Napov and Y. Notay, Numer. Lin. Alg. Appl. 18 (2011), pp. 539-564] is derived for various aggregation schemes applied to a finite element discretization of a rotated anisotropic diffusion equation. As expected, it is shown that the best aggregation scheme is one in which aggregates are aligned with the anisotropy. In practice, however, this is not what automatic aggregation procedures do. We suggest approaches for determining appropriate aggregates based on eigenvectors associated with small eigenvalues of a block splitting matrix, or based on minimizing a quantity related to the spectral radius of the iteration matrix.

1 Introduction

Recently aggregation-based algebraic multigrid methods with piecewise constant prolongation have received considerable attention. See, for instance, [8, 11]. Although these methods may require extra Krylov space iterations on coarse grids in order to perform well in a multigrid setting [14, 10], their relative simplicity compared to other algebraic multigrid methods gives them a number of advantages. They require less setup time than standard algebraic multigrid methods and maintain better sparsity in the "coarse grid" matrices. They also maintain other important properties of the original matrix such as the M-matrix property [6, Theorem 3.6].

In [9] a powerful method was derived for analyzing the convergence of two-grid aggregation methods based on the quality of individual aggregates. While the two-grid convergence rates do *not* carry over to the multigrid V-cycle, it was shown in [14, 10] that similar convergence rates could be obtained in a multigrid setting by replacing the V-cycle with a K-cycle. Hence the two-grid convergence analysis is still an important step in understanding the behavior of these methods.

*University of Washington, Dept. of Applied Mathematics, Box 352420, Seattle, WA 98195. This work was supportedd in part by NSF grant DMS-1210886.

†University of Washington, Dept. of Applied Mathematics, Box 352420, Seattle, WA 98195. This work was supported in part by NSF grant DMS-1210886.

In this paper we apply the analysis techniques in [9] to a finite element discretization of a rotated anisotropic diffusion equation with homogeneous Dirichlet boundary conditions:

$$-(\epsilon\cos^2\theta+\sin^2\theta)\frac{\partial^2 u}{\partial x^2}-2(1-\epsilon)\cos\theta\sin\theta\frac{\partial^2 u}{\partial x\partial y}-(\cos^2\theta+\epsilon\sin^2\theta)\frac{\partial^2 u}{\partial y^2}=f,$$

$$\text{on } [0,1]\times[0,1],\quad u(x,0)=u(x,1)=y(0,y)=u(1,y)=0. \tag{1}$$

This corresponds to a problem of the form $-\epsilon u_{\xi\xi}-u_{\eta\eta}=f$ in a (ξ,η) coordinate system that can be obtained by rotating the (x,y) coordinate system through angle θ. We will always assume that $\epsilon\in[0,1]$ since for $\epsilon>1$, equation (1) corresponds to $-u_{\xi\xi}-\epsilon^{-1}u_{\eta\eta}=\epsilon^{-1}f$, where the direction of stronger coupling is just rotated by $\pi/2$. We will deal with angles θ between $-\pi/2$ and $\pi/2$. Problems of this sort can be a challenge for aggregation-based algebraic multigrid methods if the direction of anisotropy is not aligned with the grid lines. Numerical experiments for the special case $\theta=\pi/4$ show that if the aggregates can be chosen to align with the direction of anisotropy, then grid-size independent and ϵ-independent convergence rates can be obtained with a two-grid method. In this paper we use the analysis in [9] to prove this. We suggest two possible strategies for automatically recognizing such cases and assembling the correct aggregates. We demonstrate that these strategies can result in smaller spectral radii for the iteration matrix in a two-grid method and that they can also be effective in a true multigrid K-cycle setting.

Problems of this type have been studied elsewhere. Different coarsening strategies were proposed, for instance, in [3, 15]. The interplay between the relaxation method and the coarsening strategy has also been explored, for example, in [2, 7]. Most of this work has been in the context of classical algebraic multigrid (where restriction and prolongation operators are more complicated) rather than aggregation-based multigrid. In [16], a new "smoothed" aggregation approach was shown to yield mesh-size independent (or nearly independent) convergence rates for problems of the form (1), without changing the coarsening strategy. Here we work with the simplest, unsmoothed aggregation procedure and try to determine if it can find aggregates for which even simple smoothers such as the damped Jacobi method can be effective. Another work that is closely related to ours is [4], where a different technique was used to show that if pairs of nodes were aggregated in or close to the direction of anisotropy, then if two of the three parameters – ϵ, θ, and the grid size – were held fixed, the convergence rate of a two-grid method using Richardson iteration as a smoother could be bounded independent of the third parameter.

2 Two-grid Convergence Analysis

Let A be an N by N symmetric positive definite (SPD) matrix and suppose we wish to solve the linear system $Ax=b$. Let $N_C << N$ be given and define an N by N_C

prolongation matrix P by

$$P = \begin{pmatrix} \mathbf{0}_{m_0} & \mathbf{0}_{m_0} & \dots & \mathbf{0}_{m_0} \\ \mathbf{1}_{m_1} & & & \\ & \mathbf{1}_{m_2} & & \\ & & \ddots & \\ & & & \mathbf{1}_{m_{N_C}} \end{pmatrix},$$

where $\mathbf{1}_{m_j}$ is a vector of 1's whose length is the number m_j of variables corresponding to the jth aggregate. The top block row of 0's corresponds to variables that are not involved in the aggregation procedure (because, for instance, their rows might be strongly diagonally dominant and hence they can be dealt with efficiently by a fine grid relaxation). If there are m_0 such variables then $\sum_{j=0}^{N_C} m_j = N$. Note that geometrically P corresponds to *piecewise constant* interpolation from the collection of aggregates to the collection of fine grid points. Define an N_C by N_C "coarse grid matrix" A_C by

$$A_C = P^T A P.$$

Let M_1 and M_2 denote pre- and post-smoothing matrices. For example, with a weighted Jacobi method $M_1 = M_2 = \omega^{-1}\text{diag}(A)$; with a symmetrized Gauss-Seidel method $M_1 = M_2^T = \text{lower triangle}(A)$. The solution procedure will be to first perform ν_1 relaxation steps with splitting M_1, so that, starting with $x^{(k)} \equiv x^{(k,0)}$, we obtain new approximations $x^{(k,j)}$ via $x^{(k,j)} = x^{(k,j-1)} + M_1^{-1}(b - Ax^{(k,j-1)})$, $j = 1, \dots, \nu_1$. The error $e^{(k,j)} \equiv A^{-1}b - x^{(k,j)}$ then satisfies $e^{(k,j)} = (I - M_1^{-1}A)e^{(k,j-1)}$ so that $e^{(k,\nu_1)} = (I - M_1^{-1}A)^{\nu_1} e^{(k,0)}$. Next a "coarse grid correction" is performed: Restrict the residual $r^{(k,\nu_1)} \equiv b - Ax^{(k,\nu_1)}$ to the coarse grid by forming $P^T r^{(k,\nu_1)}$, solve for δ in the linear system $A_C\delta = P^T r^{(k,\nu_1)}$, form the longer vector $P\delta$ and add to $x^{(k,\nu_1)}$. The result is $x^{(k,\nu)} = x^{(k,\nu_1)} + PA_C^{-1}P^T r^{(k,\nu_1)}$ so that the error now satisfies $e^{(k,\nu)} = (I - PA_C^{-1}P^T A)e^{(k,\nu_1)} = (I - PA_C^{-1}P^T A)(I - M_1^{-1}A)^{\nu_1} e^{(k,0)}$. Finally, perform ν_2 post-smoothing steps using the splitting M_2 to obtain the new iterate $x^{(k+1)}$. The resulting error is

$$e^{(k+1)} = E_{TG} e^{(k)}, \quad E_{TG} = (I - M_2^{-1}A)^{\nu_2}(I - PA_C^{-1}P^T A)(I - M_1^{-1}A)^{\nu_1}. \tag{2}$$

The matrix E_{TG} is the two-grid iteration matrix.

The *asymptotic* convergence rate of the method is determined by the *spectral radius* $\rho(E_{TG})$. In general, this is **not** the relevant quantity for studying short-time behavior of the iteration (and, hopefully, short-time behavior is all that will be observed!). In some cases, however, the spectral radius also gives a measure of the amount by which the A-norm of the error is reduced at each step. The A-norm of a vector v is $\|v\|_A \equiv \langle Av, v\rangle^{1/2} = \|A^{1/2}v\|$, where $\|\cdot\|$ without a subscript denotes the 2-norm. From (2) it follows that $A^{1/2}e^{(k+1)} = A^{1/2}E_{TG}A^{-1/2}(A^{1/2}e^{(k)})$ and hence

$$\|e^{(k+1)}\|_A \le \|A^{1/2}E_{TG}A^{-1/2}\| \cdot \|e^{(k)}\|_A.$$

The matrix $A^{1/2}E_{TG}A^{-1/2}$ has the same eigenvalues as E_{TG} so that *if* this matrix is symmetric then its norm is $\rho(E_{TG})$. The matrix $A^{1/2}E_{TG}A^{-1/2}$ can be written in the form

$$(I - A^{1/2}M_2^{-1}A^{1/2})^{\nu_2}(I - A^{1/2}PA_C^{-1}P^T A^{1/2})(I - A^{1/2}M_1^{-1}A^{1/2})^{\nu_1}.$$

Hence if $\nu_1 = \nu_2$ and $M_1 = M_2^T$, then this matrix is symmetric and the A-norm of the error is reduced at each step by at least the factor $\rho(E_{TG})$. In this paper we will study the spectral radius $\rho(E_{TG})$, whether or not these symmetry conditions hold, because that is the quantity that is analyzed in [9].

Let an N by N matrix X be defined by

$$I - X^{-1}A = (I - M_1^{-1}A)^{\nu_1}(I - M_2^{-1}A)^{\nu_2}. \tag{3}$$

For example, if one post-smoothing only is done so that $\nu_1 = 0$ and $\nu_2 = 1$, then $X = M_2$. If one step of pre-smoothing and one of post-smoothing is done then $\nu_1 = \nu_2 = 1$ and

$$I - X^{-1}A = I - M_1^{-1}A - M_2^{-1}A + M_1^{-1}AM_2^{-1}A,$$

so that $X^{-1} = M_1^{-1} + M_2^{-1} - M_1^{-1}AM_2^{-1} = M_1^{-1}(M_2 + M_1 - A)M_2^{-1}$. We always assume that this matrix is invertible and, in fact, SPD. For example, if M_1 is the lower triangular part of the SPD matrix A and M_2 is the upper triangular part (M_1^T), then $X = M_2(\mathrm{diag}(A))^{-1}M_1$.

Just as equation (2) involves the "coarse grid" approximation of the matrix $A^{-1}A = I$ through the term $PA_C^{-1}P^TA$, where $A_C = P^TAP$, there will be other matrices for which this sort of term is relevant. For any N by N matrix Y, define

$$\pi_Y \equiv P(P^TYP)^{-1}P^TY \equiv PY_C^{-1}P^TY. \tag{4}$$

The following theorem is proved in [9] and is a slight generalization of [5, Theorem 4.2]:

Theorem 3.1 (Napov and Notay [9]). Let A be an N by N SPD matrix and let P be an N by N_C matrix of rank $N_C < N$. Let M_1, ν_1, M_2, ν_2 be such that X, defined in (3), is SPD and let E_{TG} be the two-grid iteration matrix defined in (2). Then

$$\rho(E_{TG}) \le \max\left\{\lambda_{max}(X^{-1}A) - 1,\ 1 - \frac{1}{\mu_X}\right\}, \tag{5}$$

where $\lambda_{max}(\cdot)$ denotes the (algebraically) largest eigenvalue, and

$$\mu_X = \lambda_{max}(A^{-1}X(I - \pi_X)).$$

Moreover, for any N by N SPD matrix D,

$$\mu_X \le \lambda_{max}(D^{-1}X)\cdot\mu_D, \quad \mu_D \equiv \lambda_{max}(A^{-1}D(I - \pi_D)). \tag{6}$$

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

$$\rho(E_{TG}) = 1 - \frac{1}{\mu_X}, \tag{7}$$

and $\mu_X \le \omega^{-1}\mu_D$, with equality if $\nu_1 + \nu_2 = 1$.

In this paper, we will apply this theorem in the case where $M_1 = M_2 = \omega^{-1}D$ and $D = \mathrm{diag}(A)$. In that case, if $\nu_1 + \nu_2 = 1$, then $X = \omega^{-1}D$ and the first factor in the

bound (6) on μ_X is just ω^{-1}. The second, $\mu_D = \lambda_{max}(A^{-1}D(I - \pi_D))$ gives a measure of how effectively eigenvectors associated with the large eigenvalues of A^{-1} are damped by the coarse grid correction. If $\nu_1 = \nu_2 = 1$, then $X = \omega^{-1}D(2D - \omega A)^{-1}D$ and $\lambda_{max}(D^{-1}X) = \omega^{-1}\lambda_{max}((2D - \omega A)^{-1}D)$. Typically, the matrix $2D - \omega A$ is strongly diagonally dominant so that it is easy to estimate $\lambda_{max}(D^{-1}X)$. For example, if A has 2's on its main diagonal and -1's on the first sub- and super-diagonal, as in a 1D problem for the Laplacian, then $D = 2I$, $\lambda_{max}(D^{-1}A) < 2$, so $\omega^{-1} = 2$ is a reasonable choice, and then the matrix $2D - \omega A$ has 3's on its main diagonal and 1/2's on the first sub- and super-diagonal. Hence by Gerschgorin's theorem the smallest eigenvalue of this matrix is greater than or equal to 2; that is the largest eigenvalue of the inverse matrix is less than or equal to 1/2, and the bound on $\lambda_{max}(D^{-1}X)$ is $\omega^{-1} = 2$.

In general, it seems difficult to determine the quantity μ_D on the right-hand side of (6), so it is not clear that this theorem has helped much in the goal of estimating the spectral radius $\rho(E_{TG})$. However, the next theorem in [9] explains how to estimate the quantity μ_D by studying the individual aggregates.

Assume that A can be written in the form $A = A_b + A_r$, where A_b and A_r are both symmetric and nonnegative definite and A_b is block diagonal, with blocksizes equal to the corresponding aggregate size:

$$A_b = \begin{pmatrix} A^{(0)} & & & \\ & A^{(1)} & & \\ & & \ddots & \\ & & & A^{(N_C)} \end{pmatrix}, \tag{8}$$

where $A^{(k)}$, $k = 0, 1, \ldots, N_C$, is of size m_k by m_k.

If the matrix A is diagonally dominant, then one way to create this splitting is to take the diagonal blocks of A_b to be the corresponding diagonal blocks of A, with the sum of absolute values of the remaining entries in each row or column of A subtracted from the corresponding diagonal entry of A_b. For example, if

$$A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \tag{9}$$

and successive pairwise aggregation is used, then a possible choice for A_b and A_r is

$$A_b = \left(\begin{array}{cc|cc|cc} 2 & -1 & & & & \\ -1 & 1 & & & & \\ \hline & & 1 & -1 & & \\ & & -1 & 1 & & \\ \hline & & & & 1 & -1 \\ & & & & -1 & 2 \end{array}\right), \quad A_r = \left(\begin{array}{cc|cc|cc} & & & & & \\ & 1 & -1 & & & \\ \hline & -1 & 1 & & & \\ & & & 1 & -1 & \\ \hline & & & -1 & 1 & \\ & & & & & \end{array}\right). \tag{10}$$

In this way, both A_b and A_r are weakly diagonally dominant, hence nonnegative definite.

If we choose to omit the first and last variables from the aggregation procedure and reorder rows and columns so that these two variables are first, followed by variables 2 and 3 and then 4 and 5, the matrix A takes the form

$$A = \begin{pmatrix} 2 & 0 & -1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & -1 \\ -1 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & -1 & 0 & 0 & -1 & 2 \end{pmatrix}.$$

and we can split this as

$$A_b = \left(\begin{array}{cc|cc|cc} 1 & & & & & \\ & 1 & & & & \\ \hline & & 1 & -1 & & \\ & & -1 & 1 & & \\ \hline & & & & 1 & -1 \\ & & & & -1 & 1 \end{array}\right), \quad A_r = \left(\begin{array}{cc|cc|cc} 1 & & -1 & & & \\ & 1 & & & & -1 \\ \hline -1 & & 1 & & & \\ & & & 1 & -1 & \\ \hline & & -1 & & 1 & \\ & -1 & & & & 1 \end{array}\right).$$

Other such splittings are possible as well. In Section 4 we describe a splitting that is appropriate for finite element discretizations, even when the matrix is not diagonally dominant. Once the splitting is set, the following theorem from [9] enables one to estimate the 'global' parameter μ_D in (6) by 'local' quantities $\mu_D^{(k)}$ associated with each aggregate k.

Theorem 3.2 (Napov and Notay [9]). Using the notation of Theorem 3.1, let $A = A_b + A_r$ be a splitting of the SPD matrix A such that A_b and A_r are both nonnegative definite and A_b has the form (8). Assume that D in Theorem 3.1 also has the block diagonal form:

$$D = \begin{pmatrix} D^{(0)} & & & \\ & D^{(1)} & & \\ & & \ddots & \\ & & & D^{(N_C)} \end{pmatrix},$$

where each block $D^{(k)}$ is m_k by m_k, $k = 0, \ldots, N_C$. Define

$$\mu_D^{(0)} = \begin{cases} 0 & \text{if } m_0 = 0 \\ \sup_{v \in \mathbb{R}^{m_0} \setminus \mathcal{N}(A^{(0)})} \frac{v^T D^{(0)} v}{v^T A^{(0)} v} & \text{if } m_0 > 0 \end{cases},$$

and for $k = 1, \ldots, N_C$,

$$\mu_D^{(k)} = \begin{cases} 0 & \text{if } m_k = 1 \\ \sup_{v \in \mathbb{R}^{m_k} \setminus \mathcal{N}(A^{(k)})} \frac{v^T D^{(k)} (I - \pi_D^{(k)}) v}{v^T A^{(k)} v} & \text{if } m_k > 1 \end{cases}, \tag{11}$$

where

$$\pi_D^{(k)} = p^{(k)}({p^{(k)}}^T D^{(k)} p^{(k)})^{-1} {p^{(k)}}^T D^{(k)},$$

and $p^{(k)}$ is the vector of 1's in column k of P. Then

$$\mu_D \leq \max_{k=0,\ldots,N_C} \mu_D^{(k)}.$$

Note that if $m_k > 1$ then $\mathcal{N}(A^{(k)})$ must be a subset of $\mathcal{N}(D^{(k)}(I - \pi_D^{(k)})) = \text{span}\{p^{(k)}\}$, in order for this theorem to give useful information. Otherwise $\mu_D^{(k)} = \infty$ in (11). If $\mathcal{N}(A^{(k)})$ is a subset of $\text{span}\{p^{(k)}\}$, then expression (11) can be replaced by

$$\mu_D^{(k)} = \begin{cases} 0 & \text{if } m_k = 1 \\ \sup_{v \in \mathcal{R}(A^{(k)}) \setminus \{0\}} \frac{v^T D^{(k)} (I - \pi_D^{(k)}) v}{v^T A^{(k)} v} & \text{if } m_k > 1 \end{cases}, \tag{12}$$

In example (9-10), taking $D = 2I$ and each $p^{(k)} = (1,1)^T$, we find

$$\pi_D^{(k)} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad k = 1, 2, 3,$$

so that

$$D^{(k)}(I - \pi_D^{(k)}) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad k = 1, 2, 3.$$

For $k = 1$, $A^{(1)}$ is nonsingular, so $\mu_D^{(1)}$ is the largest eigenvalue of ${A^{(1)}}^{-1}(D^{(1)}(I - \pi_D^{(1)}))$, which is easily seen to be

$$\lambda_{max} \left[\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right] = 1.$$

Similarly, $\mu_D^{(3)} = 1$. For $k = 2$, $A^{(2)}$ is singular but its null space is the span of $p^{(2)}$ and its range is the span of $(1, -1)^T$. In fact $A^{(2)} = D^{(2)}(I - \pi_D^{(2)})$, so it is clear that the quantity in (12) is again 1. Thus in this example the bound from Theorem 3.2 is $\mu_D \leq 1$. Combining this with the bound on $\lambda_{max}(D^{-1}X)$ when $\nu_1 + \nu_2 = 1$ or $\nu_1 = \nu_2 = 1$, we obtain the estimate $\mu_X \leq 2$. Hence from Theorem 3.1,

$$\rho(E_{TG}) \leq \frac{1}{2}.$$

In (9-10), we took A to be a 6 by 6 matrix, but if we made it much larger, the blocks not connected to the two end points would be identical to $A^{(2)}$, so the theorems would still give the bound $\rho(E_{TG}) \leq 1/2$. In fact, this *is* the limit as $N \to \infty$ of the spectral radius of the two-grid iteration matrix when A is an N by N matrix with 2's on its main diagonal and -1's on the first sub- and super-diagonals. In this case, the estimate is perfect! The optimality of the upper bound also could have been determined from a lower bound on μ_D derived in [9, Theorem 4.1], which will be discussed further later on.

3 Finite Element Discretization

Problem (1) can be solved using a piecewise bilinear finite element approximation on a square grid with uniform spacing h in each direction. Letting

$$\begin{aligned} a &= \epsilon \cos^2\theta + \sin^2\theta \\ b &= (1-\epsilon)\cos\theta\sin\theta \\ c &= \cos^2\theta + \epsilon\sin^2\theta, \end{aligned}$$

equation (1) can be written in the form

$$-\nabla \cdot \begin{pmatrix} a & b \\ b & c \end{pmatrix} \nabla u = f.$$

Letting $\phi_j(x,y)$ denote the standard bilinear basis function with value 1 at node j and 0 at all other nodes, and writing the approximate solution $u(x,y)$ as $\sum_{j=1}^{N} u_j \phi_j(x,y)$, the coefficients u_j can be determined by solving the equations

$$-\sum_{j=1}^{N} u_j \int_0^1 \int_0^1 (\nabla\phi_i) \cdot \left[\begin{pmatrix} a & b \\ b & c \end{pmatrix} \nabla\phi_j \right] dx\,dy = \int_0^1 \int_0^1 f\phi_i \, dx\,dy, \quad i = 1, \ldots, N.$$

The global stiffness matrix can be assembled from individual element matrices, which have the form

$$\frac{1}{6} \begin{pmatrix} 2a+3b+2c & -2a+c & a-2c & -a-3b-c \\ -2a+c & 2a-3b+2c & -a+3b-c & a-2c \\ a-2c & -a+3b-c & 2a-3b+2c & -2a+c \\ -a-3b-c & a-2c & -2a+c & 2a+3b+2c \end{pmatrix}.$$

When these are assembled into a global stiffness matrix (and multiplied by 6), the stencil of the finite element discretization becomes

$$\begin{bmatrix} -a+3b-c & 2(a-2c) & -a-3b-c \\ 2(-2a+c) & 8(a+c) & 2(-2a+c) \\ -a-3b-c & 2(a-2c) & -a+3b-c \end{bmatrix}. \tag{13}$$

For $\theta = \pi/4$, for example, the stencil is

$$\begin{bmatrix} \frac{1}{2} - \frac{5}{2}\epsilon & -1-\epsilon & -\frac{5}{2} + \frac{1}{2}\epsilon \\ -1-\epsilon & 8+8\epsilon & -1-\epsilon \\ -\frac{5}{2} + \frac{1}{2}\epsilon & -1-\epsilon & \frac{1}{2} - \frac{5}{2}\epsilon \end{bmatrix}. \tag{14}$$

4 The Matrix Splitting

To form the matrix splitting $A = A_b + A_r$, we will use a method suggested to us by Y. Notay [12]. We will form "potential aggregates" from 4 by 4 squares of nodes, as pictured in Figure 1.

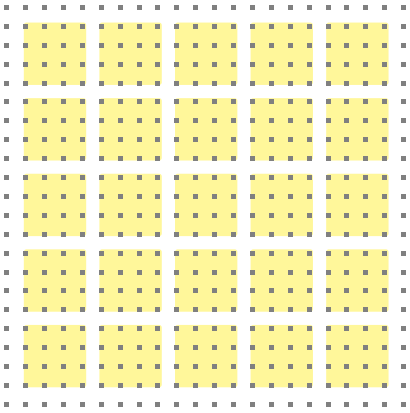


Figure 1: Potential aggregates on a grid of 22×22 interior nodes.

The nodes that are *not* included in a potential aggregate are ones whose rows are strongly diagonally dominant because they couple to Dirichlet boundary points; these make up the block $A^{(0)}$. The other blocks in A_b are then 16×16. Element matrices corresponding to each of the nine elements in the potential aggregate will be assembled to form a block of A_b, and the remaining element matrices will be assembled to form A_r. Since each of the element matrices is nonnegative definite, the matrices A_b and A_r will be as well. Note that each block of A_b (except $A^{(0)}$) is the finite element discretization of a Neumann problem on the potential aggregate, hence has null space consisting of the constant vectors (assuming $0 < \epsilon \leq 1$ in (1)). Note also that if $\epsilon = 0$ in (1) then this actually corresponds to a 1D problem: $u_{\eta\eta} = f$. A vector that represents a function u that satisfies $u_{\eta\eta} = 0$ and is constant in the ξ direction will also lie in the null space of $A^{(k)}$, and when $\epsilon > 0$ is small, this will be a near null vector for $A^{(k)}$.

Actual aggregates will be formed from subsets of the nodes in the potential aggregates. For example, when $\theta = \pi/4$, it is reasonable to aggregate points along diagonal lines, as pictured in Figure 2.

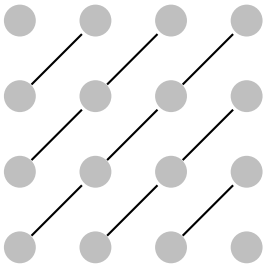


Figure 2: Diagonal aggregates.

This means that the structure is slightly different from that described in [9], where it was assumed that the aggregate size was the same as the block size in A_b.

For example, with the diagonal aggregates pictured in Figure 2, if points within a potential aggregate are numbered first along the main diagonal consisting of 4 nodes, then along the diagonal just below and the one just above, each consisting of 3 nodes,

then along the diagonals two spaces below and above the main one, etc., then the transpose of the block of P corresponding to the kth potential aggregate is

$$P^{(k)^T} = \begin{pmatrix} 1 & 1 & 1 & 1 & & & & & & & & & & & & \\ & & & & 1 & 1 & 1 & & & & & & & & & \\ & & & & & & & 1 & 1 & 1 & & & & & & \\ & & & & & & & & & & 1 & 1 & & & & \\ & & & & & & & & & & & & 1 & 1 & & \\ & & & & & & & & & & & & & & 1 & \\ & & & & & & & & & & & & & & & 1 \end{pmatrix} \tag{15}$$

Thus each $P^{(k)}$ is a 16 by 7 matrix and $\pi_D^{(k)}$ is a 16 by 16 matrix. The block size in A_b is 16 by 16, but the individual aggregate sizes range from 1 to 4. This requires a minor change in the statement of Theorem 3.2.

Theorem 3.2'. Using the notation of Theorem 3.1, let $A = A_b + A_r$ be a splitting of the SPD matrix A such that A_b and A_r are both nonnegative definite and A_b has the form

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

where each block $A^{(k)}$ is s_k by s_k, $k = 0, \ldots, N_b$. Assume that D in Theorem 3.1 also has the block diagonal form

$$D = \begin{pmatrix} D^{(0)} & & & \\ & D^{(1)} & & \\ & & \ddots & \\ & & & D^{(N_b)} \end{pmatrix},$$

where each block is s_k by s_k. Define

$$\mu_D^{(0)} = \begin{cases} 0 & \text{if } s_0 = 0 \\ \sup_{v \in \mathbb{R}^{s_0} \setminus \mathcal{N}(A^{(0)})} \frac{v^T D^{(0)} v}{v^T A^{(0)} v} & \text{if } s_0 > 0 \end{cases}, \tag{16}$$

and for $k = 1, \ldots, N_b$,

$$\mu_D^{(k)} = \begin{cases} 0 & \text{if } s_k = 1 \\ \sup_{v \in \mathbb{R}^{s_k} \setminus \mathcal{N}(A^{(k)})} \frac{v^T D^{(k)} (I - \pi_D^{(k)}) v}{v^T A^{(k)} v} & \text{if } s_k > 1 \end{cases}, \tag{17}$$

where

$$\pi_D^{(k)} = P^{(k)} (P^{(k)^T} D^{(k)} P^{(k)})^{-1} P^{(k)^T} D^{(k)},$$

and $P^{(k)}$ represents the block of P corresponding to the kth potential aggregate. Then

$$\mu_D \leq \max_{k=0,\ldots,N_b} \mu_D^{(k)}. \tag{18}$$

In order for this theorem to give useful information when some $s_k > 1$, $k = 1, \ldots, N_b$, the null space of $A^{(k)}$ must be a subset of the null space of $D^{(k)}(I - \pi_D^{(k)})$; otherwise $\mu_D^{(k)} = \infty$ in (17). The proof of this slightly generalized theorem is almost identical to the one in [9]. We include it here because it shows where an overestimate of μ_D can occur when a block $A^{(k)}$ has a near null vector that is not a near null vector of $D^{(k)}(I - \pi_D^{(k)})$, which is often the case when the anisotropy in problem (1) is large and aggregates cannot be perfectly aligned with the direction of anisotropy. For such problems, aligning aggregates close to the direction of anisotropy may yield better results than the theorem predicts.

Proof of theorem: Assume wlog that each $\mu_D^{(k)} < \infty$, $k = 0, 1, \ldots, N_b$; otherwise the result is trivial. If $s_0 > 0$ then this implies that $A^{(0)}$ is positive definite so the expression for $\mu_D^{(0)}$ becomes

$$\mu_D^{(0)} = \begin{cases} 0 & \text{if } s_0 = 0 \\ \sup_{v \neq 0} \frac{v^T D^{(0)} v}{v^T A^{(0)} v} & \text{if } s_0 > 0 \end{cases}.$$

To establish (18), first note that

$$D(I - \pi_D) = \begin{pmatrix} D^{(0)} & & & \\ & D^{(1)}(I - \pi_D^{(1)}) & & \\ & & \ddots & \\ & & & D^{(N_b)}(I - \pi_D^{(N_b)}) \end{pmatrix}.$$

Hence expression (6) for μ_D can be written as

$$\begin{aligned} \mu_D &= \lambda_{max}(A^{-1} D(I - \pi_D)) = \max_{\substack{v \in \mathbb{R}^N \\ v \neq 0}} \frac{v^T D(I - \pi_D) v}{v^T A v} \\ &= \max_{v \neq 0} \frac{v^T D(I - \pi_D) v}{v^T A_b v + v^T A_r v} \\ &= \max_{v \neq 0} \frac{{v^{(0)}}^T D^{(0)} v^{(0)} + \sum_{k=1}^{N_b} {v^{(k)}}^T D^{(k)} (I - \pi_D^{(k)}) v^{(k)}}{\sum_{k=0}^{N_b} {v^{(k)}}^T A^{(k)} v^{(k)} + v^T A_r v}, \end{aligned} \tag{19}$$

where we have written v in the block form $v = ({v^{(0)}}^T, {v^{(1)}}^T, \ldots, {v^{(N_b)}}^T)^T$.

Let $w = ({w^{(0)}}^T, {w^{(1)}}^T, \ldots, {w^{(N_b)}}^T)^T$ be a vector that attains the maximum in (19). First note that if each term ${w^{(k)}}^T A^{(k)} w^{(k)}$, $k = 0, 1, \ldots, N_b$ in the sum in the denominator of (19) were 0, then the numerator of (19) would also be 0 since each $\mu_D^{(k)}$ is finite. If this were the case, then since A is SPD, the other term in the denominator, $w^T A_r w$, would have to be positive and then $\mu_D = 0$. In this case, the result (18) holds trivially. Therefore we can assume that at least one of the terms ${w^{(k)}}^T A^{(k)} w^{(k)}$ is nonzero and, since each $A^{(k)}$ is nonnegative definite, this implies that the sum in the denominator of (19) is positive. Since A_r is nonnegative definite we can write

$$\mu_D \leq \frac{{w^{(0)}}^T D^{(0)} w^{(0)} + \sum_{k=1}^{N_b} {w^{(k)}}^T D^{(k)} (I - \pi_D^{(k)}) w^{(k)}}{\sum_{k=0}^{N_b} {w^{(k)}}^T A^{(k)} w^{(k)}}. \tag{20}$$

Since the numerator and denominator each consist of a sum of nonnegative terms (or positive terms if we eliminate any zeros), the ratio of these sums is less than or equal to the largest ratio of the terms. [To see this, note that if $a_1, \ldots, a_n, b_1, \ldots, b_n > 0$, then $(\sum_{i=1}^n a_i)/(\sum_{i=1}^n b_i) = \sum_{i=1}^n (a_i/b_i) \cdot (b_i/\sum_{j=1}^n b_j)$, which is a convex combination of the ratios a_i/b_i, hence less than or equal to the maximum of these ratios.] Thus we can write

$$\begin{aligned}
\mu_D &\leq \max_{k=0,\ldots,N_b} \frac{w^{(k)^T} D^{(k)} (I - \pi_D^{(k)}) w^{(k)}}{w^{(k)^T} A^{(k)} w^{(k)}} \quad (\text{where } \pi_D^{(0)} \equiv 0) && (21)\\
&\leq \max_{k=0,\ldots,N_b} \sup_{v \in \mathbb{R}^{s_k} \setminus \mathcal{N}(A^{(k)})} \frac{v^T D^{(k)} (I - \pi_D^{(k)}) v}{v^T A^{(k)} v} && (22)\\
&= \max_{k=0,\ldots,N_b} \mu_D^{(k)}. \quad \square
\end{aligned}$$

Note that in going from equality (19) to inequality (20), the term $w^T A_r w$ was dropped; since A_r is small compared to A_b, this will not change the ratio much *unless* w is a near null vector of A_b. In addition to the upper bound (18), a lower bound on μ_D was derived in [9, Theorem 4.1]. Based on (19), one can give a lower bound on μ_D by inserting any nonzero vector into the expression on the right-hand side. The vector suggested in [9] was $\tilde{v} = (\alpha_0 v_*^{(0)^T}, \alpha_1 v_*^{(1)^T}, \ldots, \alpha_{N_b} v_*^{(N_b)^T})^T$, where each $v_*^{(k)}$ is a unit vector that achieves the supremum in (22), and

$$\alpha_k = \gamma_k (v_*^{(k)^T} A^{(k)} v_*^{(k)})^{-1/2}, \quad k = 0, 1, \ldots, N_b, \tag{23}$$

where the γ_k's are parameters on the order of 1. Substituting $\tilde{v}$ into the right-hand side of (19), gives the lower bound

$$\begin{aligned}
\mu_D &\geq \frac{\alpha_0^2 (v_*^{(0)^T} D^{(0)} v_*^{(0)}) + \sum_{k=1}^{N_b} \alpha_k^2 (v_*^{(k)^T} D^{(k)} (I - \pi_D^{(k)}) v_*^{(k)})}{\sum_{k=0}^{N_b} \alpha_k^2 (v_*^{(k)^T} A^{(k)} v_*^{(k)}) + \tilde{v}^T A_r \tilde{v}}\\
&= \frac{\sum_{k=0}^{N_b} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0}^{N_b} \gamma_k^2 + \tilde{v}^T A_r \tilde{v}}. && (24)
\end{aligned}$$

Unfortunately, this lower bound is not very useful in some of the numerical examples presented later because $\tilde{v}^T A_r \tilde{v}$ is large compared to the other term in the denominator.

5 Convergence Rates for Various Aggregates

In this section we consider the rotated anisotropic diffusion equation (1) with $\theta = \pi/4$ and $\epsilon \in [0, 1]$. Values of $\mu_D^{(k)}$, $k = 0, 1, \ldots, N_b$, are computed for various aggregation strategies and these are used to obtain a bound on μ_D via (18). This is then used along with an estimate of $\lambda_{max}(D^{-1}X)$ to obtain a bound on μ_X via (6), and finally these values are used in (7) to bound the spectral radius $\rho(E_{TG})$ of the two-grid iteration matrix. The bounds are then compared with actual computed values to determine if they are realistic. The bounds hold when one pre-smoothing and one post-smoothing

step is performed at each iteration and also when just one pre- or one post-smoothing is performed. We consider both possibilities, using a damped Jacobi method with damping factor ω^{-1} satisfying the assumption needed for (7) in Theorem 3.1. By Gerschgorin's theorem, using the stencil in (14), the eigenvalues of $D^{-1}A$ are at most $(9+3\epsilon)/(4+4\epsilon)$ if $\epsilon \leq 1/5$ or 2 if $\epsilon > 1/5$, so these will be the values used for ω^{-1}. Thus $D = \mathrm{diag}(A)$ and $D^{(k)}$, $k = 1, \ldots, N_b$, is the 16 by 16 block of D corresponding to potential aggregate k: $D^{(k)} = 8(1+\epsilon)I_{16\times 16}$. We start with the diagonal aggregates pictured in Figure 2 and $P^{(k)}$ defined by (15).

For problems with a fixed stencil such as that in (14) (i.e., the values in the stencil do not depend on the location within the domain) and potential aggregates as pictured in Figure 1 (where the number of interior nodes in each direction is $4m+2$ for some integer m so that each potential aggregate is 4 by 4), the blocks $A^{(k)}$ and $D^{(k)}$, $k = 1, \ldots, N_b$, in Theorem 3.2' are all identical. Thus, except for the computation of $\mu_D^{(0)}$ which is easily estimated with Gerschgorin's theorem, the problem of bounding μ_D reduces to the problem of finding the largest (determined) generalized eigenvalue of a single pair of 16 by 16 matrices. [The fact that the expression on the right-hand side of (17) is a generalized eigenvalue follows from a generalization of the Courant-Fischer minimax theorem [1].]

For the problem with stencil (14) and $P^{(k)}$ defined by (15), one can look at the matrices $A^{(k)}$ and $D^{(k)}(I - \pi_D^{(k)})$ symbolically and see that for $\epsilon > 0$ the null space of $A^{(k)}$ consists of the constant vectors, while when $\epsilon = 0$ the null space of $A^{(k)}$ consists of linear combinations of the constant vector and the vector $(0,0,0,0,1,1,1,-1,-1,-1,2,2,-2,-2,3,-3)^T$ (ordering nodes along the main diagonal of the potential aggregate first, then along the first upper and first lower diagonal, etc.). The null space of $D^{(k)}(I - \pi_D^{(k)})$ consists of *all* vectors that are constant along diagonals and so contains each of these vectors.

Numerically computing generalized eigenvalues of a pair of singular matrices can be hazardous, but knowing the null spaces of the matrices involved makes this computation easy; it can be converted to a problem involving a positive definite matrix $Z^* A^{(k)} Z$ and the semidefinite matrix $Z^* D^{(k)}(I - \pi_D^{(k)})Z$ where Z has dimensions s_k by $\mathrm{rank}(A^{(k)})$ and the columns of Z span the range of $A^{(k)}$ [1]. Figure 3 (a) shows a plot of $\mu_D^{(k)}$ vs ϵ.

Using Gerschgorin's theorem one can bound $\mu_D^{(0)}$ based on the stencil (14). Since each node in $A^{(0)}$ couples to at least three boundary nodes, the smallest eigenvalue of $A^{(0)}$ is at least

$$8 + 8\epsilon - \left[3 + 3\epsilon + \frac{5}{2} - \frac{1}{2}\epsilon + \left|\frac{1}{2} - \frac{5}{2}\epsilon\right|\right] = \begin{cases} 2 + 8\epsilon & \text{if } \epsilon \leq 1/5 \\ 3 + 3\epsilon & \text{if } \epsilon > 1/5 \end{cases} .$$

Since $D^{(0)} = 8(1+\epsilon)I$, the right-hand side of (16) is $8(1+\epsilon)$ over the smallest eigenvalue of $A^{(0)}$; hence

$$\mu_D^{(0)} \leq \begin{cases} (4 + 4\epsilon)/(1 + 4\epsilon) & \text{if } \epsilon \leq 1/5 \\ 8/3 & \text{if } \epsilon > 1/5 \end{cases} .$$

Figure 3 (b) shows the bounds on $\rho(E_{TG})$ obtained from Theorems 3.1 and 3.2'. Also plotted is the actual value of $\rho(E_{TG})$ for a grid with 42 by 42 interior nodes, using one pre- and one post-smoothing step per iteration (curve marked with o's) or using only one pre- or one post-smoothing step (curve marked with $*$'s). In either case, this establishes

a mesh size independent convergence rate for the 2-grid method with the aggregates shown in Figure 2.

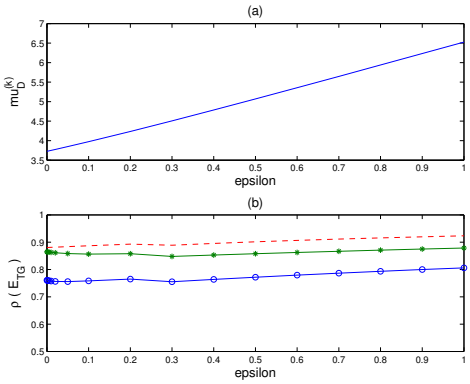


Figure 3: (a) $\mu_D^{(k)}$ for aggregates in Figure 2. (b) Bounds from Theorems 3.1 and 3.2' on $\rho(E_{TG})$ (dashed) and actual values of $\rho(E_{TG})$ on a 42 by 42 grid. Lower curve, marked with o's, is for pre- and post-smoothing; upper curve, marked with $*$'s, is for pre-smoothing only or post-smoothing only.

Unfortunately, the coarsening ratio for the aggregates shown in Figure 2 is only 7/16, while the coarse grid in a geometric multigrid method for a 2-D problem usually has about 1/4 as many points as the fine grid. The next aggregate type considered is shown in Figure 4. The potential aggregate is split into four smaller box-shaped aggregates so that the coarsening ratio is 1/4. This type of aggregate is standard for isotropic problems.

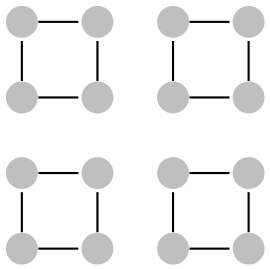


Figure 4: Box aggregates.

In this case, if nodes are again numbered consecutively in aggregates (so the lower left aggregate consists of nodes 1-4, the lower right aggregate consists of nodes 5-8, etc.) then the matrix $D^{(k)}(I - \pi_D^{(k)})$ in (17) is

$$8(1+\epsilon)\left[I - \begin{pmatrix} \frac{1}{4}e_4e_4^T & & & \\ & \frac{1}{4}e_4e_4^T & & \\ & & \frac{1}{4}e_4e_4^T & \\ & & & \frac{1}{4}e_4e_4^T \end{pmatrix}\right].$$

The constant vectors are contained in the null space of $D^{(k)}(I - \pi_D^{(k)})$, but note that when $\epsilon = 0$ the null space of $A^{(k)}$ is **not** a subset of that of $D^{(k)}(I - \pi_D^{(k)})$. The vector $(0,0,0,0,1,1,1,-1,-1,-1,2,2,-2,-2,3,-3)^T$ (using a diagonal ordering of nodes), was shown to be in the null space of $A^{(k)}$, but with the current ordering of nodes this vector becomes $(0,1,-1,0,2,3,1,2,-2,-1,-3,-2,0,1,-1,0)^T$ and this is *not* in the null space of $D^{(k)}(I - \pi_D^{(k)})$ given above; the product is $2(1+\epsilon) \cdot (0,4,-4,0,0,4,-4,0,0,4,-4,0,0,4,-4,0)^T$. Thus it can be expected that for small ϵ the right-hand side of (17) will be large; $\mu_D^{(k)}$ approaches ∞ as $\epsilon \to 0$.

Figure 5 shows the spectral radius $\rho(E_{TG})$ on a 42 by 42 grid using one pre- and one post-smoothing step per cycle (o's) and using only one pre-smoothing or one post-smoothing step per cycle ($*$'s). The left graph uses a linear scal in ϵ, while the right one displays the same results using a logarithmic scale for ϵ between $1.0e-3$ and 1. The spectral radius is indeed close to 1 when ϵ is small, but it is not as bad as indicated by the upper bound in terms of $\mu_D^{(k)}$. For $\epsilon = 0.001$, for example, the computed value of $\mu_D^{(k)}$ was 595.12, while the value of μ_D on the 42 by 42 grid was only 17.95. The spectral radius of the iteration matrix was 0.9655 for $\nu_1 = \nu_2 = 1$ and 0.9752 for $\nu_1 + \nu_2 = 1$, while the estimate based on $\mu_D^{(k)}$ would have been $\mu_X \leq (9.003/4.004) \cdot 595.12 = 1338.1$, $\rho(E_{TG}) \leq 1 - 1/1338.1 = 0.9993$.

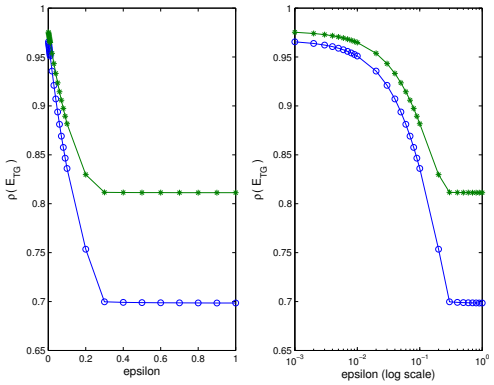


Figure 5: Actual values (o's and $*$'s) for $\rho(E_{TG})$ with box aggregates on a 42 by 42 grid. Lower curve with o's is for pre- and post-smoothing, upper curve with $*$'s is for pre-smoothing only or post-smoothing only.

For these aggregates, when ϵ is small, the value $\mu_D^{(k)}$ is a large overestimate of μ_D, at least on a 42 by 42 grid. It is shown in Figure 6, however, that μ_D increases with the grid size. It is not clear whether it asymptotes to the value $\mu_D^{(k)} = 595.12$ or to something smaller, but in either case it is large enough to indicate poor convergence of the two grid method. The lower bound (24) is not of much help here because the vectors $v_*^{(k)}$, $k = 1, \ldots, N_b$ used to derive that bound are near null vectors of $A^{(k)}$. This means that the coefficients $\alpha_1, \ldots, \alpha_{N_b}$ in (23) will be huge if the γ_k's are of order 1 and hence that $\tilde{v}^T A_r \tilde{v}$ will be huge unless the γ_k's can be chosen in such a way that $|\tilde{v}^T A_r \tilde{v}| << \|A_r\| \, \|\tilde{v}\|^2$. We have found no such γ_k's and, for very small ϵ, the best lower bounds of this form appear to be obtained by taking $\gamma_0 = 1$ and $\gamma_1 \approx \ldots \approx \gamma_{N_b} \approx 0$, in

which case one obtains a lower bound on the order of $\mu_D^{(0)} << \mu_D$.

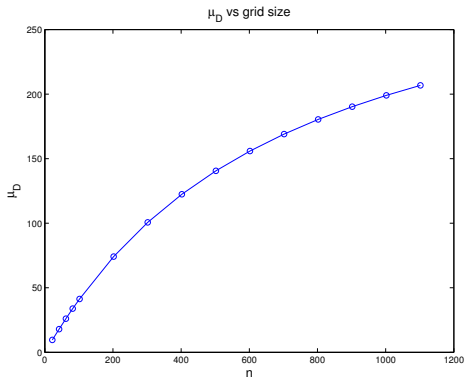


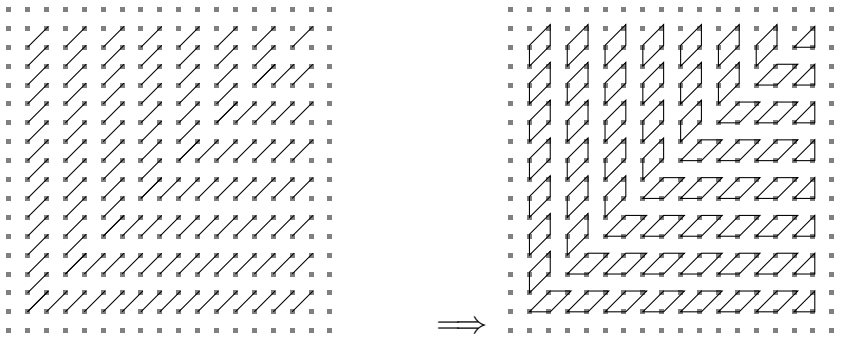
Figure 6: Computed values of μ_D vs grid size n for box aggregates, $\theta = \pi/4$, $\epsilon = 0.001$.

In general, when $D^{(k)}$ is a scalar multiple of the identity, or, more generally, a diagonal matrix with positive diagonal entries, in order for a vector to be in the null space of $D^{(k)}(I - \pi_D^{(k)})$, it *must* be constant on individual aggregates. But when $\epsilon = 0$, the null space of $A^{(k)}$ contains a vector that is constant only on the diagonal aggregates pictured in Figure 2. Thus, for small ϵ, one can expect poor behavior (or at least a poor bound on the convergence rate) of the method for any aggregates other than those (or a refinement of those) shown in Figure 2.

The automatic aggregation procedure described in [10] and implemented in [13] does not use "potential aggregates" but tries various pairwise aggregations successively, in order to minimize a quantity related to the aggregate quality. For the problem with $\theta = \pi/4$ and $\epsilon = 0.001$, in the first pass it formed pairwise aggregates aligned with the anisotropy, as pictured in Figure 7(a). On the second pass, however, the algorithm grouped pairs to the north and south above the 45° line and pairs to the east and west below the 45° line, forming the parallelogram shaped aggregates shown in Figure 7(b). In this case, the spectral radius of the iteration matrix with pre- and post-smoothing on a 42 by 42 grid was 0.9307, indicating rather slow convergence of the two-grid method. This spectral radius increased to 0.9614 on an 82 by 82 grid, and while it will asymptote to something less than 1, the asymptotic value as $N \to \infty$ may be quite close to 1.

6 Determining Appropriate Aggregates

The negative results of the previous section suggest a possible way to determine which points to aggregate. Form the potential aggregates and the splitting $A_b - A_r$. Compute the eigensystem of each block $A^{(k)}$ and look at the eigenvectors corresponding to any small eigenvalues. Aggregate only points on which those eigenvectors are near constant. For example, when $\theta = \pi/6$ and $\epsilon = 0.001$, one finds that the matrix $A^{(k)}$ again has a zero eigenvalue corresponding to the constant vector and it has an eigenvalue of 0.0027 corresponding to a vector whose first and tenth components are close (0.13 and 0.15), whose second and eleventh components are close (-0.07 and -0.04), whose third and



(a) First pass (b) Second pass

Figure 7: First pass and second pass of the automatic aggregation procedure in [10]

twelfth components are close (-0.26 and -0.23), etc., with the close components lying along lines at an angle $\pi/6$ with the vertical axis, as pictured in Figure 8. Additionally, components four and eight are fairly close (-0.46 and -0.35), as are components nine and thirteen (0.35 and 0.46), so these might be considered for aggregation as well.

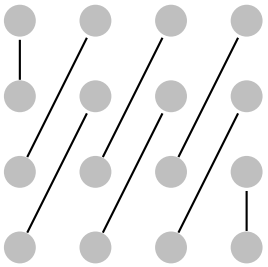


Figure 8: Aggregates based on nearly equal components of eigenvector of $A^{(k)}$ corresponding to eigenvalue 0.0027 when $\theta = \pi/6$, $\epsilon = 0.001$.

Before proceeding with this example, it should be noted that with this parameter set, the finite element matrix defined by (13) is not an M-matrix; couplings to points to the left and right, $(2 - 4\epsilon)\cos^2\theta + (2\epsilon - 4)\sin^2\theta = .4975$, are positive. Hence not all points next to the boundary of the domain correspond to diagonally dominant rows, and such points should be included among the potential aggregates. In order to keep the blocks identical, we will now include all points next to the boundary among the potential aggregates. The number of interior grid points in each direction will be taken to be a multiple of 4 so that each block $A^{(k)}$ still corresponds to a 4 by 4 group of nodes.

Using the aggregates in Figure 8 does not result in a very small value for $\mu_D^{(k)}$ since the eigenvector of $A^{(k)}$ associated with the small eigenvalue is not extremely close to the null space of $D^{(k)}(I - \pi_D^{(k)})$; we found $\mu_D^{(k)} = 50$. The computed value of μ_D was considerably smaller, however. For a 44 by 44 grid (where now all nodes are considered for aggregation), we computed $\mu_D = 3.71$, $\rho(E_{TG}) = 0.83$ for pre- and post-smoothing, and $\rho(E_{TG}) = 0.88$ for pre-smoothing only or post-smoothing only. Figure 9 shows μ_D for larger grid sizes. Again it grows with the grid size and the asymptotic limit is not clear, even from grid sizes up to $n = 1100$.

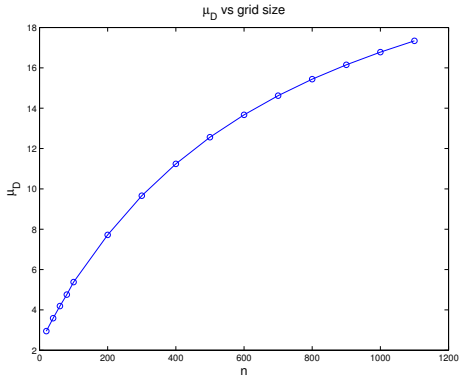


Figure 9: μ_D vs grid size n for aggregates corresponding to $\theta = \pi/6$; $\epsilon = 0.001$.

Based on the above results it appears that eigenvectors corresponding to small eigenvalues of $A^{(k)}$ do, indeed, indicate the direction of anisotropy and can be used to determine appropriate aggregates. However, to automate this procedure, one must define exactly what a "small" eigenvalue of $A^{(k)}$ is and how "nearly equal" eigenvector components must be in order to be aggregated. We will return to this later, but for now we consider another idea for automating the choice of aggregates. That is to try several different choices from among the potential aggregates and see which gives the smallest value for $\mu_D^{(k)}$. As noted previously, μ_D may be small even when $\mu_D^{(k)}$ is not, but it is reasonable to expect that even when $\mu_D^{(k)}$ cannot be made small, aggregates that make it less large will be more likely to yield a good value for μ_D.

In the upper left block of Figure 11, we have plotted the spectral radius of the iteration matrix using pre- and post-smoothing on a 44 by 44 grid as a function of ϵ and θ, where the aggregates were chosen from among the 13 possible choices pictured in Figure 10. For these tests, we let ϵ range between 0 and 1; more specifically, $\epsilon = 0.001,\ 0.002,\ 0.005,\ 0.01,\ 0.02,\ 0.05,\ 0.1,\ 0.2,\ 0.3,\ \ldots,\ 1$. The angle θ ran from $-\pi/2$ to $\pi/2$, in steps of 0.1.

Although some aggregate choices give a better coarsening ratio than others, we did not take this into account but simply chose the aggregate type that gave the smallest value for $\mu_D^{(k)}$. The upper middle block of Figure 11 shows the aggregate types that were chosen for each (ϵ, θ) pair. For $\epsilon > 0.5$, box aggregates (type 13) were always chosen; these problems have only mild anisotropy. For more strongly anisotropic problems, the aggregates tended to align with the anisotropy, as can be seen by comparing the aggregate type numbers in Figure 11 with their definitions in Figure 10. Also plotted in Figure 11 are the values of $\mu_D^{(k)}$ (lower left) and μ_D (lower right). It can be seen that the value of μ_D, which determines the spectral radius, is significantly less than $\mu_D^{(k)}$ for ϵ near 0.

For comparison, the upper right plot in Figure 11 shows $\rho(E_{TG})$ using box aggregates only. It can be seen that for ϵ near 0, box aggregates give significantly larger values for $\rho(E_{TG})$. The maximum value using box aggregates was 0.9965, while for aggregates that minimize $\mu_D^{(k)}$, it was 0.9324. Moreover, while $\rho(E_{TG})$ using box aggregates was large for all angles θ when $\epsilon = 0.001$ (greater than 0.9668), for values of θ that could

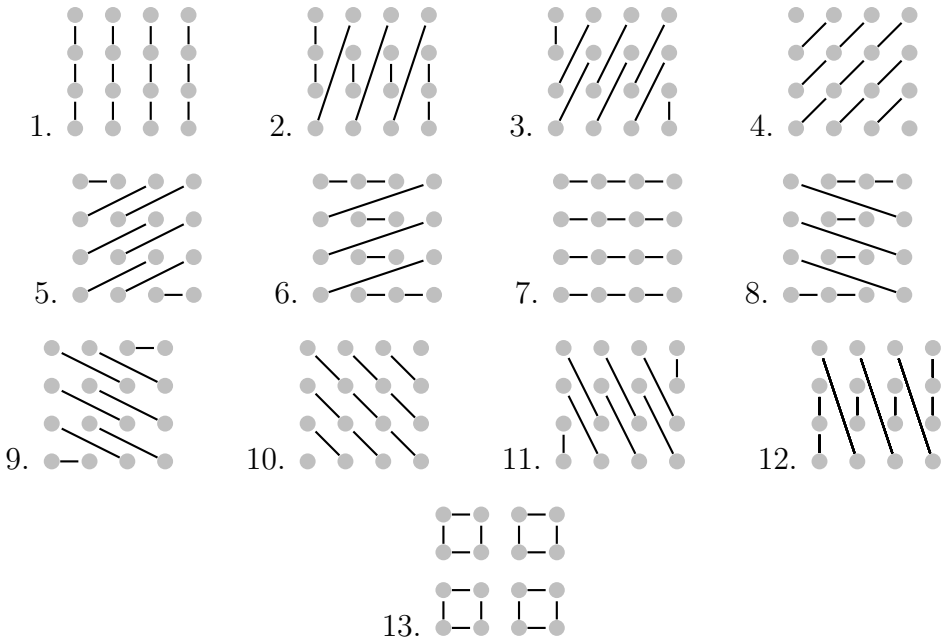


Figure 10: Possible aggregate choices.

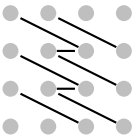
be well-represented in the potential aggregates, $\rho(E_{TG})$ dropped as low as 0.7618 when appropriate aggregate types were chosen.

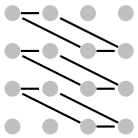
With information from these runs in hand, we decided to choose a definition of "small" eigenvalue and "nearly equal" components of the corresponding eigenvector and attempt to choose aggregates based on these notions, rather than having fixed aggregate choices. Recall that the difficulty with these problems stems from the fact that when $\epsilon = 0$, each block $A^{(k)}$ has a second null vector, in addition to the constant vector, and this will be a null vector of $D^{(k)}(I - \pi_D^{(k)})$ *only* if points within an aggregate correspond to equal components of this eigenvector. We decided that the second eigenvalue of $A^{(k)}$ would be considered "small" if it was less than 1/2 times the next smallest eigenvalue. If it was greater than this, then we concluded that the problem was not highly anisotropic and chose box aggregates. If the second eigenvalue was deemed "small", then we examined the corresponding eigenvector. We set a tolerance for "nearly equal" components at 0.05. We then looked for groups of four or more entries of the eigenvector whose values ranged over no more than three times the tolerance, or, 0.15. The largest such group was aggregated and we then checked the remaining entries for any more such groups, aggregating the largest such group and repeating until there were no more groups of four or more that differed by less than 0.15. We then searched the remaining entries for any groups of three whose values differed by at most twice the tolerance, or, 0.1. Such groups of three were aggregated. Finally we searched the remaining entries of the eigenvector for pairs that differed by no more than 0.05, and we aggregated such pairs. Remaining entries were left as singletons on the "coarse grid".

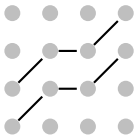
Figure 12 shows the results for $\theta = \pm\pi/3,\ \pm\pi/4,\ 0$, and an angle chosen randomly between $-\pi/2$ and $\pi/2$, which turned out to be about 0.99. [Note that the results for negative θ should be identical to those for positive θ unless roundoff plays a role; in our

experiments, the results for negative θ were the same as for positive θ.] The spectral radius of the iteration matrix (with pre- and post-smoothing) using the newly chosen aggregates is plotted with x's, while the spectral radius of the iteration matrix using aggregates chosen from the ones in Figure 10 is plotted with o's. We went out only to $\epsilon = 0.6$, because for larger values of ϵ both codes chose box aggregates and thus had identical results. For $\theta = \pm\pi/4$, the new code chose the same diagonal aggregates (types 4 and 10) as the previous one. It got somewhat better results for $\epsilon = 0.3$ and 0.4 because it deemed the second eigenvalue not to be "small" in these cases and switched to box aggregates, while the code that based its choices on the minimum value of $\mu_D^{(k)}$ did not switch to box aggregates until $\epsilon = 0.5$. Of course, this is very sensitive to the definition of a "small" second eigenvalue. For $\theta = 0$ and ϵ small (less than 0.1) both codes chose vertical aggregates (type 1). The new code continued to choose vertical aggregates until it switched to box aggregates at $\epsilon = 0.5$. Based on these results, it might have done better to switch earlier. The code that based its aggregate choice on $\mu_D^{(k)}$ chose aggregate type 12 when $\epsilon = 0.1$ and aggregate type 2 when $\epsilon = 0.2$ (near vertical aggregates), but these choices turned out not to be quite as good as the vertical aggregates. It switched to box aggregates when $\epsilon = 0.3$ and got a smaller spectral radius than the new code until it too switched to box aggregates.

For $\theta = \pm\pi/3$ and $\theta = 0.99$, the new code chose some aggregates that are not pictured in Figure 10. For example, when $\theta = -\pi/3$ and $\epsilon \leq 0.1$, it chose the aggregates below on the left (14), and for $\epsilon = 0.2$ it chose the aggregates in the middle (15). It chose the aggregates on the right (16) when $\theta = 0.99$ and $\epsilon \leq 0.1$. Here the coarsening ratio was only 10/16, and it might have been useful to force the code to aggregate some of the singletons. In general, however, both codes performed comparably, with similar coarsening ratios and with spectral radii between 0.5 and 0.9. The largest spectral radius still occurred for the smallest value of ϵ, namely, $\epsilon = 0.001$ (unless $\theta = \pm\pi/4$), but both codes obtained significantly better results than using box aggregates for small values of ϵ.

14. 

15. 

16. 

All of these results deal with a two-grid method and the spectral radius of the resulting iteration matrix. In practice, of course, more grid levels are used and the simple iteration procedure described here is usually accelerated using conjugate gradients or related iterative methods. To see if these two-grid results would carry over to a K-cycle multigrid method with flexible conjugate gradient acceleration, we decided to test the first aggregation strategy (based on minimizing $\mu_D^{(k)}$ using aggregate choices in Figure 10) in the code AGMG [10]. This aggregation strategy was used only on the finest level, where it was straightforward to identify the potential aggregates and then to choose the actual aggregates from among the choices in Figure 10 to minimize $\mu_D^{(k)}$. Aggregates at other levels were formed using the procedure in the AGMG code. Number of iterations and execution times were compared, with and without the new aggregation strategy, for $\epsilon = 0.001$ and $\theta = \pi/12$, $\pi/6$, and $\pi/4$. A right-hand side vector of all 1's was used,

$\epsilon = 0.001$			aggregates that minimize $\mu_D^{(k)}$	AGMG aggregates
θ	N	# levels	time in secs (iterations)	time in secs (iterations)
$\pi/4$	500^2	5	1.65 (60)	1.87 (66)
	1000^2	5	8.17 (72)	9.48 (80)
	2000^2	7	40.0 (82)	45.7 (93)
	3000^2	7	92.4 (87)	111 (98)
$\pi/6$	500^2	5	2.93 (97)	3.53 (124)
	1000^2	6	15.1 (121)	18.7 (155)
	2000^2	7	70.7 (139)	92.5 (182)
	3000^2	7	201 (148)	262 (196)
$\pi/12$	500^2	5	1.78 (65)	3.07 (103)
	1000^2	5	8.53 (76)	15.7 (125)
	2000^2	7	38.5 (85)	77.4 (146)
	3000^2	7	99.7 (89)	191 (155)

Table 1: Timing and iteration counts for $\epsilon = 0.001$ with and without the new aggregation procedure on the finest grid.

with a zero initial guess. Results are shown in Table 1, where it can be seen that the new aggregation strategy for the finest level did, indeed, reduce both the number of iterations and the computation time. While more experiments are needed to determine if this approach is really a viable alternative – Can we identify potential aggregates without knowing the geometry of the problem? Can the new aggregation strategy be used effectively on other levels as well? – preliminary results look promising.

7 Conclusions

We have used the analysis in [9] to show that if diagonal aggregates are used in a two-grid aggregation method to solve a rotated anisotropic diffusion equation with angle of anisotropy $\theta = \pi/4$, then one obtains a bound on the convergence rate that is independent of the mesh size and can be made independent of ϵ, the level of anisotropy, as well.

We have suggested two possible strategies for choosing appropriate aggregates, based on assembling groups of nodes that are candidates for aggregation, and then either choosing from a number of possible aggregates the one that minimizes $\mu_D^{(k)}$ or looking at the eigenvector corresponding to the smallest nonzero eigenvalue of $A^{(k)}$ and aggregating points corresponding to nearly equal components of this eigenvector. The first strategy seems more reliable, as it requires no estimation of what constitutes a "small" eigenvalue or "nearly equal" components of the corresponding eigenvector. On the other hand, the second strategy might find good aggregates that are not among the list of possibilities for the first. Using larger potential aggregates should lead to better results for angles of anisotropy that are better represented with more nodes, but this means significantly more work for the first strategy as it would not only have to solve larger generalized eigenvalue problems to compute $\mu_D^{(k)}$ but would also need to examine more possible

aggregate choices. The increase in work for the second strategy would not be as great.

The question of how to choose the "potential aggregates" has not been addressed in this paper. If the problem comes from a PDE (and this is known), then nodes that are geometrically close make appropriate potential aggregates. In an algebraic setting, such nodes might be identified using current methods of defining "strongly connected" points.

The spectral radii observed in this paper are *not* spectacular in the realm of multigrid methods. Still, it should be remembered that the problems studied here are typically very difficult for algebraic multigrid methods and that the analysis was for the simplest possible outer iteration and the simplest smoother, damped Jacobi. Early experiments with the aggregation strategy in a true multigrid K-cycle using Gauss-Seidel smoothing look promising, but much remains to be done.

Acknowledgments: The authors thank Yvan Notay for helpful comments on a first draft of this paper, and they thank the referees for many additional helpful suggestions.

References

[1] H. Avron, E. Ng, and S. Toledo, *A generalized Courant-Fischer minimax theorem*, technical report LBNL-6393E, Lawrence Berkeley National Lab, August, 2008.

[2] A. Brandt, *General hghly accurate algebraic coarsening*, ETNA 10 (2000), pp. 1-20.

[3] J. Brannick, M. Brezina, S. MacLachlan, T. Manteuffel, and S. Ruge, *An energy based AMG coarsening strategy*, Numer. Lin. Alg. Appl., 13 (2006), pp. 133-148.

[4] J. Brannick, Y. Chen, and L. Zikatanov, *An algebraic multilevel method for anisotropic elliptic equations based on subgraph matching*, Numer. Lin. Alg. Appl. 19 (2012), pp. 279-295.

[5] R. Falgout, P. Vassilevski, and L. Zikatanov, *On two-grid convergence estimates*, Num. Lin. Alg. Appl., 12 (2005), pp. 471-494.

[6] H. Kim, J. Xu, and L. Zikatanov, *A multigrid method based on graph matching for convection-diffusion equations*, Num. Lin. Alg. Appl., 10 (2003), pp. 181-195

[7] O. Livne, *Coarsening by compatible relaxation*, Numer. Lin. Alg. Appl. 11 (2004), pp. 205-227.

[8] A. Muresan and Y. Notay, *Analysis of aggregation-based multigrid*, SIAM J. Sci. Comput. 30 (2008), pp. 1082-1103.

[9] A. Napov and Y. Notay, *Algebraic analysis of aggregation-based multigrid*, Numer. Lin. Alg. Appl., 18 (2011), pp. 539-564.

[10] A. Napov and Y. Notay, *An algebraic multigrid method with guaranteed convergence rate*, SIAM J. Sci. Comput., 34 (2012), pp. 1079-1109.

[11] Y. Notay, *An aggregation based algebraic multigrid method*, ETNA 37 (2010), pp. 123-146.

[12] Y. Notay, private communication.

[13] Y. Notay, *AGMG: Iterative solution with AGgregation-based algebraic MultiGrid*, `http://homepages.ulb.ac.be/~ynotay/AGMG`.

[14] Y. Notay and P. Vassilevski, *Recursive Krylov-based multigrid cycles*, Numer. Lin. Alg. Appl., 15 (2008), pp. 473-487.

[15] L. Olson, J. Schroder, R. Tuminaro, *A new perspective on strength measures in algebraic multigrid*, Numer. Lin. Alg. Appl., 17 (2010), pp. 713-733.

[16] J. Schroder, *Smoothed aggregation solvers for anisotropic diffusion*, Numer. Lin. Alg. Appl., 19 (2012), pp. 296-312.

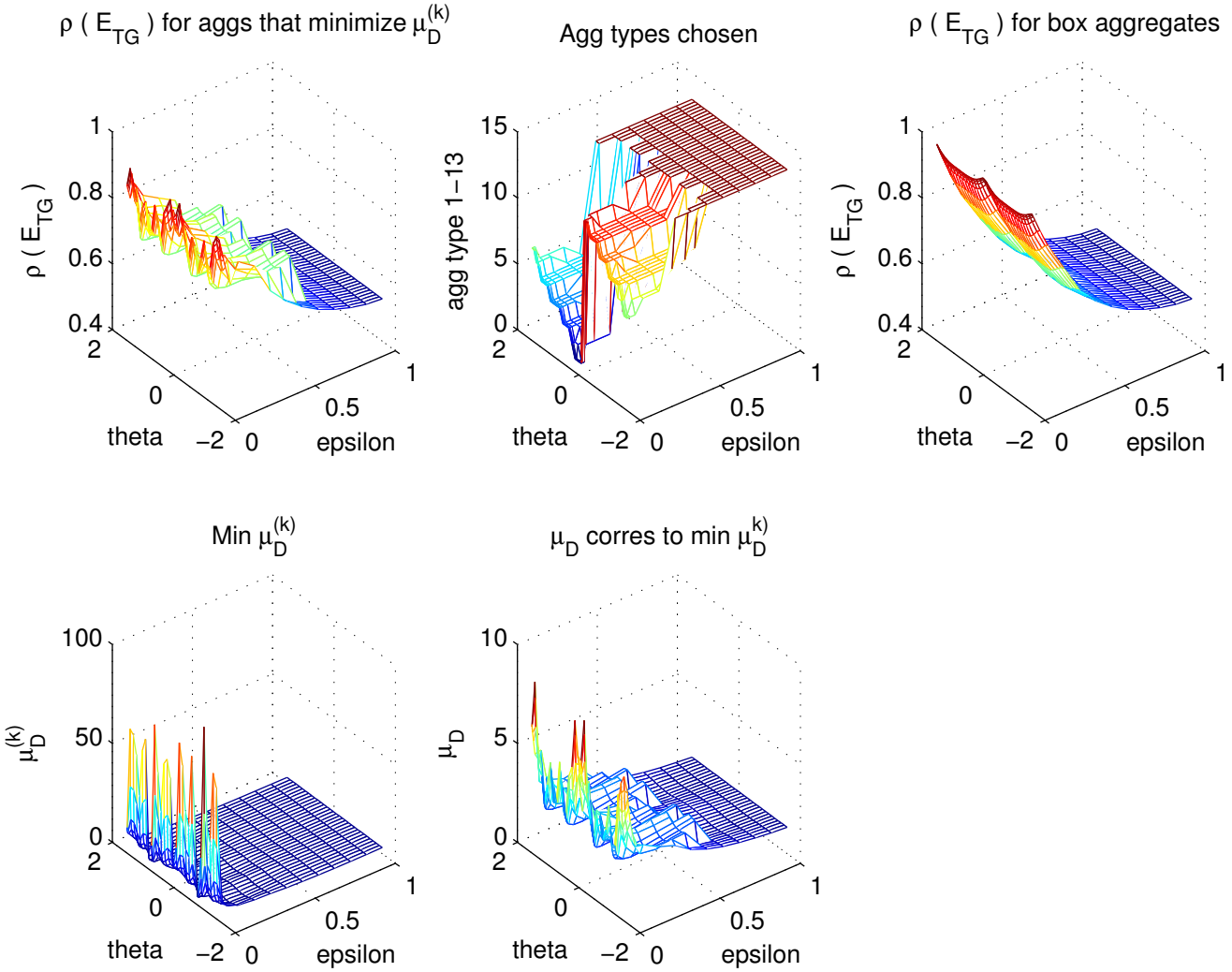


Figure 11: Upper left: $\rho(E_{TG})$ on a 44 by 44 grid using pre- and post-smoothing and aggregates that minimize $\mu_D^{(k)}$. Upper middle: Aggregate types (1-13) that were chosen. Upper right: $\rho(E_{TG})$ using box aggregates only. Lower left: $\mu_D^{(k)}$ is still fairly large for ϵ near 0. Lower right: μ_D is significantly smaller than $\mu_D^{(k)}$ for ϵ near 0.

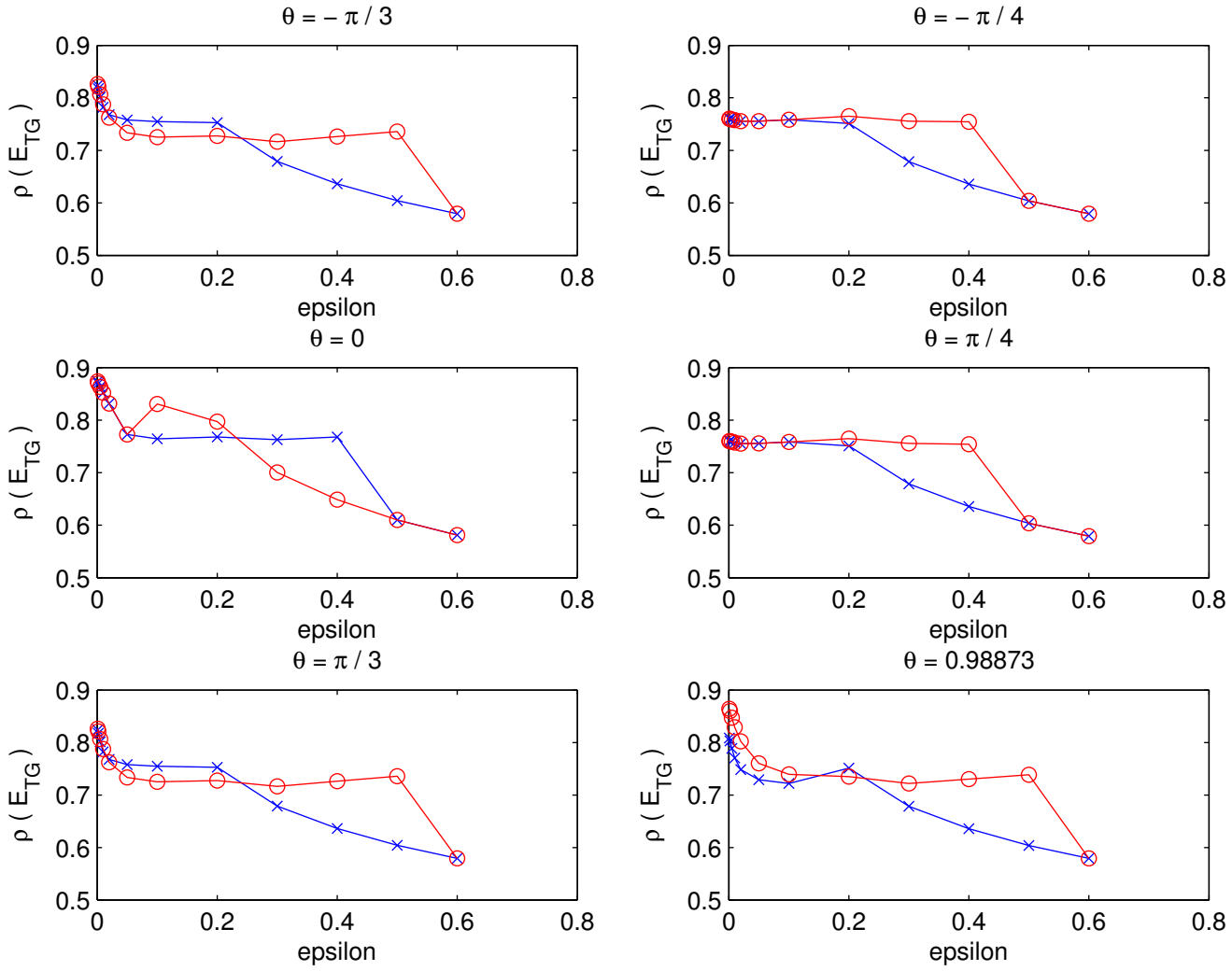


Figure 12: $\rho(E_{TG})$ on a 44 by 44 grid using pre- and post-smoothing and aggregates chosen based on "nearly equal" components of the eigenvector corresponding to a "small" eigenvalue of $A^{(k)}$ (curve marked with x's) and using aggregates chosen from among those in Figure 10 to minimize $\mu_D^{(k)}$ (curve marked with o's).