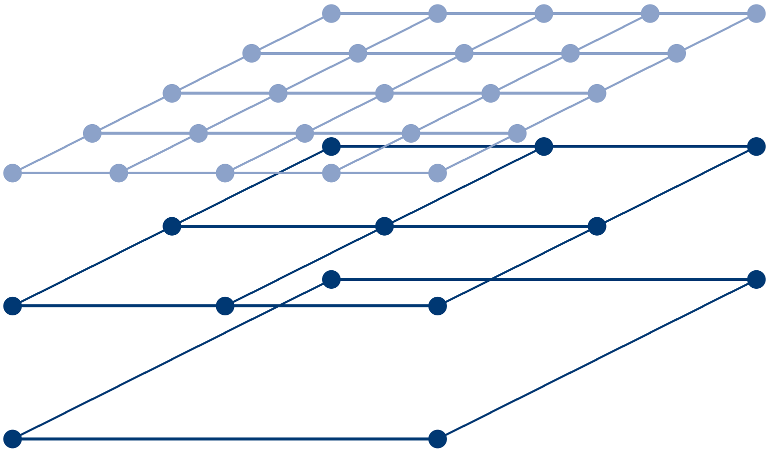
UNIVERSITÉ LIBRE DE BRUXELLES ULB

UNIVERSITÉ D'EUROPE

Algebraic analysis of V-cycle multigrid and aggregation-based two-grid methods

Artem NAPOV



2010 Bruxelles

Algebraic analysis of V-cycle multigrid and aggregation-based two-grid methods

Artem Napov

Directeur de thèse :
Prof. Yvan NOTAY
Membres du joury :
Prof. Robert BEAUWENS
Prof. Anne DELANDTSHEER
Prof. Pierre-Etienne LABEAU
Prof. Yvan NOTAY
Prof. Cornelis W. OOSTERLEE
Prof. Daniel TUYTTENS
Prof. Stefan VANDEWALLE

Thèse présentée
en vue de l'obtention du grade
de Docteur en Sciences de l'Ingénieur

Université Libre de Bruxelles
Faculté des Sciences Appilquées
Service de Métrologie Nucléaire

Bruxelles
janvier 2010

"Understanding is, after all, what science is all about – and science is a great deal more than mindless computation."

Sir Roger Penrose

"Of course everything in computerology is new; that is at once its attraction, and its weakness."

James H. Wilkinson

Remerciements / Acknowledgements

Je remercie tout d'abord mon directeur de thèse Yvan Notay pour sa disponibilité, le temps qu'il m'a consacré, pour les discussions enrichissantes et pour son encouragement, mais par dessus tout pour m'avoir fait découvrir ce vaste et surprenant chantier des idées en perfectionnement perpétuel qui est l'analyse des méthodes numériques.

Une pensée particulière va à tous les membres du service de Métrologie Nucléaire qui ont contribué, chacun à leur manière, à une ambiance de travail unique dans laquelle j'ai été immergé dès les premiers jours. Je remercie en particulier mon collègue Nicolas Pauly pour les discussions enrichissantes sur la science, l'enseignement et sur bien d'autres choses encore.

Je remercie également les membres du laboratoire Ampère de l'école centrale de Lyon, et en particulier Ronan Perrussel, qui m'ont accueilli chaleureusement en leur sein durant les quelques semaines de séjour à Lyon. Ce séjour fut enrichissant et fructueux et j'ai l'espoir que notre collaboration se poursuivra dans le futur avec la même intensité.

I am very grateful to the members of the jury (especially to Cornelius Oosterlee, Stefan Vandewalle and Robert Beauwens) who have found time to read (and proofread) the thesis and made many valuable comments and suggestions.

The last but not least. Un grand merci du fond de mon coeur va à toute ma famille: ma soeur Krystyna, ma mère Iryna et mon père Oleh. Sans leur aide et leur soutien je ne serai pas arrivé aussi loin.

Contents

Chapter 1

Introduction

1.1 Preliminaries

In this thesis we consider multigrid methods for the solution of linear systems of equations. This introductory chapter aims at situating the research material of the thesis in the general context of numerical analysis and scientific computing. In particular, the following section sheds some light on (several of the numerous) applications in which linear systems can arise. A brief overview of solutions techniques for linear systems is given in Section 1.3. Basic multigrid concepts are introduced in Section 1.4. In Section 1.5 we briefly describe the content of the following five chapters, ending up with some comments on notation.

The reader familiar with basic multigrid concepts can start directly with Section 1.5.

1.2 Why linear systems?

An important number of problems in science and engineering can be formulated in terms of linear partial differential equations (PDEs). Such equations frequently arise in:

- electrical engineering ,
- computational fluid dynamics (Stokes and Oseen equations) ,
- structural mechanics ,
- transport phenomena ,
- acoustics ,
- chemistry .

To solve numerically these PDE problems, one first performs their *discretization*; that is, the initial continuous problem, formulated at every point of the underlying domain, is reduced to a limited number of equations with usually the same number of unknowns. If the initial PDE is linear, so are the resulting equations; otherwise, it is a common practice to linearize the obtained equations using some suitable Newton-like scheme. In other words, discrete PDEs usually lead to a linear system, stated in vector-matrix notation as

$$A\mathbf{x} = \mathbf{b}\,. \tag{1.1}$$

The main discretization techniques are:

- *finite element methods*, which use a linear combination of appropriately chosen shape functions to approximate the solution; the unknowns are the weights of shape functions and linear system results from application of a minimization principle to the discretization error [12, 76];
- *finite volume methods*, based on the subdivision of the underlying domain into cells, on which unknown function(s) (often describing physical quantities) are assumed constant; linear system is then formed by balance equations that account on sources inside cells and on the transport of physical quantities between them;
- *finite difference methods*, which consider unknown function(s) in a given number of nodes inside or on the boundary of the domain; the linear system arises from PDE(s) when derivatives of each unknown function are approximated by its differences [56, 40].

Systems arising from a discretization of PDEs are often *sparse*; that is, each of their equations relate together only a small number of unknowns, and the major part of the entries of A equals zero. It then makes sense to keep in memory only the nonzero entries and their position in the matrix, which further enables to tackle problems with an important number of unknowns (10^7 for a usual PC).

Besides PDE applications, a number of problems are already discrete and formulated as a linear system of equations. Such problems arise, for instance, in image restoration or signal processing [43].

1.3 Linear system solvers

The solution of linear system(s) is the most time-consuming process in the majority of scientific computing applications and therefore should not be neglected. When regular systems are considered, it can be performed either by *direct* or by *iterative* methods.

Direct methods are usually variants of Gaussian elimination. In practice, this latter is often performed by factorizing the system matrix into a product of lower and upper triangular matrices (LU factorization), the process being finished by the consecutive solution of two related triangular systems. Even if the initial system is sparse, the triangular factors rarely have the same sparsity: direct methods often have important memory requirements.

The idea behind iterative methods is to solve the linear system (1.1) approximately using a suitable procedure, which we formally denote

$$\tilde{\mathbf{x}} = B(\mathbf{b})\,.$$

The system is then solved (exactly) if we recover the correction vector $\mathbf{e}$ such that

$$A(\tilde{\mathbf{x}} + \mathbf{e}) = \mathbf{b}\,,$$

or, equivalently,

$$A\mathbf{e} = \mathbf{r}\,, \tag{1.2}$$

where $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ is called residual. This latter equation (also called *correction equation*) is equivalent to the initial system (1.1) and can again by solved approximately. The procedure is repeated until the required precision is reached.

Note that iterative methods rarely give the exact solution of the linear system (1.1). However, if properly designed, they allow to come closer to the solution at each iteration step. This feature is particulary relevant since the solution with only a limited accuracy is often required.

An important characterization of iterative solvers is their *optimality* with respect to a given class $A(n)$ of linear system matrices, where n denote the system size. An optimal iterative method, when applied to systems with system matrix $A(n)$, should have

- its cost per iteration proportional to the system size n,
- its *convergence rate* (gain in precision per iteration step) bounded above by a constant that does not depend on n.

Clearly, if the solution of the linear system (1.1) is determined up to a desired precision ϵ with an optimal iterative method, the computational cost is proportional to $n\log(\epsilon)$. Using direct methods for the same purposes amounts to $\mathcal{O}\left(n^3\right)$ operations if the matrix is dense (not sparse) and to $\mathcal{O}\left(n^2\right)$ operations if it arises from discretization of typical 2-dimensional PDEs [62, p.9] [61, p.14]. Therefore, for system size n large enough, optimal (and even some suboptimal) iterative methods become more attractive than direct solvers.

Among the most popular iterative techniques, we should mention:

- *Krylov subspace methods*, that can be viewed as simple iterative methods where the approximation $B(\mathbf{r})$ of correction is weighted after each iteration in order to satisfy some minimization principle. The approximate solution procedure $B(\cdot)$ can still be chosen freely and is then called *preconditioner*.

- *Multigrid (multilevel) methods*, which we introduce below, have been the first numerical techniques to reach the optimal convergence for usual applications. They are considered as the most efficient methods for the solution of system arising from discretization of elliptic PDEs and among the most efficient approaches for other PDE applications.

- *Domain decomposition methods*, that correspond to a class of approaches specially designed for parallel computer architecture. Their main idea is to split the unknowns into a number of sets such that communication between such sets is reduced during the solution process.

- *Incomplete factorizations (ILU)*, often used as preconditioners by default for Krylov subspace methods. The main idea is to reduce the cost and memory requirements of direct methods that perform complete LU factorization by dropping some entries in the triangular factors. Due to their purely algebraic nature, ILU techniques can be of interest when applied to problems for which the other methods fail.

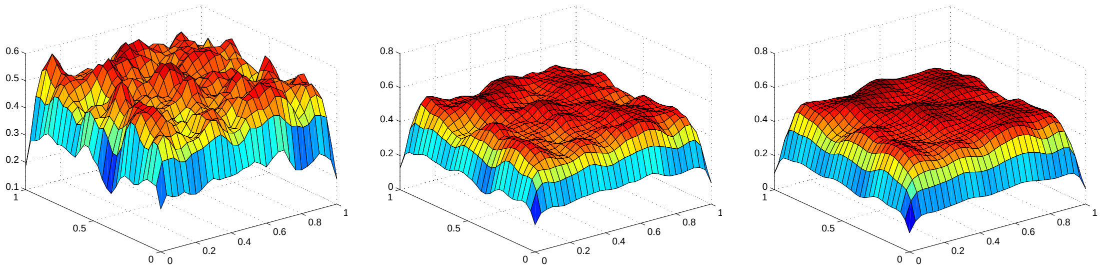
For further details on linear system solvers, we refer to corresponding chapters in [24]. Introductory material on iterative methods (including the main variants listed here) can be found in [54], whereas more advance subjects are treated in [3]. For further information on the preconditioning techniques we refer to [7], whereas a broad presentation of Krylov methods from the historical perspective can be found in [55]

1.4 Multigrid methods

The efficiency of multigrid methods depends on the interplay between its two main components: *smoother* and *coarse grid correction.* The smoother is often a simple iterative method, and, if used alone, has poor convergence properties. For Poisson-like problems

$$\frac{\partial}{\partial x}\left(\alpha_x \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\alpha_y \frac{\partial u}{\partial y}\right) + \beta u = f \tag{1.3}$$

the two well known examples are Jacobi and Gauss-Seidel smoothers [61, Chapters 1-2]; both correspond to a linear approximation procedure $B(\mathbf{v}) = B\mathbf{v}$, where B is, respectively, the diagonal and (up to some permutations) the lower triangular parts of A. When applied to the linear system (1.1), such schemes reduce the magnitude of oscillatory modes in the correction $\mathbf{e}$, while keeping the smooth components unchanged. After several smoothing iterations, the correction becomes geometrically smooth; that is, it varies slowly from one point to another (see Figure 1.1 for illustration). Other examples are block smoothers [61, Section 5.1] for anisotropic problems, ILU smoothers [71, 70] in computational fluid dynamics applications and hybrid smoothers for problems in electromagnetics [29](see also Chapter 6).



(a) (b) (c)

FIGURE 1.1: An example of correction $\mathbf{e}$ smoothed by Gauss-Seidel scheme; (a) initial correction (b) correction after 1 iteration (c) correction after 2 iterations. The corresponding linear system A was obtained by discretization of constant-coefficient isotropic Poisson PDE (1.3) with Dirichlet boundary conditions on rectangular grid 33×33.

The smooth character of the correction $\mathbf{e}$ can then be exploited, approximating it by a smaller *coarse* vector $\mathbf{e}_c$ of size $n_c < n$ which still reproduces the essential part of the correction's behaviour. This coarse correction vector is obtained by solving a smaller *coarse* $n_c \times n_c$ system

$$A_c \mathbf{e}_c = \mathbf{r}_c \tag{1.4}$$

that approximates the initial *fine* correction system (1.2). This solution corresponds to the second main multigrid ingredient, known as *coarse grid correction.*

If the coarse grid correction step is performed by a direct solver, its combination with a smoothing scheme is called *two-grid* method. Whereas it is often cheaper than a direct method, the system to be solved is smaller than the fine one only by a modest factor (4 in usual applications from two-dimensional PDE problems); the two-grid scheme is therefore still not optimal. The coarse system (1.4) can however be solved approximately by (recursively) applying γ iterations of the two-grid method; the recursion argument can be repeated, forming coarser and coarser systems, until a small enough system size is reached. If $\gamma = 1$, the resulting algorithm is called V–cycle whereas if $\gamma = 2$, we talk about W–cycle (these denominations come from the schematic representation of the recursion calls). Note that if one solves the coarse system (1.4) by γ iterations of a relevant Krylov scheme using the two-grid method as a preconditioner, one obtains the so-called K-cycle [49].

So far, we have not specified how to construct the (hierarchy of) coarse system(s) (1.4). In case of discretized PDE applications, system matrices of various size can often be generated for the problem at hand. Combining this with geometrical interpolation to pass from the coarser correction $\mathbf{e}_c$ to its finer approximation, we obtain the required ingredients. This approach is known as *geometric multigrid.* It is also possible to construct the multigrid hierarchy in a black box fashion, based only on the knowledge of the system matrix A. Such setup phase is usually called *coarsening* and the black-box multigrid which uses it is *algebraic multigrid.* Whereas it is slower than its geometric counterpart (because of the additional cost of coarsening), algebraic multigrid can be applied to a variety of problems, even those which have no PDE or geometric background.

An excellent introduction to the multigrid techniques can be found in [19]. For more details on practical aspects we refer to [61], whereas a more formal presentation can be found in [67, 27].

1.5 Overview

The remaining five chapters of this thesis treat two essentially different subjects: V-cycle schemes are considered in Chapters 2-4, whereas the aggregation-based coarsening is analyzed in Chapters 5-6. As a matter of paradox, these two multigrid ingredients, when combined together, can hardly lead to an optimal algorithm. Indeed, a V-cycle needs more accurate prolongations than the simple piecewise-constant one, associated to aggregation-based coarsening. On the other hand, aggregation-based approaches use almost exclusively piecewise constant prolongations, and therefore need more involved cycling strategies, K-cycle [49] being an attractive alternative in this respect.

Chapter 2 considers more precisely the well-known V-cycle convergence theories: the approximation property based analyses by Hackbusch [27] and by McCormick [38] and the successive subspace correction theory, as presented in [73] by Xu and in [75] by Yserentant. Under the constraint that the resulting upper bound on the convergence rate must be expressed with respect to parameters involving two successive levels at a time, these theories are compared. Unlike [75], where the comparison is performed on the basis of underlying assumptions in a particular PDE context, we compare directly the upper bounds. We show that these analyses are equivalent from the qualitative point of view. From the quantitative point of view, we show that the bound due to McCormick is always the best one.

When the upper bound on the V-cycle convergence factor involves only two successive levels at a time, it can further be compared with the two-level convergence factor. Such comparison is performed in Chapter 3, showing that a nice two-grid convergence (at every level) leads to an optimal McCormick's bound (the best bound from the previous chapter) if and only if a norm of a given projector is bounded on every level.

In Chapter 4 we consider the Fourier analysis setting for scalar PDEs and extend the comparison between two-grid and V-cycle multigrid methods to the smoothing factor. In particular, a two-sided bound involving the smoothing factor is obtained that defines an interval containing both the two-grid and V-cycle convergence rates. This interval is narrow when an additional parameter α is small enough, this latter being a simple function of Fourier components.

Chapter 5 provides a theoretical framework for coarsening by aggregation. An upper bound is presented that relates the two-grid convergence factor with local quantities, each being related to a particular aggregate. The bound is shown to be asymptotically sharp for a large class of elliptic boundary value problems, including problems with anisotropic and discontinuous coefficients.

In Chapter 6 we consider problems resulting from the discretization with edge finite elements of 3D curl-curl equation. The variables in such discretization are associated with edges. We investigate the performance of the Reitzinger and Schöberl algorithm [52], which uses aggregation techniques to construct the edge prolongation matrix. More precisely, we perform a Fourier analysis of the method in two-grid setting, showing its optimality. The analysis is supplemented with some numerical investigations.

All chapters are independent from each other and can be read in any order. We recommend however the reading of Chapters 2-4 in the ascending order since the results demonstrated in the earlier chapters are used in the following ones.

Chapters 2 through 5 have appeared as separate papers or reports. Their presentation have been only slightly modified in this thesis. In particular, *Chapter 2* corresponds to

> A. Napov and Y. Notay *Comparison of bounds for V-cycle multigrid*,
> published online in *Appl. Numer. Math.*
> DOI: 10.1016/j.apnum.2009.11.003, 2009,

Chapter 3 is taken from

> A. Napov and Y. Notay *When does two-grid optimality carry over to the V-cycle?*, accepted for publication in *Numer. Lin. Alg. Appl.*, 2009,

whereas *Chapter 4* is a slightly modified version of

> A. Napov and Y. Notay *Smoothing factor and actual multigrid convergence*,
> Report GANMN 09-03, Université Libre de Bruxelles, Brussels, Belgium, 2009,

and *Chapter 5* reproduces the content of

> A. Napov and Y. Notay *Algebraic analysis of aggregation-based multigrid*,
> Report GANMN 09-04, Université Libre de Bruxelles, Brussels, Belgium, 2009.

Regarding the Chapter 6, its content is a result of author's collaboration with Ronan Perrussel from Laboratoire Ampère, Ecole Centrale de Lyon. The corresponding paper is still in preparation and the content of this chapter it the author's contribution to the common research. Numerical experiments in the multilevel setting are limited here to the model problem setting; the algebraic multilevel implementation of the presented approach and the related numerical experiments correspond to the contribution of Ronan Perrussel and will appear in the final manuscript.

1.6 Notation

We use bold lowercase Roman letters (e.g., $\mathbf{v}$) to denote vectors and uppercase Roman (e.g., A) to denote matrices. Capital calligraphic letters (e.g., $\mathcal{V}$) represent vector subspaces, except $\mathcal{O}$, which stands for Landau big Oh symbol, and symbols in Chapter 6, which are used to denote Fourier block matrices and index sets.

We use I to denote the identity matrix and O the zero matrix. When the dimensions are not obvious from the context, we write more specifically I_m for the $m \times m$ identity matrix, and $O_{m\times l}$ for the $m \times l$ zero matrix.

For any real α, $\lfloor \alpha \rfloor$ is the largest integer not greater than α. For any set Γ, $|\Gamma|$ is its size. For any real matrix B, $\mathcal{R}(B)$ is the range of B and $\mathcal{N}(B)$ is its null space; B^T stands for its transpose and B^H for its transpose complex conjugate. For any square real matrix C, $\rho(C)$ is its spectral radius (that is, its largest eigenvalue in modulus), $\|C\| = \sqrt{\rho(C^TC)}$ is the usual 2–norm and $\|C\|_{\mathcal{F}} = \sqrt{\sum_{i,j} C_{ij}^2}$ the Frobenius norm. For an SPD matrix D, $\|\mathbf{v}\|_D = \left(\mathbf{v}^T D\mathbf{v}\right)^{1/2} = \|D^{1/2}\mathbf{v}\|$ is the associated D-norm of a vector $\mathbf{v}$ (if $D = A$, it is also called energy norm) and

$$\|C\|_D = \max_{\mathbf{v}} \frac{\|C\mathbf{v}\|_D}{\|\mathbf{v}\|_D} = \|D^{1/2}CD^{-1/2}\|$$

is the induced matrix D-norm.

We finish this section by giving the list of acronyms and the list of symbols below.

List of Acronyms

Acronym	**Meaning**
ARPACK	Arnoldi package [36]
FCG	flexible conjugated gradient
GS	Gauss-Seidel (smoother)
PDE	partial differential equation
RS	Reitzinger and Schöberl multigrid method [52]
SPD	symmetric positive definite
SSC	successive subspace correction

List of Symbols

Symbol	Meaning	Reference
A	generic system matrix	e.g., (1.1)
A_c, A_k	coarse grid matrix, coarse grid matrix on level k	p.86, p.13
c_A	approximation property constant in Hackbush's theory	(2.31)
$E_{MG}^{(k)}$	V-cycle multigrid iteration matrix on kth grid	e.g., (2.2)
E_{TG}, $E_{TG}^{(k)}$	two-grid iteration matrix (between kth and $(k-1)$th grid)	e.g., (3.3)
G_k	auxiliary matrix inducing a decomposition in SSC theory	p.15
h	mesh size on a regular grid	
J	index of the finest level in multilevel setting	p.13
K	parameter in SSC convergence theory	(2.10)
M	Chap. 6: mass matrix in edge-element discretization of (6.2)	p.116
$M^{(\cdot)}$	Chap. 2-4: equivalent pre– or post–smoothing matrix	e.g., (2.3)
n, n_k	size of A, size of A_k	
$n^{(k)}$	Chap. 5: size of kth aggregate	p.86
N	Chap. 2-3 and 5-6: number of grid unknowns in one direction	
$N^{(\cdot)}$	Chap. 4: $I - N_k^{(\nu)} A_k = (I - R_k^{-1} A_k)^\nu$	(4.3)
P, P_k	prolongation matrix (from kth and $(k-1)$th grid)	e.g., p.13
R	smoother matrix of elementary smoothers (e.g., Gauss-Seidel)	e.g., p.13
S	Chap. 5-6: smoothing iteration matrix	e.g., p.119
X	Chap. 5-6: equivalent pre– and post–smoothing matrix	(5.6), (6.13)
α, β	Chap. 4: V-cycle convergence parameters	e.g., (4.14), (4.16)
	Chap. 5-6: PDE coefficients	(5.37),(6.1)
Γ	auxiliary matrix in SSC convergence theory	(2.11)
Γ_k, $\Gamma_{\mathbf{k}}$	aggregate k or $\mathbf{k}$	p.117, p.125
δ	approximation property constant in McCormick's theory	e.g., (2.34)
θ	Chap. 2-4, 6: "frequency" in Fourier analysis	
μ, $\mu^{(k)}$	Chap. 4: smoothing factor (on kth grid)	p.60
	Chap. 5: two-grid quality (of kth aggregate)	p.87, (5.19)
ν	number of smoothing steps	p.13
π_C	projector, generally of the form $P(P^TCP)^{-1}P^TC$	e.g., (3.5)
$\omega^{(\cdot)}$	parameter in V-cycle convergence theories	(2.4)
ω, ω_{Jac}	smoother weighting	
Ω	PDE domain	

Chapter 2

Comparison of bounds for V-cycle multigrid

Summary

We consider multigrid methods with V-cycle for symmetric positive definite linear systems. We compare bounds on the convergence factor that are characterized by a constant which is the maximum over all levels of an expression involving only two consecutive levels. More particularly, we consider the classical bound by Hackbusch, a bound by McCormick, and a bound obtained by applying the successive subspace correction convergence theory with so-called a-orthogonal decomposition. We show that the constants in these bounds are closely related, and hence that these analyses are equivalent from the qualitative point of view. From the quantitative point of view, we show that the bound due to McCormick is always the best one. We also show on an example that it can give satisfactory sharp prediction of actual multigrid convergence.

2.1 Introduction

We consider multigrid methods for solving symmetric positive definite (SPD) $n \times n$ linear systems:

$$A\mathbf{x} = \mathbf{b}. \tag{2.1}$$

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 such as Gauss-Seidel, and a *coarse grid correction*, which involves solving a smaller problem on a coarser grid. A V–cycle multigrid method is obtained when this coarse problem is solved approximately with 1 iteration of the two–grid scheme on that level, and so on, until the coarsest level, where an exact solve is performed. Other cycles may be defined, including the W–cycle based on two recursive applications of the two-grid scheme at each level; see, e.g., [61].

When the system (2.1) stems from the discretization of an elliptic PDE, the V-cycle multigrid has often optimal convergence properties; that is, the convergence is independent of the number of levels and of the mesh discretization parameter h. There are two classical ways for proving this. One way consists in checking the so-called smoothing and approximation properties [10,13,26,27,37,38,53]. Another possibility consists in defining an appropriate subspace decomposition and then analyze the constants involved in the successive subspace correction (SSC) convergence theory [50, 51, 25, 73, 75, 74]. So far, these approaches have only been compared (e.g., in [75]) on the basis of the regularity assumptions that an elliptic boundary value problem should fulfill in order to guarantee optimal bounds for the multigrid method applied to its finite element discretization. This allows only qualitative conclusions which are further restricted to a specific context. For instance, such comparison does not cover V-cycle multigrid for structured linear systems [1]. In fact, a detailed comparison of the convergence theories for V-cycle is difficult because they may be (and have been) formulated diversely. There is some freedom in choosing the subspace decomposition for the SSC convergence theory and there is no unique definition of the smoothing and approximation properties.

The smoothing and approximation property ideas form the basis of the early proofs [10, 13, 26] of h-independent V-cycle convergence. For the case when A is SPD, the classical proof is presented in [27, Theorem 7.2.2] by Hackbusch. The convergence estimate is then characterized by the approximation property constant c_A, which is a maximum over all levels of an expression involving only two consecutive levels.

An alternative approach has been developed by McCormick in [38] (see also [37,53]). Here again, the convergence estimate depends on a constant δ which is a minimum over all levels of an expression involving two consecutive levels.

The SSC convergence theory is more recent and also more general, since by tuning the choice of the space decomposition one can prove some results for elliptic PDEs without requiring regularity assumptions [14]. The comparison with other approaches is not easy because this theory is traditionally formulated in an abstract setting. In this chapter, we first develop an algebraic formulation of the theory, resulting in a bound which also depends on freely chosen quantities. Next, we justify that this degree of freedom seemingly disappears if one adds the constraint that one must be able to assess the main constant in the bound considering only two levels at a time. Note that this latter constraint is not only mandatory to develop the comparison with the other two approaches. It is also very sensible in view of a quantitative analysis, where, as we illustrate on an example, the Fourier analysis setting is used to numerically calculate the bounds and compare them with the actual convergence factor.

Transferred back into the original SSC setting, the choice for which this two-level assessment is possible corresponds to the so-called a-orthogonal decomposition, which is also the decomposition that has been most extensively used when analyzing multigrid methods for the class of (H^2-) regular problems. Then, the bound depends mainly on a constant K and, in this chapter, we show that the three constants c_A, δ and K are in fact closely related, namely

$$K = \max(1, c_A)$$

and

$$\delta^{-1} = c_A^{(2)} \,,$$

where $c_A^{(2)}$ is a Hackbusch approximation property constant for the number of smoothing steps being doubled. Hence the three approaches are qualitatively equivalent, in the sense that they simultaneously succeed or fail to prove optimal convergence. From the quantitative point of view, it further turns out that McCormick's bound is the best one.

The reminder of this chapter is organized as follows. In Section 2.2, we state the general setting of this study and gather the needed assumptions. In Section 2.3, we develop our algebraic variant of the SSC theory and recall the results of Hackbusch and McCormick. The comparison is performed in Section 2.4, and an example is analyzed in Section 2.5.

2.2 General setting

We consider a multigrid method with $J+1$ levels ($J \geq 1$); index J refers to the finest level (on which the system (2.1) is to be solved), and index 0 to the coarsest level. The number of unknowns at level k, $0 \leq k \leq J$, is denoted n_k (hence $n_J = n$).

Our analysis applies to symmetric multigrid schemes based on the Galerkin principle for the SPD system (2.1); that is, restriction is the transpose of prolongation and the matrix A_k at level k, $k = J-1, \ldots, 0$, is given by $A_k = P_k^T A_{k+1} P_k$, where P_k is the prolongation operator from level k to level $k+1$; we also assume that the smoother R_k is SPD and that the number of pre–smoothing steps ν ($\nu > 0$) is equal to the number of post–smoothing steps. The algorithm for V–cycle multigrid is then as follows.

```
Multigrid with V–cycle at level $k$: $\mathbf{x}_{n+1} = \mathrm{MG}(\mathbf{b}, A_k, \mathbf{x}_n, k)$
 (1) Relax $\nu$ times with smoother $R_k$: $\mathbf{x}_n \leftarrow Smooth(\mathbf{x}_n, A_k, R_k, \nu, \mathbf{b})$
 (2) Compute residual: $\mathbf{r}_k = \mathbf{b} - A_k\mathbf{x}_n$
 (3) Restrict residual: $\mathbf{r}_{k-1} = P_{k-1}^T\mathbf{r}_k$
 (4) Coarse grid correction: if $k = 1$, $\mathbf{e}_0 = A_0^{-1}\mathbf{r}_0$
                             else $\mathbf{e}_{k-1} = \mathrm{MG}(\mathbf{r}_{k-1}, A_{k-1}, 0, k-1)$
 (5) Prolongate coarse grid correction: $\mathbf{x}_n \leftarrow \mathbf{x}_n + P_{k-1}\mathbf{e}_{k-1}$
 (6) Relax $\nu$ times with smoother $R_k$: $\mathbf{x}_{n+1} \leftarrow Smooth(\mathbf{x}_n, A_k, R_k, \nu, \mathbf{b})$
```

When applying this algorithm, the error satisfies

$$A_k^{-1}\mathbf{b} - \mathbf{x}_{n+1} = E_{MG}^{(k)}\left(A_k^{-1}\mathbf{b} - \mathbf{x}_n\right)$$

where the iteration matrix $E_{MG}^{(k)}$ is recursively defined from

$$\begin{array}{l} E_{MG}^{(0)} = 0 \quad \text{and, for} \;\; k = 1, 2, \ldots, J \;\; : \\ E_{MG}^{(k)} = (I - R_k^{-1} A_k)^{\nu} \left(I - P_{k-1}(I - E_{MG}^{(k-1)}) A_{k-1}^{-1} P_{k-1}^T A_k \right) (I - R_k^{-1} A_k)^{\nu} \end{array} \tag{2.2}$$

(see, e.g., [61, p. 48]). Our main objective is the analysis of the spectral radius of $E_{MG}^{(J)}$, which governs convergence on the finest level. Our analysis makes use of the following general assumptions.

General assumptions

- $n = n_J > n_{J-1} > ... > n_0$;
- P_k is an $n_{k+1} \times n_k$ matrix of rank n_k , $k = J-1, \ldots, 0$;
- $A_J = A$ and $A_k = P_k^T A_{k+1} P_k$, $k = J-1, \ldots, 0$;
- R_k is SPD and such that $\rho(I - R_k^{-1} A_k) < 1$, $k = J, \ldots, 1$.

Note also that most of our results do not refer explicitly to the smoother R_k , but are stated with respect to the matrices $M_k^{(\nu)}$ defined from

$$I - {M_k^{(\nu)}}^{-1} A_k = (I - R_k^{-1} A_k)^{\nu} \; . \tag{2.3}$$

That is, $M_k^{(\nu)}$ is the smoother that provides in 1 step the same effect as ν steps with R_k . The results stated with respect to $M_k^{(\nu)}$ may then be seen as results stated for the case of 1 pre– and 1 post–smoothing step, which can be extended to the general case via the relations (2.3).

Most results depend on the following parameter:

$$\omega^{(\nu)} = \max\left(1 \, , \, \max_{1 \le k \le J} \; \max_{\mathbf{w}_k \in \mathbb{R}^{n_k}} \; \frac{\mathbf{w}_k^T A_k \mathbf{w}_k}{\mathbf{w}_k^T \, M_k^{(\nu)} \, \mathbf{w}_k} \right) . \tag{2.4}$$

From $\rho(I - R_k^{-1} A_k) < 1$, it follows that $\omega^{(1)} < 2$, whereas (2.3) implies

$$\omega^{(\nu)} = \left\{ \begin{array}{cl} 1 & \text{if } \nu \text{ is even} \\ 1 + (\omega^{(1)} - 1)^{\nu} & \text{if } \nu \text{ is odd.} \end{array} \right. \tag{2.5}$$

Hence one has also $\omega^{(\nu)} < 2$ for all ν. Further, if $\omega^{(1)} = 1$, then $\omega^{(\nu)} = 1$ for all ν.

We close this subsection by introducing the projector π_{A_k} which plays an important role throughout this chapter:

$$\pi_{A_k} \;\; = \;\; P_{k-1} A_{k-1}^{-1} P_{k-1}^T A_k \; . \tag{2.6}$$

2.3 Bounds on the V-cycle multigrid convergence factor

2.3.1 SSC theory

We consider the SSC convergence analysis as presented in Theorem 4.4 and Lemma 4.6 in [73], and Theorem 5.1 in [75]. Of course, there are more recent versions of this theory, e.g., in [74] an *identity* (known as *XZ-identity*) is obtained which provides the exact convergence factor. However, we do not see how to transform these further versions so that, according to the focus of this chapter, they deliver a bound that could be assessed considering only two levels at a time (while being significantly different from the bound given by Theorem 2.1 together with Theorem 2.3). In particular, it seems clear that the exact convergence factor is a global quantity whose knowledge necessarily involves information from all levels. Note that SSC ideas are also treated in an algebraic setting in [65, Section 5], where both the XZ-identity and approximation property approaches are presented, without however comparing them.

Now, we first develop in Theorem 2.1 below an algebraic version of Theorem 5.1 in [75]. We give a complete proof since this version slightly improves the original formulation, which uses a matrix Γ with the same entries in the strict upper part, but non-negative entries in the strict lower part and positive entries on the diagonal.

Observe that in Theorem 2.1 below the freedom left in choosing the pseudo restrictions G_k corresponds, in the original formulation, to the freedom associated with the choice of the space decomposition. More precisely, given a set of G_k, $k = 0, \ldots, J-1$, we can construct a corresponding space decomposition as defined in [75]. In Appendix A we show that the converse is also true; that is, with any admissible space decomposition in the original theory, one may associate a set of pseudo restrictions G_k such that Theorem 2.1 will yield the same bound as Theorem 5.1 in [75], except for the improvement associated with the refined definition of Γ.

Theorem 2.1. *Let $E_{MG}^{(J)}$ be defined by* (2.2) *with P_k , $k = 0, \ldots, J-1$, A_k , $k = 0, \ldots, J$, and R_k , $k = 1, \ldots, J$, satisfying the general assumptions stated in Section 2.2. For $k = 1, \ldots, J$, let $M_k^{(\nu)}$ be defined by* (2.3)*, and set $M_0^{(\nu)} = A_0$.*

Let G_k , $k = 0, \ldots, J-1$, be $n_k \times n_{k+1}$ matrices, and, for $k = 0, \ldots, J$, let $\check{P}_k$ and $\check{G}_k$ be defined by, respectively,

$$\begin{aligned} &\check{P}_J = I \\ &\check{P}_k = \check{P}_{k+1}\, P_k \ , \quad k = J-1, \ldots, 0 \ , \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} &\check{G}_J = I \\ &\check{G}_k = G_k\, \check{G}_{k+1} \ , \quad k = J-1, \ldots, 0 \ , \end{aligned} \tag{2.8}$$

with $P_{-1} = G_{-1} = O$ *.*

There holds

$$\rho(E_{MG}^{(J)}) \;\le\; 1 - \frac{2-\omega^{(\nu)}}{K^{(\nu)}(1+\|\Gamma\|)^2} \;, \tag{2.9}$$

where $\omega^{(\nu)}$ *is defined by* (2.4)*,*

$$K^{(\nu)} \;=\; \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\sum_{k=0}^{J} \mathbf{v}^T\check{G}_k^T(I-P_{k-1}G_{k-1})^T\, M_k^{(\nu)}\,(I-P_{k-1}G_{k-1})\check{G}_k\mathbf{v}}{\mathbf{v}^TA\mathbf{v}} \;, \tag{2.10}$$

and

$$\Gamma \;=\; \begin{pmatrix} 0 & \gamma_{01} & \cdots & & \gamma_{0J} \\ & 0 & \cdots & & \gamma_{1J} \\ & & \ddots & & \vdots \\ & & & 0 & \gamma_{(J-1)J} \\ & & & & 0 \end{pmatrix} \;, \tag{2.11}$$

with, for $k = 0,\dots,J-1$ *and* $l = k+1,\dots,J$ *,*

$$\gamma_{kl} \;=\; \max_{\mathbf{w}_k\in\mathbb{R}^{n_k}} \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T\check{G}_l^T(I-P_{l-1}G_{l-1})^T\check{P}_l^TA\check{P}_k\mathbf{w}_k}{(\mathbf{w}_k^T\,M_k^{(\nu)}\,\mathbf{w}_k)^{1/2}(\mathbf{v}^T\check{G}_l^T(I-P_{l-1}G_{l-1})^T\,M_l^{(\nu)}\,(I-P_{l-1}G_{l-1})\check{G}_l\mathbf{v})^{1/2}} \;. \tag{2.12}$$

Moreover,

$$\|\Gamma\| \;\le\; \omega^{(\nu)}\sqrt{J(J+1)/2} \;. \tag{2.13}$$

Proof. In what follows, we omit the superscript (ν) in $M_k^{(\nu)}$. We first gather some useful definitions:

$$Q_k = (I-P_{k-1}G_{k-1})\check{G}_k \qquad , \quad k = 0,\dots,J \;; \tag{2.14}$$

$$T_k = \check{P}_k(M_k)^{-1}\check{P}_k^TA \qquad , \quad k = 0,\dots,J \;; \tag{2.15}$$

$$F_k = (I-T_k)(I-T_{k-1})\cdots(I-T_1)(I-T_0) \qquad , \quad k = 0,\dots,J \;. \tag{2.16}$$

In addition we set $F_{-1} = I$.

As shown in [65, Proposition 5.1.1] there holds

$$E_{MG}^{(J)} = (I-T_J)(I-T_{J-1})\dots(I-T_1)(I-T_0)(I-T_1)\dots(I-T_{J-1})(I-T_J) \;.$$

Further, since $A^{-1}(I-T_k)^T = (I-T_k)A^{-1}$ and $(I-T_0)^2 = I-T_0$, one has $E_{MG}^{(J)} = F_JA^{-1}F_J^TA$, showing that

$$\rho(E_{MG}^{(J)}) = \|F_J\|_A^2 = \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\|F_J\mathbf{v}\|_A^2}{\mathbf{v}^TA\mathbf{v}} \;. \tag{2.17}$$

Using this relation, we first show that (2.9) holds if

$$\mathbf{v}^T A\mathbf{v} \leq K\,(1+\|\Gamma\|)^2\,\left(\sum_{l=0}^{J}\mathbf{v}^T F_{l-1}^T A T_l F_{l-1}\mathbf{v}\right)\qquad \forall \mathbf{v}\in\mathbb{R}^n\ . \tag{2.18}$$

Indeed, since $AT_k = T_k^T A$ and using (2.4), one has, $\forall \mathbf{v}\in\mathbb{R}^n$,

$$\begin{aligned}||F_{k-1}\mathbf{v}||_A^2 - ||F_k\mathbf{v}||_A^2 &= (F_{k-1}\mathbf{v})^T AF_{k-1}\mathbf{v} - (F_{k-1}\mathbf{v})^T(I-T_k)^T A(I-T_k)F_{k-1}\mathbf{v}\\ &= 2\mathbf{v}^T F_{k-1}^T AT_kF_{k-1}\mathbf{v} - (F_{k-1}\mathbf{v})^T T_k^T AT_k(F_{k-1}\mathbf{v})\\ &= 2\mathbf{v}^T F_{k-1}^T AT_kF_{k-1}\mathbf{v} - (F_{k-1}\mathbf{v})^T A\check{P}_kM_k^{-1}\check{P}_k^T A\check{P}_kM_k^{-1}\check{P}_k^T A(F_{k-1}\mathbf{v})\\ &= 2\mathbf{v}^T F_{k-1}^T AT_kF_{k-1}\mathbf{v} - (F_{k-1}\mathbf{v})^T A\check{P}_kM_k^{-1}A_kM_k^{-1}\check{P}_k^T A(F_{k-1}\mathbf{v})\\ &\geq 2\mathbf{v}^T F_{k-1}^T AT_kF_{k-1}\mathbf{v} - \omega^{(\nu)}\ (F_{k-1}\mathbf{v})^T A\check{P}_kM_k^{-1}\check{P}_k^T A(F_{k-1}\mathbf{v})\\ &= (2-\omega^{(\nu)})\ \mathbf{v}^T F_{k-1}^T AT_kF_{k-1}\mathbf{v}\ .\end{aligned}$$

Summing both sides for $k=0,\dots,J$ shows that, $\forall \mathbf{v}\in\mathbb{R}^n$,

$$\|\mathbf{v}\|_A^2 - \|F_J\mathbf{v}\|_A^2 \geq (2-\omega^{(\nu)})\left(\sum_{l=0}^{J}\mathbf{v}^T F_{l-1}^T AT_lF_{l-1}\mathbf{v}\right)\ ,$$

and it is straightforward to check that this relation, together with (2.18) and (2.17), implies (2.9).

We now prove (2.18). Observe that, using (2.14), there holds

$$\sum_{l=0}^{J}\check{P}_lQ_l\ =\ \sum_{l=0}^{J}\check{P}_l(I-P_{l-1}G_{l-1})\check{G}_l\ =\ \sum_{l=0}^{J}\left(\check{P}_l\check{G}_l-\check{P}_{l-1}\check{G}_{l-1}\right)\ =\ \check{P}_J\check{G}_J-\check{P}_{-1}\check{G}_{-1}\ =\ I\ .$$

For any $\mathbf{v}\in\mathbb{R}^n$, one may then decompose $\mathbf{v}^T A\mathbf{v}$ as the sum of two terms (remembering that $F_{-1}=I$):

$$\mathbf{v}^T A\mathbf{v}\ =\ \sum_{l=0}^{J}\mathbf{v}^T A\check{P}_lQ_l\mathbf{v}\ =\ \sum_{l=0}^{J}\mathbf{v}^T F_{l-1}^T A\check{P}_lQ_l\mathbf{v}\ +\ \sum_{l=1}^{J}\mathbf{v}^T(I-F_{l-1}^T)A\check{P}_lQ_l\mathbf{v}\ . \tag{2.19}$$

In order to prove (2.18), we bound separately the two terms in the right hand side of (2.19).

Regarding the first term, one has, applying twice the Cauchy-Schwartz inequality,

$$\begin{aligned}\sum_{l=0}^{J}\mathbf{v}^T F_{l-1}^T A\check{P}_lQ_l\mathbf{v} &\leq \sum_{l=0}^{J}(\mathbf{v}^T Q_l^T M_lQ_l\mathbf{v})^{1/2}(\mathbf{v}^T F_{l-1}^T A\check{P}_lM_l^{-1}\check{P}_l^T AF_{l-1}\mathbf{v})^{1/2}\\ &\leq \left(\sum_{l=0}^{J}\mathbf{v}^T Q_l^T M_lQ_l\mathbf{v}\right)^{1/2}\left(\sum_{l=0}^{J}\mathbf{v}^T F_{l-1}^T AT_lF_{l-1}\mathbf{v}\right)^{1/2}\ .\end{aligned} \tag{2.20}$$

To estimate the second term, first observe that

$$I - F_{l-1} = I - (I - T_{l-1})F_{l-2} = (I - F_{l-2}) + T_{l-1}F_{l-2} = \cdots = \sum_{k=0}^{l-1} T_k F_{k-1} \ .$$

Therefore,

$$\sum_{l=1}^{J} \mathbf{v}^T (I - F_{l-1}^T) A\check{P}_l Q_l \mathbf{v} = \sum_{l=1}^{J}\sum_{k=0}^{l-1} \mathbf{v}^T F_{k-1}^T T_k^T A \check{P}_l Q_l \mathbf{v} \ ,$$

whereas, for any $0 \leq k < l \leq J$, using successively (2.15) and (2.12) with $\mathbf{w}_k = M_k^{-1}\check{P}_k^T A F_{k-1}\mathbf{v}$,

$$\begin{aligned}
\mathbf{v}^T F_{k-1}^T T_k^T A \check{P}_l Q_l \mathbf{v} &= (\mathbf{v}^T F_{k-1}^T A \check{P}_k M_k^{-1}) \check{P}_k^T A \check{P}_l Q_l \mathbf{v} \\
&\leq \gamma_{kl} (\mathbf{v}^T Q_l^T M_l Q_l \mathbf{v})^{1/2} (\mathbf{v}^T F_{k-1}^T A \check{P}_k M_k^{-1} \check{P}_k^T A F_{k-1} \mathbf{v})^{1/2} \\
&= \gamma_{kl} (\mathbf{v}^T Q_l^T M_l Q_l \mathbf{v})^{1/2} (\mathbf{v}^T F_{k-1}^T A T_k F_{k-1} \mathbf{v})^{1/2} .
\end{aligned}$$

Hence, since $\|\Gamma\| = \max_{\mathbf{y}} \frac{\|\Gamma \mathbf{y}\|}{\|\mathbf{y}\|} = \max_{\mathbf{x},\mathbf{y}} \frac{\mathbf{x}^T \Gamma \mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$ and using the definition (2.11) of Γ, there holds

$$\begin{aligned}
\sum_{l=1}^{J} \mathbf{v}^T (I - F_{l-1}^T) A \check{P}_l Q_l \mathbf{v} &\leq \sum_{l=1}^{J}\sum_{k=0}^{l-1} \gamma_{kl} (\mathbf{v}^T Q_l^T M_l Q_l \mathbf{v})^{1/2} (\mathbf{v}^T F_{k-1}^T A T_k F_{k-1} \mathbf{v})^{1/2} \\
&\leq \|\Gamma\| \left(\sum_{l=0}^{J} \mathbf{v}^T Q_l^T M_l Q_l \mathbf{v} \right)^{1/2} \left(\sum_{k=0}^{J} \mathbf{v}^T F_{k-1}^T A T_k F_{k-1} \mathbf{v} \right)^{1/2} .
\end{aligned}$$

Combining the latter result with (2.20), one gets

$$\mathbf{v}^T A \mathbf{v} \ \leq \ (1 + \|\Gamma\|) \left(\sum_{l=0}^{J} \mathbf{v}^T Q_l^T M_l Q_l \mathbf{v} \right)^{1/2} \left(\sum_{l=0}^{J} \mathbf{v}^T F_{l-1}^T A T_l F_{l-1} \mathbf{v} \right)^{1/2} .$$

Taking the square of both sides, and using (2.10) (which amounts to $\sum_{l=0}^{J} \mathbf{v}^T Q_l^T M_l Q_l \mathbf{v} \leq K\, \mathbf{v}^T A \mathbf{v}$) straightforwardly leads to (2.18), which completes the proof of (2.9).

It remains to prove (2.13). Note that $\|\Gamma\| \leq \|\Gamma\|_{\mathcal{F}} = \left(\sum_{l=1}^{J} \sum_{k=0}^{l-1} \gamma_{kl}^2 \right)^{1/2}$. Further, for any $0 \leq k < l \leq J$ and for any $\mathbf{w} \in \mathbb{R}^n$ and $\mathbf{w}_k \in \mathbb{R}^{n_k}$,

$$\begin{aligned}
\mathbf{w}^T Q_l^T \check{P}_l^T A \check{P}_k \mathbf{w}_k &\leq (\mathbf{w}^T Q_l^T \check{P}_l^T A \check{P}_l Q_l \mathbf{w})^{1/2} (\mathbf{w}_k^T \check{P}_k^T A \check{P}_k \mathbf{w}_k)^{1/2} \\
&= (\mathbf{w}^T Q_l^T A_l Q_l \mathbf{w})^{1/2} (\mathbf{w}_k^T A_k \mathbf{w}_k)^{1/2} \\
&\leq \omega^{(\nu)} \, (\mathbf{w}^T Q_l^T M_l Q_l \mathbf{w})^{1/2} \, (\mathbf{w}_k^T M_k \mathbf{w}_k)^{1/2} \ ,
\end{aligned}$$

showing that $\gamma_{kl} \leq \omega^{(\nu)}$. The required result straightforwardly follows. ■

Now, in this chapter, we focus on bounds that can be estimated considering only two consecutive levels at a time. The following theorem helps to see when the main constant

$K^{(\nu)}$ in Theorem 2.1 can be set in that form.

Theorem 2.2. *Let $\check{P}_k$ and $\check{G}_k$ be defined by* (2.7) *and* (2.8) *with P_k , $k = 0, \ldots, J-1$ and A_k , $k = 0, \ldots, J$, satisfying the general assumptions stated in Section 2.2. Then, for all $\mathbf{v} \in \mathbb{R}^n$*

$$\begin{aligned}
\mathbf{v}^T A\mathbf{v} &= \sum_{k=0}^{J} \mathbf{v}^T \check{G}_k^T (I - P_{k-1}G_{k-1})^T A_k (I - P_{k-1}G_{k-1})\check{G}_k \mathbf{v} && (2.21)\\
&\quad + \; 2\sum_{k=0}^{J} \mathbf{v}^T \check{G}_{k-1}^T P_{k-1}^T A_k \left(I - P_{k-1}G_{k-1}\right) \check{G}_k \mathbf{v} \\
&= \sum_{k=0}^{J} \mathbf{v}^T \check{G}_k^T (I - P_{k-1}G_{k-1})^T A_k (I + P_{k-1}G_{k-1})\check{G}_k \mathbf{v} \; . && (2.22)
\end{aligned}$$

Moreover, if $P_{k-1}G_{k-1}$ is a projector, then

$$(I - P_{k-1}G_{k-1})^T A_k (I + P_{k-1}G_{k-1}) \tag{2.23}$$

is nonnegative definite if and only if

$$G_{k-1} = A_{k-1}^{-1} P_{k-1}^T A_k. \tag{2.24}$$

Proof. We begin, noting that $\mathbf{v}_k^T A_k P_{k-1} G_{k-1} \mathbf{v}_k = (\mathbf{v}_k^T A_k P_{k-1} G_{k-1} \mathbf{v}_k)^T = \mathbf{v}_k^T (P_{k-1}G_{k-1})^T A_k \mathbf{v}_k$ holds for all $\mathbf{v}_k \in \mathbb{R}^{n_k}$. Using this relation with $\mathbf{v}_k = \check{G}_k \mathbf{v}$, equations (2.21) and (2.22) follow from

$$\begin{aligned}
&\sum_{k=0}^{J} \mathbf{v}^T \check{G}_k^T (I + P_{k-1}G_{k-1})^T A_k (I - P_{k-1}G_{k-1})\check{G}_k \mathbf{v} \\
&\qquad = \sum_{k=0}^{J} \left(\mathbf{v}^T \check{G}_k^T \check{P}_k^T A \check{P}_k \check{G}_k \mathbf{v} - \mathbf{v}^T \check{G}_{k-1}^T \check{P}_{k-1}^T A \check{P}_{k-1} \check{G}_{k-1} \mathbf{v} \right) \\
&\qquad = \mathbf{v}^T A \mathbf{v} \, .
\end{aligned}$$

Next, $(I - P_{k-1}G_{k-1})^T A_k (I + P_{k-1}G_{k-1})$ is nonnegative definite if and only if

$$\mathbf{v}_k^T (I - P_{k-1}G_{k-1})^T A_k (I + P_{k-1}G_{k-1}) \mathbf{v}_k \geq 0 \;\; \forall \mathbf{v}_k \in \mathbb{R}^{n_k}$$

which in turn is equivalent to

$$\mathbf{v}_k^T A_k \mathbf{v}_k \geq \mathbf{v}_k^T (P_{k-1}G_{k-1})^T A_k P_{k-1} G_{k-1} \mathbf{v}_k \;\; \forall \mathbf{v}_k \in \mathbb{R}^{n_k} \, ,$$

this latter being nothing else but

$$\|P_{k-1}G_{k-1}\|_{A_k} \;\leq\; 1.$$

Hence, if $P_{k-1}G_{k-1}$ is a projector, it has to be orthogonal, and, hence, symmetric with respect to the $(\,\cdot\,, A_k\,\cdot\,)$ inner product (see [39, Section 5.13]); that is, $P_{k-1}G_{k-1} = B_k A_k$ for some symmetric B_k. This implies $G_{k-1} = C_{k-1}P_{k-1}^T A_k$ with C_{k-1} symmetric. Since P_{k-1} has full rank, $P_{k-1}G_{k-1}$ is then a projector if and only if $C_{k-1} = A_{k-1}^{-1}$; hence the required result. ∎

Now, consider the definition (2.10) of $K^{(\nu)}$. To obtain an expression that can be assessed considering only two levels at a time, the only possibility we have found is to express the denominator $\mathbf{v}^T A\mathbf{v}$ as a sum over all levels similar to the sum in the numerator, and, assuming each term involved to be non-negative, to bound the ratio of both these sums $\sum_k a_k/\sum_k b_k$ by the maximum of the ratios $\max_k(a_k/b_k)$. The first result of Theorem 2.2 tells us that such a splitting of $\mathbf{v}^T A\mathbf{v}$ always exists, but the second result tells us that it is exploitable only with $G_{k-1} = A_{k-1}^{-1}P_{k-1}^T A_k$, since otherwise there would be negative terms in the sum of the denominator, at least for certain $\mathbf{v}$.[1] Note that these G_k are such that $P_{k-1}G_{k-1} = \pi_{A_k}$ and correspond to the so-called a-orthogonal decomposition in the original abstract theory. This choice is further analyzed in the following theorem, where we prove in particular that one has then $\Gamma = 0$. Note that with the original formulation of [75, Theorem 5.1], one could only prove $\|\Gamma\| \leq \omega^{(\nu)}$.

Theorem 2.3. *Let the assumptions of Theorem 2.1 hold, and let G_k , $k = 0,\dots,J-1$, be defined by* (2.24)*. Then, $K^{(\nu)}$ and Γ , defined as in Theorem 2.1, satisfy, respectively*

$$K^{(\nu)} = \max\left(1,\; \max_{1\leq k\leq J}\; \max_{\mathbf{w}_k\in\mathbb{R}^{n_k}} \frac{\mathbf{w}_k^T(I-\pi_{A_k})^T\, M_k^{(\nu)}\, (I-\pi_{A_k})\mathbf{w}_k}{\mathbf{w}_k^T(I-\pi_{A_k})^T A_k (I-\pi_{A_k})\mathbf{w}_k}\right) \tag{2.25}$$

$$= \max\left(1,\; \max_{1\leq k\leq J}\; \max_{\mathbf{w}_k\in\mathbb{R}^{n_k}} \frac{\mathbf{w}_k^T(I-\pi_{A_k})^T\, M_k^{(\nu)}\, (I-\pi_{A_k})\mathbf{w}_k}{\mathbf{w}_k^T A_k\mathbf{w}_k}\right) \tag{2.26}$$

and

$$\Gamma = 0\ , \tag{2.27}$$

where π_{A_k} is defined by (2.6)*.*

[1] Theorem 3.2 proves this under the additional assumption that P_kG_k is a projector, but we did not found any usable bound based on G_k for which P_kG_k would not be a projector.

Proof. We first prove (2.27). Note that (2.24) implies $\check{G}_l = A_l^{-1}\check{P}_l^T A,\ l = 0, ..., J-1$. Hence, for any $0 \le k < l \le J$ and all $\mathbf{w}_k \in \mathbb{R}^{n_k}$, $\mathbf{v} \in \mathbb{R}^n$,

$$\begin{aligned}
\mathbf{w}_k^T \check{P}_k^T A \check{P}_l (I - P_{l-1}G_{l-1})\check{G}_l \mathbf{v} &= \mathbf{w}_k^T \check{P}_k^T A \check{P}_l A_l^{-1} \check{P}_l^T A v - \mathbf{w}_k^T \check{P}_k^T A \check{P}_{l-1} A_{l-1}^{-1} \check{P}_{l-1}^T A \mathbf{v} \\
&= \mathbf{w}_k^T P_k^T \cdots P_{l-1}^T \left(\check{P}_l^T A \check{P}_l A_l^{-1} \right) \check{P}_l^T A \mathbf{v} \\
&\quad - \mathbf{w}_k^T P_k^T \cdots P_{l-2}^T \left(\check{P}_{l-1}^T A \check{P}_{l-1} A_{l-1}^{-1} \right) \check{P}_{l-1}^T A \mathbf{v} \\
&= \mathbf{w}_k^T P_k^T \cdots P_{l-1}^T \check{P}_l^T A v - \mathbf{w}_k^T P_k^T \cdots P_{l-2}^T \check{P}_{l-1}^T A \mathbf{v} \\
&= \mathbf{w}_k^T \check{P}_k^T A v - \mathbf{w}_k^T \check{P}_k^T A \mathbf{v} \\
&= 0 \ ;
\end{aligned}$$

$\gamma_{kl} = 0$ and therefore $\Gamma = 0$ readily follows.

We next prove (2.25) and (2.26). Using (2.22) and $P_{k-1}G_{k-1} = \pi_{A_k}$ together with $(I\ +\ \pi_{A_k})^T\ A_k\ (I\ -\ \pi_{A_k})\ = (I - \pi_{A_k})^T A_k (I - \pi_{A_k})$ in the definition (2.10) of $K^{(\nu)}$, one has

$$\begin{aligned}
K^{(\nu)} &= \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\sum_{k=0}^J\ \mathbf{v}^T \check{G}_k^T (I - P_{k-1}G_{k-1})^T\ M_k^{(\nu)}\ (I - P_{k-1}G_{k-1})\check{G}_k \mathbf{v}}{\sum_{k=0}^J\ \mathbf{v}^T \check{G}_k^T (I - P_{k-1}G_{k-1})^T A_k (I - P_{k-1}G_{k-1})\check{G}_k \mathbf{v}} \qquad (2.28) \\
&= \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\sum_{k=1}^J\ \mathbf{v}^T \check{G}_k^T (I - \pi_{A_k})^T\ M_k^{(\nu)}\ (I - \pi_{A_k})\check{G}_k \mathbf{v}\ +\ \mathbf{v}^T \check{G}_0^T A_0 \check{G}_0 \mathbf{v}}{\sum_{k=1}^J\ \mathbf{v}^T \check{G}_k^T (I - \pi_{A_k})^T A_k (I - \pi_{A_k})\check{G}_k \mathbf{v}\ +\ \mathbf{v}^T \check{G}_0^T A_0 \check{G}_0 \mathbf{v}} \\
&\le \max\left(1\,,\ \max_{1\le k\le J}\ \max_{\mathbf{w}_k\in\mathbb{R}^{n_k}} \frac{\mathbf{w}_k^T (I - \pi_{A_k})^T\ M_k^{(\nu)}\ (I - \pi_{A_k})\mathbf{w}_k}{\mathbf{w}_k^T (I - \pi_{A_k})^T A_k (I - \pi_{A_k})\mathbf{w}_k} \right) .
\end{aligned}$$

This proves that the right hand side of (2.25) is an upper bound on $K^{(\nu)}$; the right hand side of (2.26) is a further upper bound since

$$\max_{\mathbf{w}_k\in\mathbb{R}^{n_k}} \frac{\mathbf{w}_k^T (I - \pi_{A_k})^T M_k^{(\nu)} (I - \pi_{A_k})\mathbf{w}_k}{\mathbf{w}_k^T A_k \mathbf{w}_k} \ \ge\ \max_{\mathbf{v}_k\in\mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T (I - \pi_{A_k})^T M_k^{(\nu)} (I - \pi_{A_k})\mathbf{v}_k}{\mathbf{v}_k^T (I - \pi_{A_k})^T A_k (I - \pi_{A_k})\mathbf{v}_k}\ ,$$

as seen by restricting the maximum in the left hand side to $\mathbf{w}_k = (I - \pi_{A_k})\mathbf{v}_k$ (taking into account that $(I - \pi_{A_k})^2 = (I - \pi_{A_k})$).

To prove that the right hand sides of (2.25), (2.26) are also lower bounds on $K^{(\nu)}$, let, for $k = 0, \dots, J$, $\check{Q}_k = (I - P_{k-1}G_{k-1})\check{G}_k$. Then rewrite (2.28) as

$$K^{(\nu)} = \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\sum_{k=0}^J \mathbf{v}^T \check{Q}_k^T\ M_k^{(\nu)}\ \check{Q}_k \mathbf{v}}{\sum_{k=0}^J \mathbf{v}^T \check{Q}_k^T A_k \check{Q}_k \mathbf{v}}. \qquad (2.29)$$

Since $G_k P_k = I_{n_k}$ for $k = 0, \dots, J-1$, Lemma 2.1 in Appendix B proves that, for $0 \le l, k \le J$ with $k \ne l$,

$$\check{Q}_l \check{P}_l \check{Q}_l = \check{Q}_l \quad \text{and} \quad \check{Q}_k \check{P}_l \check{Q}_l = O_{n_k\times n}\ .$$

Restricting the maximum in (2.29) to $\mathbf{v} = \check{P}_l \check{Q}_l \mathbf{w}$ for some $0 \leq l \leq J$ yields

$$\begin{aligned} K^{(\nu)} &\geq \max_{\mathbf{w}\in\mathbb{R}^n} \frac{\mathbf{w}^T \check{Q}_l^T \, M_l^{(\nu)} \, \check{Q}_l \mathbf{w}}{\mathbf{w}^T \check{Q}_l^T A_l \check{Q}_l \mathbf{w}} \\ &= \max_{\mathbf{w}\in\mathbb{R}^n} \frac{\mathbf{w}^T \check{G}_l^T (I - P_{l-1}G_{l-1})^T \, M_l^{(\nu)} \, (I - P_{l-1}G_{l-1}) \check{G}_l \mathbf{w}}{\mathbf{w}^T \check{G}_l^T (I - P_{l-1}G_{l-1})^T A_l (I - P_{l-1}G_{l-1}) \check{G}_l \mathbf{w}} \\ &= \max_{\mathbf{w}_l\in\mathbb{R}^{n_l}} \frac{\mathbf{w}_l^T (I - P_{l-1}G_{l-1})^T \, M_l^{(\nu)} \, (I - P_{l-1}G_{l-1}) \mathbf{w}_l}{\mathbf{w}_l^T (I - P_{l-1}G_{l-1})^T A_l (I - P_{l-1}G_{l-1}) \mathbf{w}_l} \ , \end{aligned}$$

the last equality stemming from the fact that G_l, and hence $\check{G}_l$, has full rank (from (2.24), (2.8), and because P_k has full rank by virtue of our general assumptions). The conclusion follows because

$$\begin{aligned} \mathbf{w}_l^T (I - P_{l-1}G_{l-1})^T A_l (I - P_{l-1}G_{l-1}) \mathbf{w}_l &= \mathbf{w}_l^T (I - \pi_{A_l})^T A_l (I - \pi_{A_l}) \mathbf{w}_l \\ &= \mathbf{w}_l^T (A_l - A_l P_{l-1} A_{l-1}^{-1} P_{l-1}^T A_l) \mathbf{w}_l \\ &\leq \mathbf{w}_l^T A_l \mathbf{w}_l \ . \quad \blacksquare \end{aligned}$$

2.3.2 Hackbusch bound

The bound from [27, Theorem 7.2.2] is recalled in the following theorem. Note that this analysis requires $\omega^{(\nu)} = 1$. This condition is however not too restrictive since the smoother can be scaled to satisfy it. Note also that, according to (2.5), $\omega^{(\nu)} = 1$ always holds for ν even, and that $\omega^{(1)} = 1$ entails $\omega^{(\nu)} = 1$ for all ν.

Theorem 2.4. *Let $E_{MG}^{(J)}$ be defined by (2.2) with P_k , $k = 0, \ldots, J-1$, A_k , $k = 0, \ldots, J$, and R_k , $k = 1, \ldots, J$, satisfying the general assumptions stated in Section 2.2. For $k = 1, \ldots, J$, let $M_k^{(\nu)}$ and $\omega^{(\nu)}$ be defined, respectively, by (2.3) and (2.4).*

Then, if $\omega^{(\nu)} = 1$,

$$\rho(E_{MG}^{(J)}) \ \leq \ \frac{c_A^{(\nu)}}{c_A^{(\nu)} + 2} \ , \tag{2.30}$$

where

$$c_A^{(\nu)} \ = \ \max_{1\leq k\leq J} \ \max_{\mathbf{v}_k\in\mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T (A_k^{-1} - P_{k-1} A_{k-1}^{-1} P_{k-1}^T) \mathbf{v}_k}{\mathbf{v}_k^T \, {M_k^{(\nu)}}^{-1} \, \mathbf{v}_k} \ . \tag{2.31}$$

Moreover, if $\omega^{(1)} = 1$,

$$\rho(E_{MG}^{(J)}) \ \leq \ \frac{c_A^{(1)}}{c_A^{(1)} + 2\nu} \ . \tag{2.32}$$

Note that Theorem 7.2.2 in [27] considers only (2.32). The bound (2.30) is a straightforward extension (through the replacement of $M_k^{(1)} = R_k$ by $M_k^{(\nu)}$) that will make easier the comparison with other approaches. It is not really useful in practice since, as will be seen, (2.32) is always better than (2.30). Note, however, that (2.30) is more general since one may have $\omega^{(\nu)} = 1$ while $\omega^{(1)} > 1$.

Note also that in [27] some bounds based on c_A are also proved for the W and two-grid cycle, that are better than those obtained by using just the V-cycle bound as a worst case estimate.

2.3.3 McCormick's bound

We recall in the following theorem the bound obtained in [38, Lemma 2.3, Theorem 3.4 and Section 5] (see also [37], or [53] for an alternative proof).

Theorem 2.5. *Let $E_{MG}^{(J)}$ be defined by* (2.2) *with P_k , $k = 0, \dots, J-1$, A_k , $k = 0, \dots, J$, and R_k , $k = 1, \dots, J$, satisfying the general assumptions stated in Section 2.2. For $k = 1, \dots, J$, let $M_k^{(\nu)}$ be defined by* (2.3)*.*

Then,

$$\rho(E_{MG}^{(J)}) \;\leq\; 1 - \delta^{(\nu)} \quad , \tag{2.33}$$

where

$$\delta^{(\nu)} \;=\; \min_{1\leq k\leq J} \; \min_{\mathbf{v}_k\in\mathbb{R}^{n_k}} \frac{\|\mathbf{v}_k\|_{A_k}^2 - \|(I - M_k^{(\nu)\,-1} A_k)\mathbf{v}_k\|_{A_k}^2}{\|(I-\pi_{A_k})\mathbf{v}_k\|_{A_k}^2} \tag{2.34}$$

with π_{A_k} defined by (2.6)*.*

Moreover,

$$\delta^{(\nu)\,-1} \leq \frac{1}{\nu}\left(\delta^{(1)\,-1} + \nu - 1\right) . \tag{2.35}$$

2.4 Comparison

We first state our main result, which relates the constants $K^{(\nu)}$, $c_A^{(\nu)}$ and $\delta^{(\nu)}$.

Theorem 2.6. *Let $K^{(\nu)}$, $c_A^{(\nu)}$ and $\delta^{(\nu)}$ be defined respectively by* (2.25)*,* (2.31) *and* (2.34) *where P_k , $k = 0, \dots, J-1$, A_k , $k = 0, \dots, J$, and R_k , $k = 1, \dots, J$ satisfy the general assumptions stated in Section 2.2. For $k = 1, \dots, J$, let $M_k^{(\nu)}$ be defined by* (2.3)*.*

Then

$$K^{(\nu)} = \max(\ 1,\ c_A^{(\nu)}) \,, \tag{2.36}$$

and

$$\delta^{(\nu)} = \frac{1}{c_A^{(2\nu)}} \,. \tag{2.37}$$

Proof. Let

$$\tilde{P}_k = A_k^{1/2} P_{k-1} A_{k-1}^{-1/2} \qquad , \quad k = 1, \dots, J \ .$$

One has

$$\begin{aligned}
c_A^{(\nu)} &= \max_{1\le k\le J}\max_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^T(A_k^{-1}-P_{k-1}A_{k-1}^{-1}P_{k-1}^T)\mathbf{v}}{\mathbf{v}^T\ M_k^{(\nu)\ -1}\ \mathbf{v}} \\
&= \max_{1\le k\le J}\max_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^T(I-A_k^{1/2}P_{k-1}A_{k-1}^{-1}P_{k-1}^T A_k^{1/2})\mathbf{v}}{\mathbf{v}^T A_k^{1/2}\ M_k^{(\nu)\ -1}\ A_k^{1/2}\mathbf{v}} \\
&= \max_{1\le k\le J}\max_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^T(I-\tilde{P}_k\tilde{P}_k^T)\mathbf{v}}{\mathbf{v}^T A_k^{1/2}\ M_k^{(\nu)\ -1}\ A_k^{1/2}\mathbf{v}} \\
&= \max_{1\le k\le J}\max_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^T\ M_k^{(\nu)\ 1/2}\ A_k^{-1/2}(I-\tilde{P}_k\tilde{P}_k^T)^2 A_k^{-1/2}\ M_k^{(\nu)\ 1/2}\ \mathbf{v}}{\mathbf{v}^T\mathbf{v}} \\
&= \max_{1\le k\le J}\max_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^T(I-\tilde{P}_k\tilde{P}_k^T)A_k^{-1/2}\ M_k^{(\nu)}\ A_k^{-1/2}(I-\tilde{P}_k\tilde{P}_k^T)\mathbf{v}}{\mathbf{v}^T\mathbf{v}}\,.
\end{aligned}$$

Since $(I-\tilde{P}_k\tilde{P}_k^T)A_k^{-1/2} = (I-A_k^{1/2}P_{k-1}A_{k-1}^{-1}P_{k-1}^T A_k^{1/2})A_k^{-1/2} = A_k^{-1/2}(I-\pi_{A_k})^T$, this leads to

$$c_A^{(\nu)} = \max_{1\le k\le J}\max_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^T(I-\pi_{A_k})^T\ M_k^{(\nu)}\ (I-\pi_{A_k})\mathbf{v}}{\mathbf{v}^T A_k\mathbf{v}},$$

hence (2.36).

On the other hand, observing that $M_k^{(2\nu)}$ satisfies

$$I - M_k^{(2\nu)\ -1}A_k = (I - M_k^{(\nu)\ -1}A_k)^2 \quad,\quad k=1,\dots,J\ ,$$

one has

$$\begin{aligned}
\delta^{(\nu)} &= \min_{1\le k\le J}\min_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{||\mathbf{v}||_{A_k}^2 - ||I-M_k^{(\nu)\ -1}A_k\mathbf{v}||_{A_k}^2}{||(I-\pi_{A_k})\mathbf{v}||_{A_k}^2} \\
&= \min_{1\le k\le J}\min_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^TA_k\mathbf{v} - \mathbf{v}^T\left(I-M_k^{(\nu)\ -1}A_k\right)^T A_k\left(I-M_k^{(\nu)\ -1}A_k\right)\mathbf{v}}{\mathbf{v}^T(I-\pi_{A_k})^TA_k(I-\pi_{A_k})\mathbf{v}} \\
&= \min_{1\le k\le J}\min_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^TA_k\mathbf{v} - \mathbf{v}^TA_k\left(I-M_k^{(\nu)\ -1}A_k\right)^2\mathbf{v}}{\mathbf{v}^T(I-\pi_{A_k})^TA_k(I-\pi_{A_k})\mathbf{v}} \\
&= \min_{1\le k\le J}\min_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^TA_k\mathbf{v} - \mathbf{v}^TA_k\left(I-M_k^{(2\nu)\ -1}A_k\right)\mathbf{v}}{\mathbf{v}^T(I-\pi_{A_k})^TA_k(I-\pi_{A_k})\mathbf{v}} \\
&= \min_{1\le k\le J}\min_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^TA_k\ M_k^{(2\nu)\ -1}\ A_k\mathbf{v}}{\mathbf{v}^TA_k(I-\pi_{A_k})\mathbf{v}} \\
&= \min_{1\le k\le J}\min_{\mathbf{v}\in\mathbb{R}^{n_k}} \frac{\mathbf{v}^T\ M_k^{(2\nu)\ -1}\ \mathbf{v}}{\mathbf{v}^T(I-\pi_{A_k})A_k^{-1}\mathbf{v}} \\
&= \frac{1}{c_A^{(2\nu)}}\,. \qquad \blacksquare
\end{aligned}$$

We are now ready to compare the bounds (2.9), (2.30), (2.32) and (2.33). This is done in the following theorem.

Theorem 2.7. *Let $E_{MG}^{(J)}$ be defined by* (2.2) *with P_k , $k = 0, \ldots, J-1$, A_k , $k = 0, \ldots, J$, and R_k , $k = 1, \ldots, J$, satisfying the general assumptions stated in Section 2.2.For $k = 1, \ldots, J$, let $M_k^{(\nu)}$ and $\omega^{(\nu)}$ be defined, respectively, by* (2.3) *and* (2.4)*. Moreover, let $K^{(\nu)}$, $c_A^{(\nu)}$ and $\delta^{(\nu)}$ be defined respectively by* (2.25), (2.31) *and* (2.34)*.*

Then

$$\rho(E_{MG}^{(J)}) \;\le 1 - \delta^{(\nu)} \;\le 1 - \frac{2 - \omega^{(\nu)}}{K^{(\nu)}} \,. \tag{2.38}$$

Further, if $\omega^{(\nu)} = 1$,

$$\rho(E_{MG}^{(J)}) \;\le 1 - \delta^{(\nu)} \;\le \frac{c_A^{(\nu)}}{c_A^{(\nu)} + 2} \,, \tag{2.39}$$

and, if $\omega^{(1)} = 1$,

$$\rho(E_{MG}^{(J)}) \;\le 1 - \delta^{(\nu)} \;\le\; \frac{c_A^{(1)}}{c_A^{(1)} + 2\nu} \;\le\; \frac{c_A^{(\nu)}}{c_A^{(\nu)} + 2} \,. \tag{2.40}$$

Moreover,

$$1 - \frac{2 - \omega^{(\nu)}}{K^{(\nu)}} \le 1 - \frac{2 - \omega^{(\nu)}}{2}\, \delta^{(\nu)} \;, \tag{2.41}$$

and, if $\omega^{(\nu)} = 1$,

$$\frac{c_A^{(\nu)}}{c_A^{(\nu)} + 2} \;\le\; \frac{1}{\delta^{(\nu)} + 1} \;=\; 1 - \frac{\delta^{(\nu)}}{\delta^{(\nu)} + 1} \,. \tag{2.42}$$

Proof. Let us first prove two intermediate results:

$$\frac{c_A^{(\nu)}}{2} \le c_A^{(2\nu)} \le \frac{c_A^{(\nu)}}{2 - \omega^{(\nu)}} \tag{2.43}$$

and, if $\omega^{(\mu)} = 1$,

$$\frac{c_A^{(\mu)}}{\nu} \le c_A^{(\mu\nu)} \le \frac{1}{\nu}\left(c_A^{(\mu)} + \nu - 1\right) \;,\; \mu \in \mathbb{N}_0^+ . \tag{2.44}$$

The first intermediate result (2.43) follows from

$$M_k^{(2\nu)} \;=\; M_k^{(\nu)} \left(2\, M_k^{(\nu)} - A_k\right)^{-1} M_k^{(\nu)}$$

combined with

$$2\mathbf{v}_k^T\, M_k^{(\nu)}\, \mathbf{v}_k \ge 2\mathbf{v}_k^T\, M_k^{(\nu)}\, \mathbf{v}_k - \mathbf{v}_k^T A_k \mathbf{v}_k \ge (2 - \omega^{(\nu)})\mathbf{v}_k^T\, M_k^{(\nu)}\, \mathbf{v}_k \,,\;\; \forall \mathbf{v}_k \in \mathbb{R}^{n_k} .$$

We prove the second intermediate result (2.44) for $\mu = 1$; its generalization to $\mu > 1$ is performed replacing R_k by $M_k^{(\mu)}$ in the proof below. First, the right inequality (2.44) is a consequence of (2.35) since, using (2.37) one has

$$c_A^{(\nu)} = \delta^{(\nu/2)\,-1} \le \frac{1}{\nu}\left(\delta^{(1/2)\,-1} + \nu - 1\right) = \frac{1}{\nu}\left(c_A^{(1)} + \nu - 1\right)$$

where $\delta^{(1/2)}$ corresponds to the V-cycle algorithm with a smoother $\widetilde{R}_k$ such that

$$I - R_k^{-1}A_k = (I - \widetilde{R}_k^{-1}A_k)^2\,.$$

Such $\widetilde{R}_k$ is indeed well defined since $\omega^{(1)} = 1$ entails that $I - A_k^{1/2}R_k^{-1}A_k^{1/2}$ is symmetric nonnegative definite. On the other hand, the left inequality (2.44) is a straightforward consequence of

$$\mathbf{v}_k^T\, M_k^{(\nu)\,-1}\,\mathbf{v}_k \le \nu\ \mathbf{v}_k^T R_k^{-1}\mathbf{v}_k\,,\ \ \forall \mathbf{v}_k \in \mathbb{R}^{n_k}$$

which we prove as follows. This relation holds if and only if

$$\mathbf{v}_k^T A_k^{1/2}\ M_k^{(\nu)\,-1}\ A_k^{1/2}\mathbf{v}_k \le \nu\ \mathbf{v}_k^T A_k^{1/2} R_k^{-1} A_k^{1/2}\mathbf{v}_k\,, \forall \mathbf{v}_k \in \mathbb{R}^{n_k}$$

which, in view of (2.3) and when $\omega^{(1)} = 1$, is satisfied if

$$1 - (1-x)^\nu \le \nu x \qquad \forall x \in [0,1];$$

that is, if, $\forall \lambda = 1 - x \in [0,1)$,

$$\frac{1-\lambda^\nu}{1-\lambda} \le \nu\,,$$

which is readily checked from $\frac{1-\lambda^\nu}{1-\lambda} = \sum_{i=0}^{\nu-1}\lambda^i < \nu\,.$

Now, the second inequality (2.38) follows from the right inequality (2.43) combined with (2.36) and (2.37). The second inequalities (2.39) and (2.40) are equivalent to, respectively

$$c_A^{(\nu)} c_A^{(2\nu)} \ge (c_A^{(\nu)} + 2)(c_A^{(2\nu)} - 1)$$

and

$$c_A^{(1)} c_A^{(2\nu)} \ge (c_A^{(1)} + 2\nu)(c_A^{(2\nu)} - 1)\,.$$

These inequalities follow from the right inequality (2.44), used with $(\mu\,,\,\nu) = (\nu\,,\,2)$ and $(\mu\,,\,\nu) = (1\,,\,2\nu)$, respectively, combined with (2.37). Next, the last inequality of (2.40) is a consequence of the left inequality of (2.44) used with $(\mu\,,\,\nu) = (1\,,\,\nu)$. Finally, inequalities (2.41) and (2.42) follow from the left inequality (2.43) combined with (2.37) and (2.36), because $\delta^{(\nu)\,-1} \ge 1$, as may be seen from

$$\begin{aligned}\delta^{(\nu)\,-1} &= c^{(2\nu)}\\ &= \max_{1\le k\le J}\ \max_{\mathbf{w}_k\in\mathbb{R}^{n_k}}\ \frac{\mathbf{w}_k^T(I-\pi_{A_k})^T\ M_k^{(2\nu)}\ (I-\pi_{A_k})\mathbf{w}_k}{\mathbf{w}_k^T(I-\pi_{A_k})^T A_k (I-\pi_{A_k})\mathbf{w}_k}\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{\omega^{(2\nu)}} \max_{1\leq k\leq J} \max_{\mathbf{w}_k\in\mathbb{R}^{n_k}} \frac{\mathbf{w}_k^T(I-\pi_{A_k})^T\, M_k^{(2\nu)}\, (I-\pi_{A_k})\mathbf{w}_k}{\mathbf{w}_k^T(I-\pi_{A_k})^T\, M_k^{(2\nu)}\, (I-\pi_{A_k})\mathbf{w}_k} \\
&= 1\,. \qquad \blacksquare
\end{aligned}$$

From (2.38), (2.39) and (2.40), one sees that McCormick's bound is always the best one, whereas inequalities (2.41) and (2.42) show that all approaches are nevertheless qualitatively equivalent, since they give bounds which, at worst, correspond to McCormick's bound with main constant smaller by a modest factor.

2.5 Example

We consider the linear system resulting from the 9-point finite difference discretization of the two-dimensional Poisson problem

$$\begin{aligned}
-\Delta\, u = f \quad &\text{in } \ \Omega = (0,1)\times(0,1) \\
u = 0 \quad &\text{in } \ \partial\Omega
\end{aligned}$$

on a uniform grid of mesh size $h = 1/N_J$ in both directions. The matrix corresponds then, up to some scaling factor, to the following nine point stencil

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} . \tag{2.45}$$

We assume $N_J = 2^J N_0$ for some integer N_0, allowing J steps of regular geometric coarsening. We consider prolongations in form of the standard interpolation associated with bilinear finite element basis functions. The restriction P_k^T corresponds then to "full weighting", as defined in, e.g. [61] [2]. With these choices, the stencil (2.45) is preserved throughout all grids (up to some unimportant scaling factor), and $c_A^{(\nu)}$ may be assessed by analyzing

$$\max_{\mathbf{w}_k} \frac{\mathbf{w}_k^T(I-\pi_{A_k})^T M_k^{(\nu)}(I-\pi_{A_k})\mathbf{w}_k}{\mathbf{w}_k A_k \mathbf{w}_k} \tag{2.46}$$

for a matrix A_k corresponding to stencil (2.45) applied on a grid with mesh size $h_k = 1/N_k$. Considering two successive grids is therefore sufficient, and, to alleviate notation, we let $N = N_k$, $A = A_k$, $M^{(\nu)} = M_k^{(\nu)}$, $P = P_{k-1}$, $A_c = A_{k-1} = P^T A P$ and $\pi_A = \pi_{A_k} = P A_c^{-1} P^T A$.

[2] up to some scaling factor; the scalings of the prolongation and restriction are unimportant when using coarse grid matrices of the Galerkin type.

To assess (2.46), we resort to Fourier analysis. The eigenvectors of A are, for $m, l = 1, \ldots, N - 1$, the functions

$$u_{m,l}^{(N)} = \sin(m\pi x)\,\sin(l\pi y)$$

evaluated at the grid points. The eigenvalue corresponding to $\mathbf{u}_{m,l}^{(N)}$ is

$$\lambda_{m,l}^{(N)} = 4(3s_m + 3s_l - 4s_m s_l) \tag{2.47}$$

where

$$s_m = \sin^2(m\pi/2N) \quad , \quad s_l = \sin^2(l\pi/2N)\ . \tag{2.48}$$

The prolongation P satisfies (see, e.g., [61, p. 87])

$$P^T \left\{ \begin{matrix} u_{m,l}^{(N)} \\ u_{N-m,N-l}^{(N)} \\ -u_{N-m,l}^{(N)} \\ -u_{m,N-l}^{(N)} \end{matrix} \right\} = 4 \left\{ \begin{matrix} (1-s_m)(1-s_l) \\ s_m s_l \\ s_m(1-s_l) \\ (1-s_m)s_l \end{matrix} \right\} u_{m,l}^{(N/2)}$$

for $1 \le m, l \le N/2 - 1$, with $P^T u_{m,l}^{(N)} = 0$ for $m = N/2$ or $m = N/2$. Expressed in the Fourier basis (that is, in the basis of eigenvectors of A), $I - \pi_A$ is therefore block diagonal with, for $1 \le m, l \le N/2 - 1$, 4×4 blocks

$$(I - \pi_A)_{m,l} = I_4 - P_{m,l}\left(A_{m,l}^{(c)}\right)^{-1} P_{m,l}^T A_{m,l} \tag{2.49}$$

where

$$\begin{aligned} P_{m,l}^T &= 4\Big(\ (1-s_m)(1-s_l) \quad s_m s_l \quad s_m(1-s_l) \quad (1-s_m)s_l\ \Big) \\ A_{m,l} &= \operatorname{diag}\left(\lambda_{m,l}^{(N)}\,,\ \lambda_{N-m,N-l}^{(N)}\,,\ \lambda_{m,N-l}^{(N)}\,,\ \lambda_{N-m,l}^{(N)}\right) \\ A_{m,l}^{(c)} &= P_{m,l}^T A_{m,l} P_{m,l} = 64\big(3s_m(1-s_m) + 3s_l(1-s_l) - 16s_l(1-s_l)s_m(1-s_m)\big)\ . \end{aligned}$$

For $m = N/2$, $1 \le l \le N/2 - 1$ and $l = N/2$, $1 \le m \le N/2 - 1$, $(I - \pi_A)_{m,l} = I_2$ is a 2×2 identity block, whereas $(I - \pi_A)_{\frac{N}{2},\frac{N}{2}} = 1$ reduces to the scalar identity. If $M^{(\nu)}$ in the Fourier basis has the same block diagonal structure, we are left with the analysis of

$$\rho_{m,l} = \rho\left((I - \pi_A)_{m,l}^T M_{m,l}^{(\nu)} (I - \pi_A)_{m,l} A_{m,l}^{-1}\right)\ . \tag{2.50}$$

Now, we consider more specifically damped Jacobi smoothing; that is $R_k = \omega_{Jac}^{-1}\operatorname{diag}(A) = \omega_{Jac}^{-1}\, 8\, I$, with $\omega_{Jac} \in (\,0\,,\,4/3\,)$ to ensure $\omega^{(1)} = (3/2)\omega_{Jac} < 2$. Then, for any number of pre– and post–smoothing steps ν, $M^{(\nu)}$ is diagonal in the Fourier basis, with diagonal entries depending on the eigenvalues of A; that is (see (2.47)), depending on s_m and s_l .To obtain grid independent bounds, it is then interesting to consider $\rho_{m,l} = \rho(s_m, s_l)$ as a function of s_m , s_l , and to let these parameters vary continuously in $[0,1]$, excluding the corner points where $s_m(1-s_m) = s_l(1-s_l) = 0$, which correspond to singularities. For all ν, $\rho(s_m, s_l)$ has the following symmetries: $\rho(s_m, s_l) = \rho(1-s_m, s_l) = \rho(s_m, 1-s_l) = \rho(1-s_m, 1-s_l)$. Further, numerical investigations reveal that the maximum on the considered domain is located at the boundary, i.e., corresponds to, e.g., $s_m = 0$. Because of the symmetries it is sufficient to analyze this latter case. One may check that $\rho(0, s_l)$ is the largest eigenvalue in modulus of

$$\frac{1}{4}\begin{pmatrix} \frac{s_l\mu_1+s_l\mu_4}{3} & 0 & 0 & -\frac{s_l\mu_1+s_l\mu_4}{3} \\ 0 & \frac{\mu_2}{3-(1-s_l)} & 0 & 0 \\ 0 & 0 & \frac{\mu_3}{3-s_l} & 0 \\ -\frac{\mu_1(1-s_l)+\mu_4(1-s_l)}{3} & 0 & 0 & \frac{(1-s_l)\mu_1+(1-s_l)\mu_4}{3} \end{pmatrix} ,$$

where $\{\mu_i\}_{i=1,\dots,4}$ are the 4 diagonal entries of $M_{kl}^{(\nu)}$, given by

$$\mu_i = \frac{(A_{m,l})_{i,i}}{1-(1-\frac{\omega_{Jac}}{2}(A_{m,l})_{i,i})^{\nu}} .$$

Thus

$$\rho(0, s_l) = \max\left(\frac{\mu_3}{3-s_l}, \frac{\mu_2}{3-(1-s_l)}, \frac{\mu_1+\mu_4}{3}\right) ,$$

and, injecting the expressions of μ_i,

$$\rho(0, s_l) = \max\left(\frac{1}{1-(1-\frac{\omega_{Jac}}{2}(3-s_l))^{(\nu)}}, \frac{1}{1-(1-\frac{\omega_{Jac}}{2}(2+s_l))^{\nu}}, \right.$$
$$\left. \frac{s_l}{1-(1-\frac{3\omega_{Jac}}{2}s_l)^{\nu}} + \frac{1-s_l}{1-(1-\frac{3\omega_{Jac}}{2}(1-s_l))^{\nu}}\right) .$$

Note that for $s_l \to 0$ the third term is larger that the maximum over s_l of the first and the second; hence

$$\rho(0, s_l) \le \sup_{s_l\in(0,1)}\left(\frac{s_l}{1-(1-\frac{3\omega_{Jac}}{2}s_l)^{\nu}} + \frac{1-s_l}{1-(1-\frac{3\omega_{Jac}}{2}(1-s_l))^{\nu}}\right) . \tag{2.51}$$

The right hand side of (2.51) is in fact independent of s_l for $\nu = 1$, and, for $\nu = 2$ and $\nu = 4$, one may check, using elementary function analysis (see Appendix B), that the supremum is reached for $s_l \to 0, 1$. Hence

$$c_A^{(\nu)} \le \frac{2}{3\nu\omega_{Jac}} + \frac{1}{1-(1-\frac{3\omega_{Jac}}{2})^{\nu}} \quad , \quad \nu = 1, 2, 4. \tag{2.52}$$

Using the relation (2.52) as an equality, we can evaluate the different bounds. This is

ω_{Jac}	$\omega^{(1)}$	$c_A^{(1)}$	$c_A^{(2)}$	$\frac{c_A^{(1)}}{c_A^{(1)}+2}$	$1-\frac{2-\omega^{(1)}}{K^{(1)}}$	$1-\delta^{(1)}$	$\rho(E_{MG}^{(J)})$
1/2	1	2.666	1.733	0.571	0.626	0.423	0.398
2/3	1	2	1.5	0.5	0.5	0.333	0.271
1	1.5	1.333	1.666	(*)	0.5	0.387	0.251

TABLE 2.1: Convergence factor of V–cycle (for $N_0 = 2$ and $J = 6$) and the corresponding bounds for $\nu = 1$; (*) the quantity exists, but does not correspond to the bound, since $\omega^{(1)} > 1$.

ω_{Jac}	$\omega^{(2)}$	$c_A^{(1)}$	$c_A^{(2)}$	$c_A^{(4)}$	$\frac{c_A^{(1)}}{c_A^{(1)}+4}$	$\frac{c_A^{(2)}}{c_A^{(2)}+2}$	$1-\frac{2-\omega^{(2)}}{K^{(2)}}$	$1-\delta^{(2)}$	$\rho(E_{MG}^{(J)})$
1/2	1	2.666	1.733	1.337	0.4	0.4	0.423	0.252	0.187
2/3	1	2	1.5	1.25	0.333	0.333	0.333	0.2	0.121
1	1	1.333	1.666	1.233	(*)	0.25	0.4	0.189	0.091

TABLE 2.2: Convergence factor of V–cycle (for $N_0 = 2$ and $J = 6$) and the corresponding bounds for $\nu = 2$; (*) the quantity exists, but does not correspond to the bound, since $\omega^{(1)} > 1$.

done in Table 2.1 and 2.2 for different number ν of smoothing steps, where we also compare the bounds with the actual convergence factor. One sees that McCormick's bound is indeed the best one and, further, that it gives in the considered cases a satisfactory sharp prediction of actual multigrid convergence.

2.6 Conclusion

We have considered different bounds on the V-cycle multigrid convergence factor, each depending on a parameter given by the maximum over all levels of a expression defined on two levels only. More precisely, we have considered the bound in [27, Theorem 7.2.2] by Hackbusch, the result [38, Lemma 2.3, Theorem 3.4 and Section 5] of McCormick and the Successive Subspace Correction theory [73, Theorem 4.4 and Lemma 4.6], [75, Theorem 5.1] used with a-orthogonal decomposition. Regarding the latter approach, it has been adapted here to the algebraic framework and slightly improved. We have sown that the main parameters of these three theories are related to each other and that the corresponding bounds are equivalent from the qualitative point of view; that is, they simultaneously succeed or fail to prove an optimal convergence for a given problem. From the quantitative viewpoint, we have proved that the bound of McCormick is the sharpest, and, further, that it leads to an accurate convergence estimate at least for a typical example.

Appendix A

We first show that Theorem 5.1 in [75] particularized to the matrix case (that is, applied to the case of matrix operators in $\mathbb{R}^n$ with $a(\mathbf{v},\mathbf{w}) = (\mathbf{v}, A\mathbf{w}) = \mathbf{v}^T A\mathbf{w}$) yields the same bound as Theorem 2.1 (except for the additional refinement in the definition of $\|\Gamma\|$), provided that one has $\mathcal{W}_k = \mathcal{R}(\check{P}_k)$ and $\mathcal{V}_k = \mathcal{R}(\check{P}_k\check{G}_k - \check{P}_{k-1}\check{G}_{k-1})$, where $\check{P}_k$ and $\check{G}_k$ refer to the notation in Theorem 2.1, and $\mathcal{W}_k, \mathcal{V}_k$ to notation in [75].

Firstly, note that Theorem 5.1 provides a bound on the energy norm of product iteration matrices of the form (2.16), where

$$T_k = B_k^+ Q_k A \,, \tag{2.53}$$

B_k^+ being a matrix corresponding to a invertible operator onto $\mathcal{W}_k$, and Q_k being the orthogonal projector on the subspace $\mathcal{W}_k = \mathcal{R}(\check{P}_k)$; that is, $Q_k = \check{P}_k(\check{P}_k^T\check{P}_k)^{-1}\check{P}_k^T$. It then follows that the definition (2.53) matches (2.15) by setting $B_k^+ = \check{P}_k M_k^{-1}\check{P}_k^T$. Observe also that, $\forall \mathbf{w}_k \in \mathcal{W}_k$,

$$\mathbf{z}_k = B_k^+\mathbf{w}_k \;\Leftrightarrow\; \mathbf{w}_k = \check{P}_k(\check{P}_k^T\check{P}_k)^{-1}M_k(\check{P}_k^T\check{P}_k)^{-1}\check{P}_k^T\mathbf{z}_k \,.$$

Hence

$$B_k = \check{P}_k(\check{P}_k^T\check{P}_k)^{-1}M_k(\check{P}_k^T\check{P}_k)^{-1}\check{P}_k^T \tag{2.54}$$

is the proper inverse of B_k^+ onto $\mathcal{W}_k$.

Next, the bound on $\|F_J\|_A^2$ in [75] is based on the decomposition of any vector $\mathbf{v} \in \mathbb{R}^n$ as

$$\mathbf{v} = \sum_{k=0}^{J} \mathbf{v}_k$$

where $\mathbf{v}_k \in \mathcal{V}_k$. With $\mathcal{V}_k = \mathcal{R}(\check{P}_k\check{G}_k - \check{P}_{k-1}\check{G}_{k-1})$, it means

$$\mathbf{v}_k = \check{P}_k(I - P_{k-1}G_{k-1})\check{G}_k\mathbf{v} = (\check{P}_k\check{G}_k - \check{P}_{k-1}\check{G}_{k-1})\mathbf{v}. \tag{2.55}$$

Then, the bound in [75] is

$$\|F_J\|_A^2 \;\leq\; 1 - \frac{2-\omega}{K_1(1+K_2)^2} \,, \tag{2.56}$$

where K_1 is such that

$$\sum_{k=0}^{J}(B_k\mathbf{v}_k, \mathbf{v}_k) \leq K_1\mathbf{v}^T A\mathbf{v} \quad \forall\, \mathbf{v} \in \mathbb{R}^n \,, \tag{2.57}$$

where ω satisfy

$$(A\mathbf{w}_k, \mathbf{w}_k) \leq \omega(B_k\mathbf{w}_k, \mathbf{w}_k) \quad \forall\, \mathbf{w}_k \in \mathcal{W}_k\,, k = 1, ..., J\,, \tag{2.58}$$

and where $K_2 = \|\widetilde{\Gamma}\|$, with $\widetilde{\Gamma} = (\widetilde{\gamma}_{kl})$ being the $(J+1)\times(J+1)$ matrix whose coefficients are such that

$$(A\mathbf{w}_k, \mathbf{v}_l) \leq \widetilde{\gamma}_{kl}(B_k\mathbf{w}_k, \mathbf{w}_k)^{1/2}(B_l\mathbf{v}_l, \mathbf{v}_l)^{1/2} \quad \forall\, \mathbf{v}_k \in \mathcal{V}_k\,,\ \mathbf{w}_k \in \mathcal{W}_k \tag{2.59}$$

for $k \leq l$, and $\widetilde{\gamma}_{kl} = \widetilde{\gamma}_{lk}$ for $k > l$.

With (2.54) and (2.55), it is easy to recognize that $K^{(\nu)}$ in (2.10) is the best constant K_1 satisfying (2.57). On the other hand, " $\forall \mathbf{w}_k \in \mathcal{W}_k$ " means " for all $\mathbf{w}_k = \check{P}_k\mathbf{w}$ with $\mathbf{w} \in \mathbb{R}^n$ " and " $\forall \mathbf{v}_k \in \mathcal{V}_k$ " means " for all $\mathbf{v}_k = \check{P}_k(I - P_{k-1}G_{k-1})\check{G}_k\mathbf{v}$ with $\mathbf{v} \in \mathbb{R}^n$ ". Hence, for $k < l$, γ_{kl} in (2.12) is the best $\widetilde{\gamma}_{kl}$ satisfying (2.59). Further, using the same arguments, we see that $\omega^{(\nu)}$ is the best choice for ω. Therefore, the equivalence between the bound (2.56) in [75] and (2.9) is proved, except for the additional refinement showing that the lower triangular part of Γ can be set to zero.

We next show that with any admissible choice of $\mathcal{V}_k$, one may associate valid G_k, $k = 0, ..., J$ such that $\mathcal{V}_k = \mathcal{R}(\check{P}_k\check{G}_k - \check{P}_{k-1}\check{G}_{k-1})$ (setting $P_{-1} = G_{-1} = O_{n_0\times n_0}$). In other words, any bound from Theorem 5.1 in [75] obtained using a particular decomposition can also be obtained via (2.9) (up to some additional refinement in the definition of $\|\Gamma\|$) using a particular set of matrices G_k.

We begin the proof letting

$$\mathcal{X}_k = \mathcal{V}_0 \oplus \mathcal{V}_1 \oplus \ldots \oplus \mathcal{V}_k\,.$$

Observe that the proposition holds if, given $\mathcal{X}_0 \subset \mathcal{X}_1 \subset \ldots \subset \mathcal{X}_J = \mathbb{R}^n$, one can find G_k, $k = 0, ..., J$ such that

$$\mathcal{R}(\check{P}_k\check{G}_k) = \mathcal{X}_k \tag{2.60}$$

and

$$\mathcal{R}(\check{P}_k\check{G}_k - \check{P}_{k-1}\check{G}_{k-1}) \ \cap\ \mathcal{R}(\check{P}_{k-1}\check{G}_{k-1}) = \{0\}.$$

The latter equality is checked if, for all $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^n$,

$$\left(\check{P}_k\check{G}_k - \check{P}_{k-1}\check{G}_{k-1}\right)\mathbf{v} \ = \ \check{P}_{k-1}\check{G}_{k-1}\mathbf{w} \ \Rightarrow \ \left(\check{P}_k\check{G}_k - \check{P}_{k-1}\check{G}_{k-1}\right)\mathbf{v} \ = \ \check{P}_{k-1}\check{G}_{k-1}\mathbf{w} = 0\,;$$

that is, since $\check{P}_k$ has full rank, if

$$\begin{aligned} (I - P_{k-1}G_{k-1})\left(\check{G}_k\mathbf{v}\right) \ &= \ P_{k-1}G_{k-1}\left(\check{G}_k\mathbf{w}\right) \\ \Rightarrow \ (I - P_{k-1}G_{k-1})\left(\check{G}_k\mathbf{v}\right) \ &= \ P_{k-1}G_{k-1}\left(\check{G}_k\mathbf{w}\right) = 0. \end{aligned} \tag{2.61}$$

This proposition is true when $P_{k-1}G_{k-1}$ is a projector (note that $P_{-1}G_{-1} = O_{n_0\times n_0}$ is a projector as well). The right equalities (2.61) follow then from the multiplication of (2.61) by $(I - P_{k-1}G_{k-1})$ and $P_{k-1}G_{k-1}$, respectively.

We now assume that $\check{G}_j$ has been constructed properly for $j = J, ..., k+1$ (which holds trivially for $j = J-1$), and show that one can construct G_k such that

$$\mathcal{R}(\check{P}_k G_k \check{G}_{k+1}) = \mathcal{X}_k \tag{2.62}$$

while satisfying the constraint

$$G_k P_k G_k = G_k\,, \tag{2.63}$$

yielding the required result by induction, since (2.63) implies $(P_k G_k)^2 = P_k G_k$.

Let $m_k = \dim(\mathcal{X}_k)$. Observe that $\mathcal{W}_0 \subset \ldots \subset \mathcal{W}_k$ implies $m_k \leq \dim(\mathcal{W}_k) = n_k$. Hence (2.62) holds if $\mathcal{G}_k = \mathcal{R}(\check{G}_k)$ is a prescribed m_k-dimensional subspace of $\mathbb{R}^{n_k}$ whose image by $\check{P}_k$ is $\mathcal{X}_k$. Let H_k be an $n_k \times m_k$ matrix whose columns form a basis of this subspace. We search for G_k of the form

$$G_k = H_k Z_k\,,$$

where Z_k is an $m_k \times n_{k+1}$ matrix of rank m_k. Then (2.62) holds if $Z_k\check{G}_{k+1}$ has rank m_k, which is ensured if $\mathcal{R}(\check{G}_{k+1})$ contains an m_k-dimensional subspace complementary to $\mathcal{N}(Z_k)$ (see [39, p. 199]). Note that $\dim(\mathcal{R}(\check{G}_{k+1})) = \dim(\mathcal{X}_{k+1}) \geq m_k$, hence there exists at least one m_k-dimensional subspace $\mathcal{G}_k$ of $\mathcal{R}(\check{G}_{k+1})$, and we shall enforce the null space of Z_k to be complementary to $\mathcal{G}_k$.

Consider now the constraint (2.63). With the given form of G_k, it is satisfied when

$$Z_k P_k H_k = I_{m_k}\,;$$

that is, according to the terminology in [6], if Z_k is a $\{1,2\}$-inverse of $P_k H_k$. As shown in [6, p. 59], given any subspace $\mathcal{S}_k$ complementary to $\mathcal{T}_k = \mathcal{R}(P_k H_k)$ there exist such a $\{1,2\}$-inverse having $\mathcal{S}_k$ as a null space.

Hence the required result is proven if one can always find $\mathcal{S}_k$ complementary to both $\mathcal{G}_k$ and $\mathcal{T}_k$. This, in turn, is true since $\mathcal{G}_k$ and $\mathcal{T}_k$ are subspaces of the same dimension of a finite dimensional space $\mathbb{R}^{n_k}$, see [34].

Appendix B

Lemma 2.1. *Let P_k , $k = 0, \ldots, J-1$ be $n_{k+1} \times n_k$ matrices of rank n_k with $n = n_J > n_{J-1} > \cdots > n_0$. Let G_k , $k = 0, \ldots, J-1$ be $n_{k+1} \times n_k$ matrices such that*

$$G_k\, P_k = I_{n_k} \ .$$

Set $P_{-1} = G_{-1} = O_{n_0 \times n_0}$ and let, for $k = 0, \ldots, J$, $\check{P}_k$ be defined by (2.7), $\check{G}_k$ be defined by (2.8), and $\check{Q}_k = (I - P_{k-1}G_{k-1})\check{G}_k$.

There holds, for $0 \le l, k \le J$ with $k \neq l$,

$$\check{Q}_k \check{P}_k \check{Q}_k = \check{Q}_k \quad \text{and} \quad \check{Q}_l \check{P}_k \check{Q}_k = O_{n_l \times n} \ .$$

Proof. Note that $G_k\, P_k = I_{n_k}$ implies $\check{G}_k \check{P}_k = I_{n_k}$. The first statement follows then from

$$\begin{aligned}(I - P_{k-1}G_{k-1})\check{G}_k\check{P}_k(I - P_{k-1}G_{k-1}) &= (I - P_{k-1}G_{k-1})(I - P_{k-1}G_{k-1}) \\ &= I - P_{k-1}G_{k-1} \ .\end{aligned}$$

To prove the second statement, we consider two cases. If $l > k$,

$$\begin{aligned}(I - P_{l-1}G_{l-1})\check{G}_l\check{P}_k &= (I - P_{l-1}G_{l-1})G_l \cdots G_{J-1}P_{J-1} \cdots P_l P_{l-1} \cdots P_k \\ &= (I - P_{l-1}G_{l-1})P_{l-1} \cdots P_k \\ &= P_{l-1}(I - G_{l-1}P_{l-1})P_{l-2} \cdots P_k \\ &= O_{n_l \times n_k} \ ,\end{aligned}$$

whereas, if $l < k$,

$$\begin{aligned}\check{G}_l\check{P}_k(I - P_{k-1}G_{k-1}) &= G_l \cdots G_{k-1}G_k \cdots G_{J-1}P_{J-1} \cdots P_k(I - P_{k-1}G_{k-1}) \\ &= G_l \cdots G_{k-1}(I - P_{k-1}G_{k-1}) \\ &= G_l \cdots G_{k-2}(I - G_{k-1}P_{k-1})G_{k-1} \\ &= O_{n_l \times n_k} \ . \qquad \blacksquare\end{aligned}$$

Appendix C

In this appendix we outline for even values of ν the proof of the following identity

$$\sup_{s_l\in(0,1)}\left(\frac{s_l}{1-(1-\frac{3\omega_{Jac}}{2}s_l)^\nu}+\frac{1-s_l}{1-(1-\frac{3\omega_{Jac}}{2}(1-s_l))^\nu}\right)=\frac{2}{3\nu\omega_{Jac}}+\frac{1}{1-(1-\frac{3\omega_{Jac}}{2})^\nu}\,,$$

with $\omega_{\mathrm{Jac}}\in(\,0\,,\,4/3\,)$. More precisely, we prove that

$$f(s_l)=\frac{s_l}{1-(1-\frac{3\omega_{Jac}}{2}s_l)^\nu}$$

is a convex function for $\omega_{\mathrm{Jac}}\in(\,0\,,\,4/3\,)$, and hence so is $f(s_l)+f(1-s_l)$, the prove being finished by the fact that any convex function takes it supremum at the boundary.

Now, note that

$$\tilde{f}(c)=\frac{3\omega_{Jac}}{2}f(c\,(2/3)\omega_{Jac}^{-1})=\left(1+c+\ldots+c^{\nu-1}\right)^{-1}=g(c)^{-1}$$

is convex for $c\in(-1,1)$ if and only if $f(s_l)$ is convex. However, $\tilde{f}(c)$ is convex if $\frac{d^2\tilde{f}}{dc^2}>0$ for $c\in(-1,1)$, that is, if $\frac{d^2g}{dc^2}\cdot g\;<\;2\cdot\left(\frac{dg}{dc}\right)^2$. On the other hand, one can check that

$$\frac{d^2g}{dc^2}\cdot g-2\left(\frac{dg}{dc}\right)^2\;=\;-\sum_{i=0}^{\nu/2-1}c^{2i-2}(c^2+i(\nu-2i)(c+1)^2)\,,$$

this last term being negative for $c\in(-1,1)$.

Chapter 3

When does two-grid optimality carry over to the V-cycle?

Summary

We investigate additional condition(s) that confirm that a V-cycle multigrid method is satisfactory (say, optimal) when it is based on a two-grid cycle with satisfactory (say, level-independent) convergence properties. The main tool is McCormick's bound on the convergence factor [SIAM J. Numer.Anal., 22(1985), pp.634-643], which we showed in previous work to be the best bound for V-cycle multigrid among those that are characterized by a constant that is the maximum (or minimum) over all levels of an expression involving only two consecutive levels; that is, that can be assessed considering only two levels at a time. We show that, given a satisfactorily converging two-grid method, McCormick's bound allows us to prove satisfactory convergence for the V-cycle if and only if the norm of a given projector is bounded at each level. Moreover, this projector norm is simple to estimate within the framework of Fourier analysis, making it easy to supplement a standard two-grid analysis with an assessment of the V-cycle potentialities. The theory is illustrated with a few examples that also show that the provided bounds may give a satisfactory sharp prediction of the actual multigrid convergence.

3.1 Introduction

We consider multigrid methods for the solution of symmetric positive definite (SPD) $n \times n$ linear systems:

$$A\mathbf{x} = \mathbf{b}. \tag{3.1}$$

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 such as Gauss-Seidel, and a *coarse-grid correction*, which corresponds to the solution of the residual equation on a coarser grid. A V–cycle multigrid method is obtained when the residual equation is solved approximately with one application of the two–grid scheme on that level, and so on, until the coarsest level, where an exact solve is performed. Other cycles may be defined, including the W–cycle based on two recursive applications of the two-grid scheme on each level, see, e.g., [61].

If there are only two levels, accurate bounds may be obtained either by means of Fourier analysis [60,61,68], or by using some appropriate algebraic tools [16,22,23,46,59]. This focus on two-grid schemes is motivated by the fact that, “if the two-grid method converges sufficiently well, then the multigrid method with W–cycle will have similar convergence properties” [61, p. 77] (see also [12, pp. 226–228] and [47]). This is not the case for the V–cycle since there are known examples where the two-grid method converges relatively well, whereas the multigrid method with V–cycle scales poorly with the number of levels [41]. Hence, V–cycle analysis has to be, at some point, essentially different from two-grid analysis.

In this chapter, we investigate additional condition(s) for obtaining an optimal V-cycle method from an optimal[1] two-grid method. Note that we do not base our work on a new analysis of the V-cycle. Several analyses are indeed available, which, however, have a common gap: the conditions for proving that the V-cycle converges nicely have not been compared with the two-grid convergence factor, and it is so far unclear how they are related. In fact, a number of results relate the V-cycle convergence to *sufficient* conditions for two-grid convergence; see, e.g., the two conditions (3.3) in [14], the first of which is sufficient for two-grid. Or, simply, consider V-cycle analysis particularized to the two-level case. Such sufficient conditions are, however, often stronger than needed for just two-level convergence, and, as far as we know, no comparison has been made with *necessary and sufficient* conditions or with two-grid convergence factor.

To analyze the V-cycle, one possibility consists of defining an appropriate subspace decomposition and then applying successive subspace correction (SSC) theory [50, 51, 25, 73, 75, 74]. Another possibility consists in checking so-called smoothing and approximation properties [10, 13, 26, 27, 37, 38, 53]. Regarding the latter approach, the best result for SPD matrices have been obtained by Hackbusch [27, Theorem 7.2.2] and McCormick [38]. In a Chapter 2, we show that these results are qualitatively equivalent, with McCormick's bound being always the sharpest. Note that, in both cases, the bound is characterized by a constant that is the minimum/maximum over all levels of an expression involving only two consecutive levels. This last property is important in the context of this study, since it seems at first sight not possible to compare with

[1] By “optimal”, for a two-grid method, we mean “having level-independent convergence properties”; that is, referring to a situation where the two-grid method is defined at different levels of a multigrid hierarchy, it is considered optimal if there is a level-independent bound on the convergence factor that is uniform with respect to problem size.

the two-grid convergence rate a global expression that would involve simultaneously all levels.

On the other hand, we also consider in Chapter 2 the classical formulation of the SSC theory (as stated in [73] or [75]), and discuss how to obtain a bound that could also be assessed considering only two levels at a time. It turns out that this requires the use of the so-called a-orthogonal decomposition, which corresponds to the choice most frequently made when applying the SSC theory to multigrid methods for H^2-regular problems. Then, the analysis in Chapter 2 shows that this approach is also qualitatively equivalent to the Hackbusch and McCormick ones, the latter remaining the sharpest.

Hence, regarding the goal pursued in this work, all exploitable results are superseded by (but qualitatively equivalent to) McCormick's bound, which is characterized by the constant δ; in this work, we relate this constant to the two-grid convergence factor. This reveals that a satisfactory (optimal) two-grid cycle on each level leads to a satisfactory estimate of δ if and only if a given norm of an exact coarse-grid correction (projection) operator remains bounded at each level. Moreover, it turns out that this norm is easy to assess within the framework of a Fourier analysis.

Eventually, we consider several examples, illustrating the sharpness of the bound based on two-grid convergence rates and the projector norm. It further turns out that both of these ingredients are independent and play an important role in the V-cycle convergence behavior.

The reminder of this chapter is organized as follows. In Section 3.2 we state the general setting of this study and gather the needed assumptions. The relation between the McCormick constant δ and the two-grid convergence factor is established in Section 3.3. Illustrative examples are discussed in Section 3.4.

3.2 General setting

We consider a multigrid method with $J+1$ levels ($J \geq 1$); index J refers to the finest level (on which the system (3.1) is to be solved), and index 0 to the coarsest level. The number of unknowns at level k , $0 \leq k \leq J$, is noted n_k (with thus $n_J = n$).

Our analysis applies to symmetric multigrid schemes based on the Galerkin principle for the SPD system (3.1); that is, restriction is the transpose of prolongation and the matrix A_k at level k , $k = J-1, \dots, 0$, is given by $A_k = P_k^T A_{k+1} P_k$, where P_k is the prolongation operator from level k to level $k+1$; we also assume that the smoother R_k is SPD and that the number of pre–smoothing steps ν ($\nu > 0$) is equal to the number of post–smoothing steps. The algorithm for V–cycle multigrid is then as follows.

Multigrid with V–cycle at level k: $\mathbf{x}_{n+1} = \mathrm{MG}(\mathbf{b}, A_k, \mathbf{x}_n, k)$

```
(1) Relax $\nu$ times with smoother $R_k$ : $\mathbf{x}_n \leftarrow Smooth(\mathbf{x}_n, A_k, R_k, \nu, \mathbf{b})$
(2) Compute residual: $\mathbf{r}_k = \mathbf{b} - A_k\mathbf{x}_n$
(3) Restrict residual: $\mathbf{r}_{k-1} = P_{k-1}^T\mathbf{r}_k$
(4) Coarse grid correction: if $k = 1$ , $\mathbf{e}_0 = A_0^{-1}\mathbf{r}_0$
                            else $\mathbf{e}_{k-1} = \mathrm{MG}(\mathbf{r}_{k-1}, A_{k-1}, 0, k-1)$
(5) Prolongate coarse-grid correction: $\mathbf{x}_n \leftarrow \mathbf{x}_n + P_{k-1}\mathbf{e}_{k-1}$
(6) Relax $\nu$ times with smoother $R_k$ : $\mathbf{x}_{n+1} \leftarrow Smooth(\mathbf{x}_n, A_k, R_k, \nu, \mathbf{b})$
```

When applying this algorithm, the error satisfies

$$A_k^{-1}\mathbf{b} - \mathbf{x}_{n+1} = E_{MG}^{(k)}\left(A_k^{-1}\mathbf{b} - \mathbf{x}_n\right) ,$$

where the iteration matrix $E_{MG}^{(k)}$ is recursively defined from

$$\begin{aligned} &E_{MG}^{(0)} = 0 \quad \text{and, for} \quad k = 1, 2, \ldots, J \ : \\ &E_{MG}^{(k)} = (I - R_k^{-1}A_k)^\nu \left(I - P_{k-1}(I - E_{MG}^{(k-1)})A_{k-1}^{-1}P_{k-1}^T A_k\right)(I - R_k^{-1}A_k)^\nu \end{aligned} \tag{3.2}$$

(see, e.g., [61, p. 48]). Our main objective is the analysis of the spectral radius of $E_{MG}^{(J)}$, which governs convergence on the finest level. Our analysis makes use of the following general assumptions.

General assumptions

- $n = n_J > n_{J-1} > ... > n_0$;
- P_k is an $n_{k+1} \times n_k$ matrix of rank n_k, $k = J-1, \ldots, 0$;
- $A_J = A$ and $A_k = P_k^T A_{k+1} P_k$, $k = J-1, \ldots, 0$;
- R_k is SPD and such that $\rho(I - R_k^{-1}A_k) < 1$, $k = J, \ldots, 1$.

In what follows, we make use of the two-grid cycle involving two consecutive levels k and $k-1$, which corresponds to the following iteration matrix:

$$E_{TG}^{(k)} = (I - R_k^{-1}A_k)^\nu \left(I - P_{k-1}A_{k-1}^{-1}P_{k-1}^T A_k\right)(I - R_k^{-1}A_k)^\nu, \ k = 1, \ldots, J \ . \tag{3.3}$$

Most of our results do not refer explicitly to the smoother R_k , but are stated with respect to the matrices $M_k^{(\nu)}$ defined from

$$I - {M_k^{(\nu)}}^{-1} A_k = (I - R_k^{-1}A_k)^\nu \ . \tag{3.4}$$

That is, $M_k^{(\nu)}$ is the smoother that provides in one step the same effect as ν steps with R_k . The results stated with respect to $M_k^{(\nu)}$ may then be seen as results stated for the case of one pre– and one post–smoothing step, which can be extended to the general case via the relations (3.4).

We close this subsection by introducing the projector π_{A_k}, which plays an important role throughout this chapter:

$$\pi_{A_k} \; = \; P_{k-1} A_{k-1}^{-1} P_{k-1}^T A_k \; . \tag{3.5}$$

Note that $I - \pi_{A_k}$ is the (exact) coarse-grid correction matrix at level k.

3.3 Theoretical Analysis

3.3.1 McCormick's bound

We recall in the following theorem the bound obtained in [38, Lemma 2.3, Theorem 3.4 and Section 5] (see also [37], or [53] for an alternative proof). The equivalence of (3.8) with the definition (3.7) is proved in in Theorem 2.6.

Note that convergence estimates based on regularity assumptions are also considered in [37]. These estimates are obtained when Theorem 3.1 below is applied to discretized PDEs. However, Theorem 3.1 on its own is a purely algebraic result that may by applied to any multigrid method satisfying the general assumptions in Section 3.2, without reference to a PDE context. Hence, there is no need for regularity assumptions to apply here, as may be further confirmed by the purely algebraic proof in [53].

Theorem 3.1. *Let* $E_{MG}^{(J)}$, $M_k^{(\nu)}$, *and* π_{A_k} , $k = 1, \ldots, J$, *be defined, respectively, by* (3.2), (3.4), *and* (3.5), *with* P_k , $k = 0, \ldots, J-1$, A_k , $k = 0, \ldots, J$, *and* R_k , $k = 1, \ldots, J$, *satisfying the general assumptions stated in Section 3.2.*

Then

$$\rho(E_{MG}^{(J)}) \; \leq \; 1 - \delta^{(\nu)} \quad , \tag{3.6}$$

where

$$\delta^{(\nu)} \; = \; \min_{1 \leq k \leq J} \; \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \; \frac{\|\mathbf{v}_k\|_{A_k}^2 - \|(I - {M_k^{(\nu)}}^{-1} A_k)\mathbf{v}_k\|_{A_k}^2}{\|(I - \pi_{A_k})\mathbf{v}_k\|_{A_k}^2} \tag{3.7}$$

$$\phantom{\delta^{(\nu)}} \; = \; \min_{1 \leq k \leq J} \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T A_k \mathbf{v}_k}{\mathbf{v}_k^T (I - \pi_{A_k})^T M_k^{(2\nu)} (I - \pi_{A_k}) \mathbf{v}_k} \tag{3.8}$$

3.3.2 Relationship to the two-grid convergence rate

We first recall, in the following lemma, a useful characterization of the two-grid rate obtained in [23, p. 480].

Lemma 3.1. *Let $E_{TG}^{(k)}$, $M_k^{(\nu)}$, and π_{A_k}, $k = 1, \ldots, J$, be defined, respectively, by* (3.3), (3.4), *and* (3.5), *with P_k, $k = 0, \ldots, J-1$, A_k, $k = 0, \ldots, J$, and R_k, $k = 1, \ldots, J$, satisfying the general assumptions stated in Section 3.2.*

Then

$$1 - \rho(E_{TG}^{(k)}) = \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T (I - \bar{\pi}_{A_k}) A_k^{1/2} \, {M_k^{(2\nu)}}^{-1} A_k^{1/2} (I - \bar{\pi}_{A_k}) \mathbf{v}_k}{\mathbf{v}_k^T (I - \bar{\pi}_{A_k}) \mathbf{v}_k}, \tag{3.9}$$

with $\bar{\pi}_{A_k} = A_k^{1/2} \pi_{A_k} A_k^{-1/2}$.

The next theorem contains our main result.

Theorem 3.2. *Let $E_{TG}^{(k)}$, $M_k^{(\nu)}$, and π_{A_k}, $k = 1, \ldots, J$, be defined, respectively, by* (3.3), (3.4), *and* (3.5), *with P_k, $k = 0, \ldots, J-1$, A_k, $k = 0, \ldots, J$, and R_k, $k = 1, \ldots, J$, satisfying the general assumptions stated in Section 3.2. Let $\delta^{(\nu)}$ be defined by* (3.7).

Then

$$\delta^{(\nu)} \geq \min_{1 \le k \le J} \frac{1 - \rho(E_{TG}^{(k)})}{\| I - \pi_{A_k} \|_{M_k^{(2\nu)}}^2} = \min_{1 \le k \le J} \frac{1 - \rho(E_{TG}^{(k)})}{\| \pi_{A_k} \|_{M_k^{(2\nu)}}^2}. \tag{3.10}$$

Moreover,

$$\delta^{(\nu)} \leq \min_{1 \le k \le J} \min \left(1 - \rho(E_{TG}^{(k)}), \; \frac{1}{\| \pi_{A_k} \|_{M_k^{(2\nu)}}^2} \right). \tag{3.11}$$

Proof.

Let ξ_k be defined by

$$\xi_k = \min_{\mathbf{v} \in \mathbb{R}^{n_k}} \frac{\mathbf{v}^T A_k \mathbf{v}}{\mathbf{v}^T (I - \pi_{A_k})^T M_k^{(2\nu)} (I - \pi_{A_k}) \mathbf{v}}.$$

From (3.8), there holds

$$\delta^{(\nu)} = \min_{1 \le k \le J} \xi_k. \tag{3.12}$$

On the other hand, Lemma 3.1 implies (since $A_k(I - \pi_{A_k}) = (I - \pi_{A_k})^T A_k$ and $(I - \pi_{A_k}) = (I - \pi_{A_k})^2$)

$$\begin{aligned}
1 - \rho(E_{TG}^{(k)}) &= \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T A_k^{1/2} (I - \pi_{A_k}) \, {M_k^{(2\nu)}}^{-1} A_k (I - \pi_{A_k}) A_k^{-1/2} \mathbf{v}_k}{\mathbf{v}_k^T A_k^{1/2} (I - \pi_{A_k}) A_k^{-1/2} \mathbf{v}_k} \\
&= \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T (I - \pi_{A_k}) \, {M_k^{(2\nu)}}^{-1} A_k (I - \pi_{A_k}) A_k^{-1} \mathbf{v}_k}{\mathbf{v}_k^T (I - \pi_{A_k})(I - \pi_{A_k}) A_k^{-1} \mathbf{v}_k} \\
&= \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T (I - \pi_{A_k}) \, {M_k^{(2\nu)}}^{-1} (I - \pi_{A_k})^T \mathbf{v}_k}{\mathbf{v}_k^T (I - \pi_{A_k}) A_k^{-1} (I - \pi_{A_k})^T \mathbf{v}_k}.
\end{aligned} \tag{3.13}$$

In what follows, we omit the subscripts k, as well as the superscript (k) and (2ν) in E_{TG} and M, respectively, when they are obvious from context. Using (3.13), one obtains

$$
\begin{aligned}
\xi^{-1} &= \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T(I-\pi_A)^T M (I-\pi_A)\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \\
&= \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T A^{-1/2}(I-\pi_A)^T M^{1/2} M^{1/2}(I-\pi_A)A^{-1/2}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \\
&= \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T M^{1/2}(I-\pi_A)A^{-1/2}A^{-1/2}(I-\pi_A)^T M^{1/2}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \\
&= \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T(I-\pi_A)A^{-1}(I-\pi_A)^T\mathbf{v}}{\mathbf{v}^T M^{-1}\mathbf{v}} \qquad (3.14)\\
&\le \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T(I-\pi_A)A^{-1}(I-\pi_A)^T\mathbf{v}}{\mathbf{v}^T(I-\pi_A)M^{-1}(I-\pi_A)^T\mathbf{v}} \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T(I-\pi_A)M^{-1}(I-\pi_A)^T\mathbf{v}}{\mathbf{v}^T M^{-1}\mathbf{v}} \\
&= \frac{1}{1-\rho(E_{TG})} \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T M^{1/2}(I-\pi_A)M^{-1/2}M^{-1/2}(I-\pi_A)^T M^{1/2}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \\
&= \frac{1}{1-\rho(E_{TG})} \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T M^{-1/2}(I-\pi_A)^T M^{1/2}M^{1/2}(I-\pi_A)M^{-1/2}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \\
&= \frac{1}{1-\rho(E_{TG})} \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T(I-\pi_A)^T M(I-\pi_A)\mathbf{v}}{\mathbf{v}^T M\mathbf{v}} \\
&= \frac{1}{1-\rho(E_{TG})} \left\|I-\pi_A\right\|_M^2 .
\end{aligned}
$$

The result (3.10) follows directly, using Kato's lemma (e.g., [65, Lemma 3.6]) which implies $\|I-\pi_A\|_M = \|\pi_A\|_M$, since $\pi_A \neq O, I$ by virtue of our general assumptions.

In addition, using (3.14) together with Lemma 3.1, one also has

$$
\begin{aligned}
\xi &= \min_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T M^{-1}\mathbf{v}}{\mathbf{v}^T(I-\pi_A)A^{-1}(I-\pi_A)^T\mathbf{v}} \\
&\le \min_{\mathbf{v}=(I-\pi_A)^T\mathbf{w},\ \mathbf{w}\in\mathbb{R}^n} \frac{\mathbf{v}^T M^{-1}\mathbf{v}}{\mathbf{v}^T(I-\pi_A)A^{-1}(I-\pi_A)^T\mathbf{v}} \\
&= 1-\rho(E_{TG}),
\end{aligned}
$$

which gives the first term in the right-hand side of (3.11).

On the other hand, since

$$
\mathbf{v}^T A^{1/2}\, {M^{(2\nu)}}^{-1} A^{1/2}\mathbf{v} \;=\; \mathbf{v}^T\mathbf{v} - \mathbf{v}^T(I - A^{1/2}\, {M^{(\nu)}}^{-1} A^{1/2})^2\mathbf{v}^T \;\le\; \mathbf{v}^T\mathbf{v}\ ,\quad \forall\mathbf{v}\in\mathbb{R}^n ,
$$

there holds

$$
\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T M\mathbf{v}\,,\quad \forall\mathbf{v}\in\mathbb{R}^n .
$$

Hence,

$$\begin{aligned}\xi &= \min_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T A\mathbf{v}}{\mathbf{v}^T(I-\pi_A)^T M(I-\pi_A)\mathbf{v}} \\ &\leq \min_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^T M\mathbf{v}}{\mathbf{v}^T(I-\pi_A)^T M(I-\pi_A)\mathbf{v}} \\ &= \frac{1}{\|I-\pi_A\|_M^2}\,,\end{aligned} \tag{3.15}$$

which, combined with Kato's lemma $\|I-\pi_A\|_M = \|\pi_A\|_M$, gives the second term in the right-hand side of (3.11). ■

Theorem 3.2 shows that McCormick's bound proves a satisfactory convergence rate for the V–cycle if and only if, at each level, the two-grid method converges fast enough and $\|\pi_{A_k}\|_{M_k^{(2\nu)}} = \|\, {M_k^{(2\nu)}}^{1/2}\, \pi_{A_k}\, {M_k^{(2\nu)}}^{-1/2}\, \|$ is nicely bounded. We can further show the following corollary.

Corollary 3.1. *Let the assumptions of Theorem 3.2 hold and let $E_{MG}^{(J)}$ be defined by (3.2).*

Then

$$\rho(E_{TG}^{(J)}) \leq \rho(E_{MG}^{(J)}) \leq 1-\delta^{(\nu)} \leq 1-\min_{1\leq k\leq J}\frac{1-\rho(E_{TG}^{(k)})}{\|\pi_{A_k}\|^2_{M_k^{(2\nu)}}}\,. \tag{3.16}$$

Proof.
The proof of $\rho(E_{TG}^{(k)}) \leq \rho(E_{MG}^{(k)})$ can be deduced from the relation (7.2.2a) in [27] combined with (7.2.4a) from the same reference, which proves that

$$A^{1/2}E_{MG}^{(k)}A^{-1/2} \leq A^{1/2}E_{TG}^{(k)}A^{-1/2}\,.$$

The other results follow from Theorems 3.1 and 3.2. ■

Note that the V-cycle convergence factor is bounded below by the two-grid convergence factor on the finest grid only. Indeed, $\max_{1\leq k\leq J}\rho(E_{TG}^{(k)})$ can be close to 1 even when $\rho(E_{MG}^{(J)})$ is not, for instance when the smoother alone is efficient enough on the finest level, so that poor two-grid ingredients on coarser levels will not significantly affect the convergence. In practice, however, one has often $\max_{1\leq k\leq J}\rho(E_{TG}^{(k)}) \approx \rho(E_{TG}^{(J)})$ (e.g., consider the discrete Poisson equation on many simple geometries with uniform meshes). Then (3.16) defines an interval, containing both $1-\delta^{(\nu)}$ and $\rho(E_{MG}^{(J)})$, that is narrow if and only if $\max_{1\leq k\leq J}\|\pi_{A_k}\|_{M_k^{(2\nu)}}$ is not much larger than 1.

3.3.3 Fourier analysis

Often, a multigrid method is assessed by estimating the two-grid convergence rate with Fourier analysis [60, 61, 68]. This means that one considers a model constant-coefficient

PDE for which the eigenvectors of the discrete matrix are explicitly known at all levels. Simple smoothers have the same set of eigenvectors and, hence, the matrices A_k and R_k are both diagonal whenever expressed in the corresponding basis (the Fourier basis). In more complicated situations, R_k may be only block-diagonal with small diagonal blocks; A_k may also have a block diagonal structure in case of coupled systems of PDEs. Note that $M_k^{(2\nu)}$, expressed in the Fourier basis, will then have the same block diagonal structure as A_k and R_k, and will be pointwise diagonal if A_k and R_k are pointwise diagonal.

Let

$$A_k = \begin{pmatrix} \Lambda_1^{(k)} & & & \\ & \Lambda_2^{(k)} & & \\ & & \ddots & \\ & & & \Lambda_{l_k}^{(k)} \end{pmatrix} \quad , \quad M_k^{(2\nu)} = \begin{pmatrix} \Sigma_1^{(k)} & & & \\ & \Sigma_2^{(k)} & & \\ & & \ddots & \\ & & & \Sigma_{l_k}^{(k)} \end{pmatrix}$$

be this (block) diagonal representation of A_k and $M_k^{(2\nu)}$, where the i^{th} block has size $m_i^{(k)} \times m_i^{(k)}$, $i = 1, ..., l_k$. Technically, Fourier analysis of a two-grid method at level k characterized by a given prolongation P_{k-1} is possible if there exists a basis of the coarse space (the coarse Fourier basis) such that the expression of P_{k-1} in both this basis and the (fine grid) Fourier basis has the structure

$$P_{k-1} = \begin{pmatrix} p_1^{(k-1)} & & & \\ & p_2^{(k-1)} & & \\ & & \ddots & \\ & & & p_{l_k}^{(k-1)} \end{pmatrix} ,$$

where $p_i^{(k-1)}$ are (possibly complex) rectangular matrices of size $m_i^{(k)} \times m_i^{(k-1)}$.

Here, we observe that, in this context, $M_k^{(2\nu)\;1/2}\,\pi_{A_k}\,M_k^{(2\nu)\;-1/2}$ is also block diagonal with diagonal blocks of the form

$$\Sigma_i^{(k)\;1/2}\,p_i^{(k-1)}\left(p_i^{(k-1)\;H}\,\Lambda_i^{(k)}p_i^{(k-1)}\right)^{-1}\,p_i^{(k-1)\;H}\,\Lambda_i^{(k)}\,\Sigma_i^{(k)\;-1/2} \quad . \tag{3.17}$$

Hence, $\|\pi_{A_k}\|^2_{M_k^{(2\nu)}}$ is the maximal norm of all these $m_i^{(k)} \times m_i^{(k)}$ blocks. Further, the matrices (3.17) are the product of rectangular matrices; taking the product of their norms gives an easy-to-assess upper bound:

$$\|\pi_{A_k}\|_{M_k^{(2\nu)}} \leq \max_i \|\,\Sigma_i^{(k)\;1/2}\,p_i^{(k-1)}\|\;\|\left(p_i^{(k-1)\;H}\,\Lambda_i^{(k)}p_i^{(k-1)}\right)^{-1}\,p_i^{(k-1)\;H}\,\Lambda_i^{(k)}\,\Sigma_i^{(k)\;-1/2}\,\|. \tag{3.18}$$

It is worth noting that the latter inequality becomes an equality when $m_i^{(k-1)} = 1$ for all i; that is, when the rectangular blocks $p_i^{(k-1)}$ are all simple vectors, as most often arises when analyzing scalar PDEs.

3.3.4 Finite element setting

Consider a finite element discretization of Poisson boundary value problem on a bounded domain. Such a domain is first approximated by an appropriate polygonal or polyhedral mesh, which is then refined several times. These refinements naturally induce a multigrid hierarchy (including inter-grid transfer operators P_k). It then can be shown (see [72, Theorem 4.2]) that $\|\pi_{A_k}\|$ are bounded on all levels if and only if the underlying problem possesses (full) elliptic regularity. Since $\|\cdot\|$ behaves similarly to $\|\cdot\|_{M_k^{(2\nu)}}$ for a number of smoothers, essentially the same conclusions hold with respect to $\|\pi_{A_k}\|_{M_k^{(2\nu)}}$.

With regards to the Theorem 3.2, these observations show that level independent two-grid convergence implies, in this context, a level-independent bound for V-cycle multigrid if and only if the problem has full elliptic regularity. Hence, it follows that McCormick's analysis cannot prove optimal bounds for the V-cycle if the problem does not possess full regularity. Considering the results in Chapter 2, the same conclusions hold for Hackbusch's analysis [27, Section 7.2], and the successive subspace correction theory with a-orthogonal decomposition [73, 75]. Thus, for the case when $\|\pi_{A_k}\|$ and $\|\pi_{A_k}\|_{M_k^{(2\nu)}}$ behave similarly with respect to the problem size, we show here that another type of analysis, as developed in, e.g., [50,51,25,73,75,74], is really needed to get uniform results for the V-cycle for problems with less than full regularity.

3.4 Examples

We consider three examples that represent three possible different practical situations. In the first, both $\rho(E_{TG}^{(k)})$ and $\|\pi_A\|^2_{M^{(2)}}$ are nicely bounded above. In the second example, $\rho(E_{TG}^{(k)})$ remains bounded away from one while $\|\pi_A\|^2_{M^{(2)}}$ increases rapidly with the problem size. The third example is the other way around: $\|\pi_A\|^2_{M^{(2)}}$ is nicely bounded while $\rho(E_{TG}^{(k)})$ is far from being optimal.

3.4.1 Standard multigrid with 2D Poisson

We consider the linear system resulting from the bilinear finite element discretization of the two-dimensional Poisson problem

$$\begin{aligned} -\Delta\, u = f \quad & \text{in } \ \Omega = (0,1)\times(0,1) \\ u = 0 \quad & \text{in } \ \partial\Omega \end{aligned}$$

on a uniform grid of mesh size $h = 1/N_J$ in both directions. The matrix corresponds then to the following nine point stencil:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} . \tag{3.19}$$

Up to some scaling factor, this is also the stencil obtained with 9-point finite difference discretization. We assume $N_J = 2^J N_0$ for some integer N_0 , allowing J steps of regular geometric coarsening. We consider the standard prolongation operator

$$P_k = \begin{pmatrix} J_k \\ I_{n_k} \end{pmatrix} ,$$

where J_k corresponds to the natural interpolation associated with bilinear finite element basis functions. The restriction P_k^T corresponds then to "full weighting", as defined in, e.g., [61] [2]. We consider damped Jacobi smoothing: $R_k = \omega_{\mathrm{Jac}}^{-1}\mathrm{diag}(A_k)$. Since the stencil is preserved on all levels, it is sufficient to consider only two successive grids; to alleviate notation, we therefore let $N = N_k$, $A = A_k$,$R = R_k$, $M = M_k^{(\nu)}$, $P = P_{k-1}$, $A_c = A_{k-1} = P^T AP$, and $\pi_A = \pi_{A_k} = PA_c^{-1}P^T A$.

We now use Fourier analysis to asses $\|\pi_A\|_{M^{(2\nu)}}$ via (3.18). The eigenvectors of A are, for $i, j = 1, \ldots , N - 1$, the functions

$$u_{i,j}^{(N)} = \sin(i\pi x) \sin(j\pi y)$$

evaluated at the grid points. The eigenvalue corresponding to $u_{i,j}^{(N)}$ is

$$\lambda_{i,j}^{(N)} = 4(3s_i + 3s_j - 4s_i s_j) , \tag{3.20}$$

where

$$s_i = \sin^2(\frac{i\pi}{2N}) \quad , \quad s_j = \sin^2(\frac{j\pi}{2N}) . \tag{3.21}$$

Hence, the eigenvalues of $I - R^{-1}A$ are in the interval $[1 - \omega_{Jac}\frac{3}{2}, 1)$. One has therefore $\rho(I - R^{-1}A) \leq 1$, as required by our general assumptions if $\omega_{Jac} \in (0, 4/3)$. The prolongation P satisfies (see, e.g., [61, p. 87])

$$P^T \left\{ \begin{array}{c} u_{i,j}^{(N)} \\ u_{N-i,N-j}^{(N)} \\ -u_{N-i,j}^{(N)} \\ -u_{i,N-j}^{(N)} \end{array} \right\} = 4 \left\{ \begin{array}{c} (1-s_i)(1-s_j) \\ s_i s_j \\ s_i(1-s_j) \\ (1-s_i)s_j \end{array} \right\} u_{i,j}^{(N/2)}$$

[2] up to some scaling factor; the scalings of the prolongation and restriction are unimportant when using coarse-grid matrices of the Galerkin type.

for $1 \le i,j \le N/2-1$, with $P^T u_{i,j}^{(N)} = 0$ for $i = N/2$ or $j = N/2$. Using

$$\begin{aligned}
\mathbf{p}_{i,j} &= 4\left(\begin{array}{cccc} (1-s_i)(1-s_j) & s_i s_j & s_i(1-s_j) & (1-s_i)s_j \end{array} \right)^T , \\
\Lambda_{i,j} &= \operatorname{diag}\left(\lambda_{i,j}^{(N)} , \lambda_{N-i,N-j}^{(N)} , \lambda_{N-i,j}^{(N)} , \lambda_{i,N-j}^{(N)}\right) , \\
\Sigma_{i,j}^{(\nu)} &= \operatorname{diag}\left(\left\{ \sigma^{(\nu)}(\lambda_{c,s}^{(N)}) \,\middle|\, \sigma^{(\nu)}(\lambda) = \frac{\lambda}{1-(1-\frac{\omega_{Jac}\lambda}{8})^\nu} \right\}_{(c,s)=(i,j),(N-i,N-j),(N-i,j),(i,N-j)} \right) ,
\end{aligned}$$

we can rewrite (3.18):

$$\|\pi_A\|_{M^{(2\nu)}}^2 = \max_{i,j=1,\ldots,N-1} g^{(\nu)}(s_i,s_j) ,$$

where

$$g^{(\nu)}(s_i,s_j) = \frac{\| \Sigma_{i,j}^{(2\nu)\,1/2} \mathbf{p}_{i,j} \|^2 \; \| \mathbf{p}_{i,j}{}^T \Lambda_{i,j} \, \Sigma_{i,j}^{(2\nu)\,-1/2} \|^2}{(\mathbf{p}_{i,j}{}^T \Lambda_{i,j} \; \mathbf{p}_{i,j})^2} . \tag{3.22}$$

One also has

$$\max_{i,j=1,\ldots,N-1} g^{(\nu)}(s_i,s_j) \;\le\; \sup_{(s_i,s_j)\in(0,1)\times(0,1)} g^{(\nu)}(s_i,s_j) .$$

For all ν, $g^{(\nu)}(s_i,s_m)$ exhibits the following symmetries: $g^{(\nu)}(s_i,s_j) = g^{(\nu)}(1-s_i,s_j) = g^{(\nu)}(s_i,1-s_j) = g^{(\nu)}(1-s_i,1-s_j)$. Further, numerical investigations reveal that the maximum on the considered domain is located at the boundary, i.e., corresponds to, e.g., $s_j = 0$ or, equivalently, $j = 0$ (such index values represent asymptotic behavior and do not correspond to any Fourier block). Because of the symmetries, it is sufficient to analyze this latter case. Next, since

$$\begin{aligned}
&g^{(\nu)}(s_i,0) \\
&\quad = \frac{\left((\mathbf{p}_{i,0})_1^2\sigma^{(2\nu)}(\lambda_{i,0}^{(N)}) + (\mathbf{p}_{i,0})_3^2\sigma^{(2\nu)}(\lambda_{N-i,0}^{(N)})\right)\left(\frac{(\mathbf{p}_{i,0})_1^2 \left(\lambda_{i,0}^{(N)}\right)^2}{\sigma^{(2\nu)}(\lambda_{i,0}^{(N)})} + \frac{(\mathbf{p}_{i,0})_3^2 \left(\lambda_{N-i,0}^{(N)}\right)^2}{\sigma^{(2\nu)}(\lambda_{N-i,0}^{(N)})} \right)}{\left((\mathbf{p}_{i,0})_1^2\lambda_{i,0}^{(N)} + (\mathbf{p}_{i,0})_3^2\lambda_{N-i,0}^{(N)} \right)^2} \\
&\quad = 1 + \frac{(\mathbf{p}_{i,0})_1^2(\mathbf{p}_{i,0})_3^2 \left(\frac{\sigma^{(2\nu)}(\lambda_{i,0}^{(N)})}{\sigma^{(2\nu)}(\lambda_{N-i,0}^{(N)})} \left(\lambda_{N-i,0}^{(N)}\right)^2 + \frac{\sigma^{(2\nu)}(\lambda_{N-i,0}^{(N)})}{\sigma^{(2\nu)}(\lambda_{i,0}^{(N)})} \left(\lambda_{i,0}^{(N)}\right)^2 - 2\lambda_{i,0}^{(N)}\lambda_{N-i,0}^{(N)} \right)}{\left((\mathbf{p}_{i,0})_1^2\lambda_{i,0} + (\mathbf{p}_{i,0})_3^2\lambda_{N-i,0} \right)^2} \\
&\quad = 1 + s_i(1-s_i)\left(\frac{1-\left(1-\frac{3}{2}\omega_{Jac}s_i\right)^{2\nu}}{1-\left(1-\frac{3}{2}\omega_{Jac}(1-s_i)\right)^{2\nu}} + \frac{1-\left(1-\frac{3}{2}\omega_{Jac}(1-s_i)\right)^{2\nu}}{1-\left(1-\frac{3}{2}\omega_{Jac}s_i\right)^{2\nu}} - 2 \right) ,
\end{aligned} \tag{3.23}$$

ω_{Jac}	$1-\frac{1}{\Vert\pi_A\Vert^2_{M^{(2)}}}$	$\rho(E_{TG}^{(J)})$	$\rho(E_{MG}^{(J)})$	$1-\delta^{(1)}$	$1-\frac{(1-\rho(E_{TG}^{(J)}))}{\Vert\pi_A\Vert^2_{M^{(2)}}}$
1/2	0.385	0.391	0.398	0.423	0.625
2/3	0.333	0.25	0.271	0.333	0.5
1	0.2	0.25	0.251	0.4	0.4

TABLE 3.1: The estimates of main convergence parameters for $\nu = 1$ and for different damping factors ω_{Jac}.

ω_{Jac}	$1-\frac{1}{\Vert\pi_A\Vert^2_{M^{(4)}}}$	$\rho(E_{TG}^{(J)})$	$\rho(E_{MG}^{(J)})$	$1-\delta^{(2)}$	$1-\frac{(1-\rho(E_{TG}^{(J)}))}{\Vert\pi_A\Vert^2_{M^{(4)}}}$
1/2	0.25	0.153	0.187	0.252	0.365
2/3	0.2	0.083	0.121	0.2	0.266
1	0.143	0.068	0.091	0.189	0.2

TABLE 3.2: The estimates of main convergence parameters for $\nu = 2$ and for different damping factors ω_{Jac}.

we obtain (see Appendix A for details)

$$\Vert\pi_A\Vert^2_{M^{(2\nu)}} \le \sup_{(s_i,s_j)\in(0,1)\times(0,1)} g^{(\nu)}(s_i,s_j) = \sup_{s_i\in(0,1)} g^{(\nu)}(s_i,0) \le \begin{cases} 2-\frac{3\omega_{Jac}}{4} & \text{if } \nu=1 \\ 1+\frac{1}{3\nu\omega_{Jac}} & \text{if } \nu>1\,. \end{cases}$$

Note that this bound is asymptotically sharp for $N \to \infty$ when $\nu = 1$, since $\lim_{s\to 0} g^{(1)}(s,0) = 2 - 3\omega_{Jac}/4$. In Tables 3.1 and 3.2, we use this bound and the asymptotically sharp estimate

$$\delta^{(\nu)\,-1} \le \frac{1}{3\nu\omega_{Jac}} + \frac{1}{1-(1-\frac{3\omega_{Jac}}{2})^{2\nu}}\,, \quad \forall\ \nu = 1,2,$$

obtained in Chapter 2 to illustrate inequalities (3.16), with two-grid and V-cycle multigrid convergence factors numerically assessed for $N_0 = 2$ and $J = 7$ (hence $N = 256$). Note that $\rho(E_{TG}^{(k)})$ increases with the mesh size, so that $\max_{1\le k\le J}\rho(E_{TG}^{(k)})$ corresponds to the value on the finest grid, which is close to the asymptotic one. Observe that the interval containing both $\rho(E_{MG}^{(J)})$ and $1-\delta^{(1)}$ is sharp enough. On the other hand, $1-\frac{1}{\Vert\pi_A\Vert^2_{M^{(2)}}}$ is also a lower bound on $1-\delta^{(1)}$ by (3.11), but in general not a lower bound on the effective convergence factor.

3.4.2 Aggregation-based multigrid for 1D Poisson

We consider $N \times N$ linear system associated to $A = A(\epsilon)$, where

$$A(\epsilon) = \begin{pmatrix} 2 & -1 & & \cdots & -1 \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 \\ -1 & & \cdots & -1 & 2 \end{pmatrix} + \epsilon\, N^{-1}\, I_N\,, \tag{3.24}$$

with $N = 2^J N_0$ and $\epsilon > 0$. We also assume piecewise constant prolongation of the form

$$P = \begin{pmatrix} 1 & 1 & & & & & \\ & & 1 & 1 & & & \\ & & & & \ddots & & \\ & & & & & 1 & 1 \end{pmatrix}^T .$$

Note that, with this prolongation, the successive coarse-grid matrices $A_k = A_k(\epsilon)$ are also given by (3.24) with N replaced by $N_k = 2^k N_0$, where we consider $N_0 \geq 2$. Hence, we can omit the subscript k (or $k-1$), let $A_c = A_{k-1} = P^T A P$, and set $\pi_A = \pi_{A_k} = P A_c^{-1} P^T A$.

Note that this is a 1D like problem which could be solved more efficiently using a tridiagonal solver. The analysis below can however be easily repeated in more dimensions, leading essentially to the same conclusions. We therefore continue with the 1D variant for the sake of simplicity.

The eigenvectors of $A(\epsilon)$ are, for $j = 0, \ldots, N - 1$, the functions

$$u_j^{(N)} = \frac{1}{\sqrt{N}} \exp(i\; j\pi x)$$

evaluated at the grid points, with $i = \sqrt{-1}$. The eigenvalue corresponding to $u_j^{(N)}$ is

$$\lambda_j^{(N)}(\epsilon) = 4\; \sin^2(j\pi N^{-1}) + \epsilon\; N^{-1}\,.$$

The prolongation P satisfies (see [41, p. 1087])

$$P^T \left\{ \begin{array}{c} u_j^{(N)} \\ u_{j+N/2}^{(N)} \end{array} \right\} = \sqrt{2}\; e^{i\; j\pi N^{-1}} \left\{ \begin{array}{c} \cos(j\pi N^{-1}) \\ i\; \sin(j\pi N^{-1}) \end{array} \right\} u_j^{(N/2)}\,.$$

We consider damped Jacobi smoother $R = 2\,\mathrm{diag}(A)$. Hence, the eigenvalues of $I - R^{-1}A$ are in the interval $[1 - \frac{\epsilon\; N^{-1}}{4+2\epsilon\; N^{-1}}\;,\; 1 - \frac{4+\epsilon\; N^{-1}}{4+2\epsilon\; N^{-1}}) = [\omega, 1-\omega)$ with $\omega = \frac{4+\epsilon\; N^{-1}}{4+2\epsilon\; N^{-1}} \in (0,1)$. One therefore has $\rho(I - R^{-1}A) \leq 1$, as required by our general assumptions.

Letting

$$
\begin{aligned}
&\mathbf{p}_j = \sqrt{2}\, e^{i\; j\pi N^{-1}} \begin{pmatrix} \cos(j\pi N^{-1}) & i\;\sin(j\pi N^{-1}) \end{pmatrix}^H ,\\
&\Lambda_j(\epsilon) = \operatorname{diag}\left(\lambda_j^{(N)}(\epsilon)\,,\, \lambda_{j+N/2}^{(N)}(\epsilon)\right) ,\\
&\Sigma_j^{(\nu)}(\epsilon) = \operatorname{diag}\left(\left\{ \sigma^{(\nu)}(\lambda_c^{(N)}(\epsilon)) \;\middle|\; \sigma^{(\nu)}(\lambda) = \frac{\lambda}{1-(1-\frac{\omega\lambda}{4+\epsilon\;N^{-1}}\;)^{\nu}} \right\}_{(c)=(j),(j+N/2)}\right) ,
\end{aligned}
$$

we can rewrite (3.18):

$$
\|\pi_A\|_{M^{(2\nu)}} = \max_{j=0,\dots,N/2-1} \frac{\|\;\Sigma_j^{(2\nu)}(\epsilon)^{1/2}\;\mathbf{p}_j\|\;\|\;\mathbf{p}_j{}^H\,\Lambda_j(\epsilon)\,\Sigma_j^{(2\nu)}(\epsilon)^{-1/2}\;\|}{\mathbf{p}_j{}^H\,\Lambda_j(\epsilon)\;\;\mathbf{p}_j} \,. \tag{3.25}
$$

First observe that $\sigma^{(2\nu)}(\lambda)$ is an increasing function of λ since $t(1-(1-t)^{2\nu})^{-1}$ is an increasing function of t on the interval $(0,1)$. Hence, since $\lambda_1^{(N)}(\epsilon) \le \lambda_{1+N/2}^{(N)}(\epsilon)$ for $N \ge 2N_0 \ge 4$, we have

$$
\begin{aligned}
\|\pi_A\|_{M^{(2\nu)}} &\ge \frac{\|\;\Sigma_1^{(2\nu)}(\epsilon)^{1/2}\;\mathbf{p}_1\|\;\|\;\mathbf{p}_1{}^H\,\Lambda_1(\epsilon)\,\Sigma_1^{(2\nu)}(\epsilon)^{-1/2}\;\|}{\mathbf{p}_1{}^H\,\Lambda_1(\epsilon)\;\;\mathbf{p}_1}\\
&= \frac{\sqrt{|(\mathbf{p}_1)_1|^2\frac{\sigma^{(2\nu)}(\lambda_1^{(N)}(\epsilon))}{\sigma^{(2\nu)}(\lambda_{1+N/2}^{(N)}(\epsilon))}+|(\mathbf{p}_1)_2|^2}\sqrt{|(\mathbf{p}_1)_1|^2\;\lambda_1^{(N)}(\epsilon)^2\;\frac{\sigma^{(2\nu)}(\lambda_{1+N/2}^{(N)}(\epsilon))}{\sigma^{(2\nu)}(\lambda_1^{(N)}(\epsilon))}+|(\mathbf{p}_1)_2|^2\;\lambda_{1+N/2}^{(N)}(\epsilon)^2}}{|(\mathbf{p}_1)_1|^2\;\lambda_1^{(N)}(\epsilon)+|(\mathbf{p}_1)_2|^2\;\lambda_{1+N/2}^{(N)}(\epsilon)}\\
&\ge \sqrt{\frac{\sigma^{(2\nu)}(\lambda_1^{(N)}(\epsilon))}{\sigma^{(2\nu)}(\lambda_{1+N/2}^{(N)}(\epsilon))}}\;\frac{\sqrt{|(\mathbf{p}_1)_1|^2+|(\mathbf{p}_1)_2|^2}\sqrt{|(\mathbf{p}_1)_1|^2\;\lambda_1^{(N)}(\epsilon)^2+|(\mathbf{p}_1)_2|^2\;\lambda_{1+N/2}^{(N)}(\epsilon)^2}}{|(\mathbf{p}_1)_1|^2\;\lambda_1^{(N)}(\epsilon)+|(\mathbf{p}_1)_2|^2\;\lambda_{1+N/2}^{(N)}(\epsilon)}\\
&= \sqrt{\frac{\sigma^{(2\nu)}(\lambda_1^{(N)}(\epsilon))}{\sigma^{(2\nu)}(\lambda_{1+N/2}^{(N)}(\epsilon))}}\;\frac{\sqrt{\cos^4(\pi N^{-1})\sin^2(\pi N^{-1})+\cos^2(\pi N^{-1})\sin^4(\pi N^{-1})+\mathcal{O}(\epsilon)}}{2\cos^2(\pi N^{-1})\sin^2(\pi N^{-1})+\mathcal{O}(\epsilon)}\,.
\end{aligned}
$$

Further, using again the monotonicity of $\sigma^{(2\nu)}$, there holds

$$
\frac{\sigma^{(2\nu)}(\lambda_1^{(N)}(\epsilon))}{\sigma^{(2\nu)}(\lambda_{1+N/2}^{(N)}(\epsilon))} \ge \frac{\lim_{\lambda\to 0}\sigma^{(2\nu)}(\lambda)}{\sigma^{(2\nu)}(4+\epsilon N^{-1})} = \frac{(4+\epsilon N^{-1})}{\nu\omega}\;\frac{1-(1-\omega)^{2\nu}}{(4+\epsilon N^{-1})} = \frac{1-(1-\omega)^{2\nu}}{\nu\omega}
$$

with $\omega \in (0,1)$. Hence, for $\epsilon \to 0$, we have

$$
\|\pi_A\|^2_{M^{(2\nu)}} \ge \frac{1-(1-\omega)^{2\nu}}{\nu\omega}\,\frac{1}{4\cos^2(\pi N^{-1})\sin^2(\pi N^{-1})} = \mathcal{O}(N^2)\,.
$$

Thus, $\|\pi_A\|^2_{M_k^{(2)}}$ increases with the problem size when ϵ is small enough, whereas, as shown in [41], the two-grid convergence factor remains bounded. Hence, we have an example of optimal two-grid method for which the V-cycle convergence estimate is poor. As seen in Table 3.3, it turns out that the actual convergence factor also deteriorates with the number of levels, showing that the analysis based on $\|\pi_A\|^2_{M^{(2)}}$ is qualitatively correct.

$J(N)$	1(8)	3(32)	5(128)	7(512)	9(2048)
$\|\pi_A\|^2_{M_J^{(2)}}$	1.471	13.58	208.0	3312	52575
$\rho(E_{TG}^{(J)})$	0.375	0.490	0.499	0.5	0.5
$\rho(E_{MG}^{(J)})$	0.375	0.800	0.947	0.986	0.997

TABLE 3.3: The values of main parameters for $\epsilon = 10^{-4}$ and for different problem sizes; the coarsest grid corresponds to $N_0 = 4$.

3.4.3 Positive off-diagonal entries

We consider the $(2N_J - 1) \times (2N_J - 1)$ matrix

$$A = \begin{pmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ & & & 1 & 2 \end{pmatrix},$$

with $N_k = N_0 \cdot 2^k$, corresponding to the one-dimensional stencil

$$\begin{bmatrix} 1 & 2 & 1 \end{bmatrix} . \tag{3.26}$$

We also consider the $(2N_k - 1) \times (N_k - 1)$ prolongation matrix

$$P_k = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & & & & & \\ & & 1 & 0 & 1 & & & \\ & & & & & \ddots & & \\ & & & & & 1 & 0 & 1 \end{pmatrix}^T \tag{3.27}$$

and the damped Jacobi smoother $R_k = \frac{1}{2}\,\mathrm{diag}(A_k)$ with one pre- and one post-smoothing step at each level. Note that the stencil (3.26) is preserved on all levels.

The values of $\|\pi_A\|^2_{M_J^{(2)}}$ and $\rho(E_{TG}^{(J)})$ on the finest grid, which are also the maximal values of these parameters over all grids, are given in the Table 3.4 together with the V-cycle convergence factor $\rho(E_{MG}^{(J)})$.

$J(N)$	1(4)	3(16)	5(64)	7(256)	9(1024)
$\|\pi_A\|^2_{M_J^{(2)}}$	1.235	1.479	1.498	1.5	1.5
$\rho(E_{TG}^{(J)})$	0.625	0.971	0.998	0.9999	0.99999
$\rho(E_{MG}^{(J)})$	0.625	0.971	0.998	0.9999	0.99999

TABLE 3.4: The values of main parameters for different problem sizes; the coarsest grid corresponds to $N_0 = 2$.

This example illustrates that $\|\pi_{A_k}\|^2_{M_k^{(2\nu)}}$ is a parameter essentially independent of $\rho(E_{TG}^{(k)})$, since it remains nicely bounded while both the two-grid and the V-cycle convergence factor deteriorate rapidly with the problem size.

3.5 Conclusion

We have presented a two-sided inequality (3.16) on the McCormick's estimate of V-cycle convergence factor (which is the best bound from the previous chapter). The inequality proves that the bound predicts an optimal V-cycle convergence *if and only if* the related two-grid scheme has level-independent convergence properties *and* the M_k-norm of a given projector π_{A_k} is bounded on on all levels. As a straightforward consequence, if the latter norm condition is checked, level-independent two-grid convergence implies optimal convergence properties for V-cycle multigrid. We have also shown on examples that both these conditions (level-independent convergence of the two-grid scheme and on the boundness of the π_{A_k} norm) are independent; that is, each of them can be satisfied whereas the other is not.

In the finite element context, when multigrid hierarchy is induced by successive mesh refinements, and considering well conditioned smoothers, we have shown that the bound of McCormick (as well as the other bounds in Chapter 2) provides an optimal estimate for V-cycle multigrid if and only if the underlying problem possesses (full) elliptic regularity.

Considering the Fourier analysis, we have observed that the norm of π_{A_k} can be easily assessed, allowing to supplement the two-grid estimate with an indication of V-cycle potentialities.

Appendix A

In this appendix, we outline the proof of the following inequality:

$$\sup_{s_i\in(0,1)} g^{(\nu)}(s_i,0) \leq \begin{cases} 2-\frac{3\omega_{Jac}}{4} & \text{if } \nu=1 \\ 1+\frac{1}{3\nu\omega_{Jac}} & \text{if } \nu>1\,, \end{cases} \tag{3.28}$$

with $g^{(\nu)}$ defined by (3.23) and $\omega_{\text{Jac}} \in [\,0\,,\,4/3\,)$.

Note that $g^{(\nu)}(s_i,0) = g^{(\nu)}(1-s_i,0)$ and it is sufficient to seek a supremum for $s_i \in (0,0.5)$. Next, exchanging $(3/2)\omega_{Jac}$ for α (hence, $\alpha \in [0,2)$), one has

$$\begin{aligned} g^{(\nu)}(s_i,0) =& 1+s_i(1-s_i)\left(\left[\frac{1-(1-\alpha s_i)^{2\nu}}{1-(1-\alpha(1-s_i))^{2\nu}}-1\right]+\left[\frac{1-(1-\alpha(1-s_i))^{2\nu}}{1-(1-\alpha s_i)^{2\nu}}-1\right]\right) \\ =& 1+s_i(1-s_i)\left[(1-\alpha s_i)^{2\nu}-(1-\alpha(1-s_i))^{2\nu}\right] \\ & \times\left(\frac{1}{1-(1-\alpha s_i)^{2\nu}}-\frac{1}{1-(1-\alpha(1-s_i))^{2\nu}}\right) \end{aligned}$$

$$\begin{aligned}
&\leq 1 + s_i(1-s_i)\left[(1-\alpha s_i)^{2\nu} - (1-\alpha(1-s_i))^{2\nu}\right]\left(\frac{1}{1-(1-\alpha s_i)^{2\nu}}\right) &(3.29)\\
&\leq 1 + s_i(1-\alpha s_i)^{2\nu}\left(\frac{1}{1-(1-\alpha s_i)^{2\nu}}\right)\\
&= 1 + \frac{(1-\alpha s_i)^{2\nu}}{\alpha\sum_{k=0}^{2\nu-1}(1-\alpha s_i)^k}\\
&\leq 1 + \frac{1}{2\nu\alpha}\,,
\end{aligned}$$

the last inequality coming from the fact that $\alpha s_i \in [0,1)$. This proves (3.28) for $\nu > 1$.

On the other hand, if $\nu = 1$, (3.29) further gives

$$\begin{aligned}
g^{(1)}(s_i,0) &\leq 1 + s_i(1-s_i)\left[(1-\alpha s_i)^2 - (1-\alpha(1-s_i))^2\right]\left(\frac{1}{1-(1-\alpha s_i)^2}\right)\\
&= 1 + s_i(1-s_i)\left[\alpha(2-\alpha)(1-2s_i)\right]\left(\frac{1}{\alpha s_i(2-\alpha s_i)}\right)\\
&= 1 + (2-\alpha)(1-2s_i)\left(\frac{1}{\alpha} - \frac{2-\alpha}{\alpha(2-\alpha s_i)}\right) &(3.30)\\
&\leq 1 + (2-\alpha)\left(\frac{1}{\alpha} - \frac{2-\alpha}{2\alpha}\right) &(3.31)\\
&= 2 - \frac{\alpha}{2}\,,
\end{aligned}$$

where the inequality (3.31) comes from the fact that the expression (3.30) is a decreasing function of s_i. This concludes the proof.

Chapter 4

Smoothing factor and actual multigrid convergence

Summary

We consider the Fourier analysis of multi-grid methods for symmetric positive definite and semi-positive definite linear systems arising from the discretizations of scalar PDEs. In this framework, the smoothing factor is frequently used to estimate the potential of a multigrid approach. In this chapter, the smoothing factor is related to the actual two-grid convergence rate and also to the V-cycle convergence estimate based on McCormick theory in [SIAM J. Numer.Anal., 22(1985), pp.634-643]. A two-sided bound is obtained that defines an interval containing both the two-grid and V-cycle convergence rate. This interval is narrow when an additional parameter is small enough, which is a simple function of quantities available in standard Fourier analysis.

From a qualitative viewpoint, it turns out that, besides the smoothing factor, the convergence mainly depends on the angle between the eigenvectors of the matrix associated with small eigenvalues and the range of the prolongation. Nice V-cycle convergence is guaranteed if the tangent of this angle has an upper bound proportional to the eigenvalue, whereas nice two-grid convergence requires the tangent to be bounded by an expression proportional only to the square root of the eigenvalue.

The presented results apply to rigorous Fourier analysis for regular discrete PDEs, and also to local Fourier analysis via the discussion of semi-definite systems as may arise from the discretization of PDEs with periodic boundary conditions.

4.1 Introduction

We consider Fourier analysis of multigrid methods for symmetric positive definite (SPD) or, more generally, symmetric semi-positive definite $n \times n$ linear systems

$$A\mathbf{x} = \mathbf{b}. \tag{4.1}$$

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 such as Gauss-Seidel, and a *coarse-grid correction*, which corresponds to the solution of the residual equation on a coarser grid. A multigrid method is obtained when the residual equation is solved approximately applying few iterations of the two–grid scheme on that level, and so on, until the coarsest level when an exact solve is performed. If the two–grid method is used recursively once on each level, the resulting algorithm is called V–cycle multigrid, whereas more involved cycling strategies (like W– or F–cycle) correspond to more iterations of two–grid method on given levels (see, e.g., [61, 27, 67]).

Fourier analysis [60, 61, 68] is a widely used tool that helps to design efficient multigrid approaches. It exploits the fact that the discretization of a constant-coefficient (elliptic) boundary value problem on simple domains often leads to a system (4.1) of which discrete Fourier modes are eigenvectors. If, in addition, other multigrid components also have a simple block structure in this Fourier basis, the analysis of a multigrid approach can be reduced to the analysis of diagonal blocks of small size, which can be done either analytically or numerically. The multigrid components designed for such simple cases are then adapted to more complex problems.

Fourier analysis is in practice limited to a few consecutive grids: generally two, rarely three [69]. Often Fourier analysis is further reduced to the computation of a simpler (one-grid) *smoothing factor*. When assessing this latter, the coarse-grid correction is assumed to annihilate the so-called *smooth* (or *low frequency*) error modes, while leaving *rough* (or *high frequency*) modes unchanged. Since this is the limit case of the desired behavior of a coarse-grid correction, the smoothing factor is often considered as an ideal two-grid convergence estimate. However, it is so far unclear which condition(s) are to be satisfied by the coarse-grid correction for having the actual two-grid convergence close to this ideal. Further, nice two-grid convergence does not necessarily imply optimal convergence of the multigrid method with V-cycle [41], hence the latter likely requires additional conditions.

In this chapter we investigate these questions for symmetric multigrid schemes of Galerkin type. The coarse-grid correction is essentially determined by the prolongation, and we establish a simple connection between the smoothing factor and the actual two-grid convergence via an additional parameter α that mainly depends on the coefficients of the prolongation in the Fourier basis. Regarding the V-cycle convergence rate, we use as main tool McCormick's bound [38] (see also [37, 53]) which is shown in Chapter 2 to be the best convergence estimate among those that can be assessed considering only two consecutive levels at a time. In a previous chapter, we show that optimal two-grid convergence implies optimal V-cycle convergence if the norm of the (two-grid) coarse-grid correction operator is bounded at each level. However, although it is sketched how to compute this norm within the framework of Fourier analysis, no simple criterion is given nor a connection is made with the smoothing factor. Here we prove a simple relation between McCormick's bound and the smoothing factor, using the same easy-to-compute parameter α that relates the smoothing factor with the two-grid convergence rate.

When the constant α and the smoothing factor are nicely bounded at each level, our analysis essentially proves that the two-grid and the V-cycle convergence factors are both in a narrow interval, which further goes towards zero as the number of smoothing steps is increased. On the other hand, from a more qualitative viewpoint, we deduce easy-to-check conditions to be satisfied by the prolongation for optimal two-grid or V-cycle convergence. Doing so, we give in some sense a more precise meaning to statements like "Interpolation must be able to approximate an eigenvector with error bound proportional to the size of the associated eigenvalue" [18, p. 1573], [21, p. 4]. We also highlight that the conditions for guaranteed optimal V-cycle convergence are in fact stronger that the conditions for optimal two-grid convergence.

In a number of practical cases, when Fourier analysis cannot be applied directly, it is still possible to replace boundary conditions, for instance, by the periodic ones, to make Fourier analysis work. Provided that some negligible extra smoothing is performed on the boundary, such modification has little influence on the convergence rate [17, 57]. These approaches are closely related to *local Fourier analysis*, which can often be viewed [68, Remark 5.3] [61, Section 3.4.4] as a (rigorous) Fourier analysis for problems with periodic boundary conditions. Since such boundary conditions often lead to semi-positive definite (singular) systems (4.1), our treatment should be valid for them as well. This is addressed in this work via the extension of McCormick's bound to the semi-positive definite case.

The reminder of this chapter is organized as follows. In Section 4.2 we state the general setting of this study for SPD systems and gather the needed assumptions. McCormick's bound is introduced in Section 4.3 and Fourier analysis for SPD problems is discussed in Section 4.4. The approach is extended to symmetric semi-positive definite systems in Section 4.5. Illustrative examples are discussed in Section 4.6.

4.2 General setting

We consider a multigrid method with $J+1$ levels; $J > 1$ corresponds to a truly multigrid method, whereas $J = 1$ leads to a mere two-grid scheme. Index J refers to the finest level (on which the system (4.1) is to be solved), and index 0 to the coarsest level. The number of unknowns at level k, $0 \leq k \leq J$, is noted n_k (with thus $n_J = n$).

Our analysis applies to symmetric multigrid schemes based on the Galerkin principle for the SPD system (4.1); that is, restriction is the transpose of prolongation and the matrix A_k at level k, $k = J-1, \dots, 0$, is given by $A_k = P_k^T A_{k+1} P_k$, where P_k is the prolongation operator from level k to level $k+1$; we also assume that the smoother R_k is SPD and that the number of pre–smoothing steps ν ($\nu > 0$) is equal to the number of post–smoothing steps.

The algorithm for V–cycle multigrid is defined as follows.

```
Multigrid with V–cycle at level $k$: $\mathbf{x}_{n+1} = \mathrm{MG}(\mathbf{b}, A_k, \mathbf{x}_n, k)$
 (1) Relax $\nu$ times with smoother $R_k$:
         repeat $\nu$ times    $\mathbf{x}_n \leftarrow \mathbf{x}_n + R_k^{-1}(\mathbf{b}_k - A_k\mathbf{x}_n)$
 (2) Compute residual: $\mathbf{r}_k = \mathbf{b} - A_k\mathbf{x}_n$
 (3) Restrict residual: $\mathbf{r}_{k-1} = P_{k-1}^T\mathbf{r}_k$
 (4) Coarse grid correction: if $k = 1$, $\mathbf{e}_0 = A_0^{-1}\mathbf{r}_0$
                             else $\mathbf{e}_{k-1} = \mathrm{MG}(\mathbf{r}_{k-1}, A_{k-1}, \mathbf{0}, k-1)$
 (5) Prolongate coarse-grid correction: $\mathbf{x}_n \leftarrow \mathbf{x}_n + P_{k-1}\mathbf{e}_{k-1}$
 (6) Relax $\nu$ times with smoother $R_k$:
         repeat $\nu$ times    $\mathbf{x}_{n+1} \leftarrow \mathbf{x}_n + R_k^{-1}(\mathbf{b}_k - A_k\mathbf{x}_n)$
```

Observe that for $k = 1$ this algorithm corresponds to a standard two-grid method with exact coarse-grid solve. Our analysis makes use of the following general assumptions.

General assumptions

- $n = n_J > n_{J-1} > \cdots > n_0$;
- P_k is an $n_{k+1} \times n_k$ matrix of rank n_k, $k = J-1, \ldots, 0$;
- $A_J = A$ and $A_k = P_k^T A_{k+1} P_k$, $k = J-1, \ldots, 0$;
- R_k is SPD and such that $\rho(I - R_k^{-1}A_k) < 1$, $k = J, \ldots, 1$.

Most of our results do not refer explicitly to the smoother R_k, but are stated with respect to the matrices $N_k^{(\nu)}$ defined from

$$N_k^{(\nu)} = \sum_{j=0}^{\nu-1} (I - R_k^{-1}A_k)^j R_k^{-1} , \tag{4.2}$$

which also satisfy

$$I - N_k^{(\nu)} A_k = (I - R_k^{-1}A_k)^\nu . \tag{4.3}$$

That is, $N_k^{(\nu)}$ is the relaxation operator that provides in 1 step the same effect as ν steps with R_k^{-1}. The results stated with respect to $N_k^{(\nu)}$ may then be seen as results stated for the case of 1 pre– and 1 post–smoothing step, which can be extended to the general case via the relation (4.3). If $N_k^{(\nu)}$ is nonsingular, it plays the same role as $M_k^{(\nu)\,-1}$ from the two previous chapters; however, in Section 4.5 potentially singular $N_k^{(\nu)}$ are considered.

When applying the V-cycle algorithm, the error satisfies

$$A_k^{-1}\mathbf{b} - \mathbf{x}_{n+1} = E_{MG}^{(k)}\left(A_k^{-1}\mathbf{b} - \mathbf{x}_n\right)$$

where the iteration matrix $E_{MG}^{(k)}$ is recursively defined from

$$\begin{aligned} &E_{MG}^{(0)} = O \quad \text{and, for} \quad k = 1, 2, \ldots, J \ : \\ &E_{MG}^{(k)} = (I - R_k^{-1}A_k)^{\nu}\left(I - P_{k-1}(I - E_{MG}^{(k-1)})A_{k-1}^{-1}P_{k-1}^T A_k\right)(I - R_k^{-1}A_k)^{\nu} \end{aligned} \tag{4.4}$$

(see, e.g., [61, p. 48]). Note that for $J = 1$ (4.4) reduces to the two-grid iteration matrix:

$$E_{TG}^{(J)} = (I - R_J^{-1}A_J)^{\nu}\left(I - P_{J-1}A_{J-1}^{-1}P_{J-1}^T A_J\right)(I - R_J^{-1}A_J)^{\nu}\,. \tag{4.5}$$

The convergence on finest level is governed by the spectral radius $\rho(E_{MG}^{(J)})$, or, in case of two-grid, $\rho(E_{TG}^{(J)})$. In this chapter, we want to discuss assessment of these spectral radii within the framework of a Fourier analysis, as may be developed for systems arising from the discretization of scalar PDEs. It means that the eigenvectors of A_k are explicitly known at each level and form the *Fourier basis.* We further assume that the smoother shares the same set of eigenvectors; i.e., is also diagonal when expressed in this Fourier basis.[1] According to (4.2), $N_k^{(\nu)}$ will be diagonal as well for all ν.

Technically, a Fourier analysis is then possible if, expressing P_{k-1} in both the coarse (level $k-1$) and fine (level k) Fourier basis, it has the form

$$P_{k-1} \;=\; \begin{pmatrix} \mathbf{p}_1^{(k-1)} & & & \\ & \mathbf{p}_2^{(k-1)} & & \\ & & \ddots & \\ & & & \mathbf{p}_{l_k-1}^{(k-1)} \\ & & & O \end{pmatrix}\,, \tag{4.6}$$

where $\mathbf{p}_j^{(k-1)}$ are (possibly complex) vectors of size $m_j^{(k)}$, $j = 1, ..., l_k - 1$. Note that this form induces a block partitioning of A_k, R_k and N_k when these matrices are expressed in Fourier basis. More precisely we write

$$A_k = \begin{pmatrix} \Lambda^{(k,1)} & & & \\ & \Lambda^{(k,2)} & & \\ & & \ddots & \\ & & & \Lambda^{(k,l_k)} \end{pmatrix}, \quad R_k = \begin{pmatrix} \Gamma^{(k,1)} & & & \\ & \Gamma^{(k,2)} & & \\ & & \ddots & \\ & & & \Gamma^{(k,l_k)} \end{pmatrix},$$

[1] Here we exclude cases for which the smoother is block diagonal as, e.g., when using red-black Gauss-Seidel relaxation for 5-pont discretizations of Poisson equation [61, Section 4.5].

$$N_k^{(2\nu)} = \begin{pmatrix} \Sigma^{(k,1)} & & & \\ & \Sigma^{(k,1)} & & \\ & & \ddots & \\ & & & \Sigma^{(k,l_k)} \end{pmatrix} , \tag{4.7}$$

where $\Lambda^{(k,j)} = \operatorname{diag}(\lambda_1^{(k,j)}, \dots, \lambda_{m_j^{(k)}}^{(k,j)})$, $\Gamma^{(k,j)} = \operatorname{diag}(\gamma_1^{(k,j)}, \dots, \gamma_{m_j^{(k)}}^{(k,j)})$, and $\Sigma^{(k,j)} = \operatorname{diag}(\sigma_1^{(k,j)}, \dots, \sigma_{m_j^{(k)}}^{(k,j)})$. Note that the block l_k corresponds to all eigenmodes of A_k that have no corresponding block in P_{k-1}; that is, all modes such that the associated eigenvector $\mathbf{v}_k$ satisfies $P_{k-1}^T \mathbf{v}_k = \mathbf{0}$. Whereas the separate treatment of the "non-prolongated" block is not compulsory (it can, for instance, be merged with one of the regular blocks), we adopt it here because such block (which can also be a group of "non-prolongated" blocks put together) often arises in practice.

Observe that the partitioning induced by (4.6) associates in a same block other than l_k the different Fourier modes that, on the coarse-grid, are mapped by P_{k-1}^T onto the same coarse Fourier mode. To develop our analysis, we don't need to enter the details about which modes are associated. It is important to note, however, that in usual setting of Fourier analysis (see, e.g., [61,68]), inside each set of associated modes (that is, inside each block other than l_k), there is a unique mode classified as "low frequency", all other modes being labelled as "high frequency". Moreover, the modes that belong to the block l_k are all classified as "high frequency". Then, the smoothing factor is the worst factor by which high frequency components are reduced per relaxation step; that is,

$$\tilde{\mu}^{(k)} = \max_{j=1,\dots,l_k} \max_{\substack{i=1,..,m_j^{(k)} \\ i \text{ is a "high frequency" mode}}} |1 - \gamma_i^{(k,j)\,-1} \lambda_i^{(k,j)}| \,.$$

In our study, we assume that the ordering inside each block is such that

$$|1 - \gamma_1^{(k,j)\,-1} \lambda_1^{(k,j)}| \;\geq\; |1 - \gamma_2^{(k,j)\,-1} \lambda_2^{(k,j)}| \;\geq\; \cdots \;\geq\; |1 - \gamma_{m_j^{(k)}}^{(k,j)\,-1} \lambda_{m_j^{(k)}}^{(k,j)}| \tag{4.8}$$

and report results with respect to

$$\begin{aligned} \mu^{(k)} &= \max_{j=1,\dots,l_k} \max_{\substack{i=2,..,m_j^{(k)} \text{ if } j<l_k \\ i=1,..,m_{l_k}^{(k)} \text{ if } j=l_k}} |1 - \gamma_i^{(k,j)\,-1} \lambda_i^{(k,j)}| \\ &= \max\left(\max_{j=1,\dots,l_k-1} |1 - \gamma_2^{(k,j)\,-1} \lambda_2^{(k,j)}| \,,\; |1 - \gamma_1^{(k,l_k)\,-1} \lambda_1^{(k,l_k)}| \right) . \end{aligned} \tag{4.9}$$

Clearly, $\mu^{(k)}$ coincides with the classical smoothing factor if, inside each block j other than l_k, the low frequency component is also the one less efficiently relaxed by the smoother. This corresponds to usual situations, but may be not true in whole generality.

Note, however, that one has $\mu^{(k)} \leq \widetilde{\mu}^{(k)}$ as soon as each block other than l_k contains at least one low frequency mode. In the following, we call $\mu^{(k)}$ the smoothing factor without checking further if $\mu^{(k)} = \widetilde{\mu}^{(k)}$.

Eventually, observe that (4.3) implies $1 - \sigma_i^{(k,j)} \lambda_i^{(k,j)} = (1 - \gamma_i^{(k,j)\,-1} \lambda_i^{(k,j)})^{2\nu}$ and hence (4.8) is equivalent to

$$\sigma_1^{(k,j)} \lambda_1^{(k,j)} \leq \sigma_2^{(k,j)} \lambda_2^{(k,j)} \leq \cdots \leq \sigma_{m_j^{(k)}}^{(k,j)} \lambda_{m_j^{(k)}}^{(k,j)} .$$

4.3 V–cycle analysis and McCormick's bound

We recall here the bound obtained in [38, Lemma 2.3, Theorem 3.4 and Section 5] (see also [37], or [53] for an alternative proof). The equivalence of definition (4.11) with (4.12) is proved in Theorem 2.6.

Theorem 4.1. *Let $E_{MG}^{(J)}$ and $N_k^{(\nu)}$, $k = 1, \ldots, J$, be defined respectively by (4.4) and (4.2) with A being symmetric positive definite and with P_k , $k = 0, \ldots, J-1$, A_k , $k = 0, \ldots, J$, and R_k , $k = 1, \ldots, J$, satisfying the general assumptions stated in Section 4.2.*

Then, letting $\pi_{A_k} = P_{k-1} A_{k-1}^{-1} P_{k-1}^T A_k$, there holds

$$\rho(E_{MG}^{(J)}) \leq 1 - \min_{1 \leq k \leq J} \delta_k^{(\nu)} \quad , \tag{4.10}$$

where

$$\delta_k^{(\nu)} = \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\|\mathbf{v}_k\|_{A_k}^2 - \|(I - N_k^{(\nu)} A_k)\mathbf{v}_k\|_{A_k}^2}{\|(I - \pi_{A_k})\mathbf{v}_k\|_{A_k}^2} \tag{4.11}$$

$$= \min_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T N_k^{(2\nu)} \mathbf{v}_k}{\mathbf{v}_k^T (A_k^{-1} - P_{k-1} A_{k-1}^{-1} P_{k-1}^T) \mathbf{v}_k} . \tag{4.12}$$

Moreover,

$$\delta_k^{(\nu)\,-1} \leq \frac{1}{\nu} \left(\delta_k^{(1)\,-1} + \nu - 1 \right) . \tag{4.13}$$

As already mentioned, it was shown in Chapter 2 that the McCormick's bound is the best bound for V-cycle multigrid among those characterized by a constant which is a maximum over all levels of an expression involving only two consecutive levels at a time. This latter feature is the key property that allows us, in the next section, to assess the bound in standard Fourier analysis setting, and relate it to the smoothing factor.

4.4 Rigorous Fourier analysis for SPD problems

Let $\widetilde{A}_k = N_k^{(2\nu)\,1/2} A_k\, N_k^{(2\nu)\,1/2}$ and $\widetilde{P}_{k-1} = N_k^{(2\nu)\,-1/2} P_{k-1}$, with corresponding block structure

$$\widetilde{A}_k = \begin{pmatrix} \widetilde{\Lambda}^{(k,1)} & & & \\ & \widetilde{\Lambda}^{(k,2)} & & \\ & & \ddots & \\ & & & \widetilde{\Lambda}^{(k,l_k)} \end{pmatrix} , \ \widetilde{P}_{k-1} = \begin{pmatrix} \widetilde{\mathbf{p}}_1^{(k-1)} & & & \\ & \widetilde{\mathbf{p}}_2^{(k-1)} & & \\ & & \ddots & \\ & & & \widetilde{\mathbf{p}}_{l_k-1}^{(k-1)} \\ & & & O \end{pmatrix} ,$$

where $\widetilde{\Lambda}^{(k,j)} = \mathrm{diag}(\widetilde{\lambda}_i^{(k,j)})$ with $\widetilde{\lambda}_i^{(k,j)} = \sigma_i^{(k,j)} \lambda_i^{(k,j)}$. Setting

$$\tilde{\pi}^{(k,j)} = \widetilde{\mathbf{p}}_j^{(k-1)} (\widetilde{\mathbf{p}}_j^{(k-1)\,H}\, \widetilde{\Lambda}^{(k,j)} \widetilde{\mathbf{p}}_j^{(k-1)})^{-1}\, \widetilde{\mathbf{p}}_j^{(k-1)\,H}\, \widetilde{\Lambda}^{(k,j)} ,$$

there holds

$$\begin{aligned} \rho(E_{TG}^{(k)}) &= \rho\left(((I - P_{k-1} A_{k-1}^{-1} P_{k-1}^T A_k)\, (I - N_k^{(\nu)}\, A_k)^2 \right) \\ &= \rho\left((I - P_{k-1} A_{k-1}^{-1} P_{k-1}^T A_k)\, (I - N_k^{(2\nu)}\, A_k) \right) \\ &= \rho\left((I - \widetilde{P}_{k-1} \widetilde{A}_{k-1}^{-1} \widetilde{P}_{k-1}^T \widetilde{A}_k)(I - \widetilde{A}_k) \right) \\ &= \max\left(\max_{j=1,\ldots,l_k-1} \rho\left((I - \tilde{\pi}^{(k,j)})(I - \widetilde{\Lambda}^{(k,j)}) \right) \ , \ \rho\left(I - \widetilde{\Lambda}^{(k,l_k)} \right) \right) , \end{aligned}$$

and

$$\begin{aligned} \delta_k^{(\nu)\,-1} &= \max_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T (A_k^{-1} - P_{k-1} A_{k-1}^{-1} P_{k-1}^T) \mathbf{v}_k}{\mathbf{v}_k^T\, N_k^{(2\nu)\,-1}\, \mathbf{v}_k} \\ &= \max_{\mathbf{v}_k \in \mathbb{R}^{n_k}} \frac{\mathbf{v}_k^T (I - \widetilde{P}_{k-1} \widetilde{A}_{k-1}^{-1} \widetilde{P}_{k-1}^T \widetilde{A}_k) \widetilde{A}_k^{-1} \mathbf{v}_k}{\mathbf{v}_k^T \mathbf{v}_k} \\ &= \ \max\left(\max_{j=1,\ldots,l_k-1} \rho\left((I - \tilde{\pi}^{(k,j)})\, \widetilde{\Lambda}^{(k,j)\,-1} \right) \ , \ \rho\left(\widetilde{\Lambda}^{(k,l_k)\,-1} \right) \right) . \end{aligned}$$

Now, for each individual block other than l_k, the quantities one has to take the maximum of may be assessed by applying the following theorem with $\widetilde{\Lambda} = \widetilde{\Lambda}^{(k,j)}$ and $\widetilde{\mathbf{p}} = \widetilde{\mathbf{p}}_j^{(k-1)}$ (this vector is not equal to zero since the block l_k is not considered). Observe that the assumption $0 < \widetilde{\lambda}_i \le 1$ is then not restrictive since $I - \widetilde{A}_k$ and $(I - N_k^{(\nu)} A_k)^2$ have the same spectra, and hence the eigenvalues of $\widetilde{\Lambda}^{(k,j)}$, being a subset of the eigenvalues of $\widetilde{A}_k$, belong to $(0, 1]$ by virtue of our general assumptions.

Theorem 4.2. *Let* $\widetilde{\Lambda} = \mathrm{diag}(\widetilde{\lambda}_i)$ *be a* $m \times m$ *real matrix with* $0 < \widetilde{\lambda}_1 \leq \widetilde{\lambda}_2 \leq \cdots \leq \widetilde{\lambda}_m \leq 1$ *, and let* $\widetilde{\mathbf{p}} = (\widetilde{p}_1 \quad \dots \quad \widetilde{p}_m)^T$ *be a nonzero complex vector. Set*

$$\widetilde{\pi} = \widetilde{\mathbf{p}}\ (\widetilde{\mathbf{p}}^H \widetilde{\Lambda} \widetilde{\mathbf{p}})^{-1}\ \widetilde{\mathbf{p}}^H \widetilde{\Lambda}\,,$$
$$\rho_{TG} = \rho\Big(\ (I_m - \widetilde{\pi})\ \Big(I_m - \widetilde{\Lambda}\Big)\ \Big)\ ,$$

and

$$\delta^{-1} = \rho\Big(\ (I_m - \widetilde{\pi})\ \widetilde{\Lambda}^{-1}\Big)\ .$$

Letting

$$\widetilde{\alpha} = \sum_{i=2}^{m} \frac{\widetilde{\lambda}_i^2 |\widetilde{p}_i|^2}{\widetilde{\lambda}_1^2 \|\widetilde{\mathbf{p}}\|^2}\,, \tag{4.14}$$

then

$$\frac{\widetilde{\lambda}_2}{1 + \widetilde{\alpha}(1 - \widetilde{\lambda}_1/\widetilde{\lambda}_2)} \ \leq\ \delta\ \leq\ \left(\frac{4}{\widetilde{\alpha}}\right)^{1/3}\ . \tag{4.15}$$

Moreover, if $|\widetilde{p}_1| > 0$*, letting*

$$\widetilde{\beta} = \sum_{i=2}^{m} \frac{\widetilde{\lambda}_i^2 |\widetilde{p}_i|^2}{\widetilde{\lambda}_1^2 |\widetilde{p}_1|^2}\,, \tag{4.16}$$

then

$$\widetilde{\lambda}_1 + \frac{\widetilde{\lambda}_2 - \widetilde{\lambda}_1}{1 + \widetilde{\beta}} \ \leq\ \delta\ \leq\ \widetilde{\lambda}_1 + \frac{\widetilde{\lambda}_m - \widetilde{\lambda}_1}{1 + \widetilde{\beta}} \tag{4.17}$$

and

$$\widetilde{\lambda}_1 + \frac{\widetilde{\lambda}_2 - \widetilde{\lambda}_1}{1 + \widetilde{\lambda}_2^{-1} \widetilde{\lambda}_1 \widetilde{\beta}} \ \leq\ 1 - \rho_{TG}\ \leq\ \min\left(\widetilde{\lambda}_1 + \frac{\widetilde{\lambda}_m - \widetilde{\lambda}_1}{1 + \widetilde{\lambda}_m^{-1} \widetilde{\lambda}_1 \widetilde{\beta}}\ ,\ \widetilde{\lambda}_2\right)\ , \tag{4.18}$$

whereas, if $|\widetilde{p}_1| = 0$*,*

$$\delta\ =\ 1 - \rho_{TG}\ =\ \widetilde{\lambda}_1\,. \tag{4.19}$$

Proof. Set $\widetilde{\lambda}_c = \widetilde{\mathbf{p}}^H \widetilde{\Lambda} \widetilde{\mathbf{p}} = \sum_{i=1}^m \widetilde{\lambda}_i |\widetilde{p}_i|^2$. First, observe that, according to Lemma 2.2 in [41],

$$\widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} \widetilde{\lambda}_m^{-1} + (1 - \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1}) \widetilde{\lambda}_1^{-1} \leq \delta^{-1} \leq \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} \widetilde{\lambda}_2^{-1} + (1 - \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1}) \widetilde{\lambda}_1^{-1}\,, \tag{4.20}$$

and, similarly,

$$\widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1}\ \eta_m + (1 - \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1}) \eta_1 \leq \rho_{TG} \leq \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1}\ \eta_2 + (1 - \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1}) \eta_1\,, \tag{4.21}$$

where $\eta_i = (1 - \widetilde{\lambda}_i)$. Equality (4.19) then readily follows. Moreover, (4.20) and (4.21) can be further rewritten as

$$\widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} \left(\widetilde{\lambda}_m^{-1} - \widetilde{\lambda}_1^{-1} \right) + \widetilde{\lambda}_1^{-1} \le \delta^{-1} \le \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} \left(\widetilde{\lambda}_2^{-1} - \widetilde{\lambda}_1^{-1} \right) + \widetilde{\lambda}_1^{-1} \,, \tag{4.22}$$

$$\widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} \left(\eta_m - \eta_1 \right) + \eta_1 \le \rho_{TG} \le \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} \left(\eta_2 - \eta_1 \right) + \eta_1 \,. \tag{4.23}$$

We now prove the inequalities (4.17) and (4.18). Note that

$$\widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} = \frac{\widetilde{\lambda}_1 |\widetilde{p}_1|^2}{\sum_{i=1}^m \widetilde{\lambda}_i |\widetilde{p}_i|^2} = \left(1 + \widetilde{\lambda}_1 \frac{\sum_{i=2}^m \widetilde{\lambda}_i |\widetilde{p}_i|^2}{\widetilde{\lambda}_1^2 |\widetilde{p}_1|^2} \right)^{-1}$$

implies

$$(1 + \xi_2)^{-1} \le \widetilde{\lambda}_1 |\widetilde{p}_1|^2 \widetilde{\lambda}_c^{-1} \le (1 + \xi_m)^{-1} \,, \tag{4.24}$$

where $\xi_i = \widetilde{\widetilde{\beta}\widetilde{\lambda}_1}/\widetilde{\lambda}_i$. Using these last inequalities in (4.22) and (4.23) one obtains (since $\widetilde{\lambda}_1^{-1} \ge \cdots \ge \widetilde{\lambda}_m^{-1}$ and $\eta_1 \ge \cdots \ge \eta_m$)

$$\begin{aligned} \frac{1}{\xi_m + 1} \left(\widetilde{\lambda}_m^{-1} - \widetilde{\lambda}_1^{-1} \right) + \widetilde{\lambda}_1^{-1} \le \ \delta^{-1} \ &\le \frac{1}{\xi_2 + 1} \left(\widetilde{\lambda}_2^{-1} - \widetilde{\lambda}_1^{-1} \right) + \widetilde{\lambda}_1^{-1} \,, \\ \frac{1}{\xi_m + 1} \left(\eta_m - \eta_1 \right) + \eta_1 \le \ \rho_{TG} \ &\le \frac{1}{\xi_2 + 1} \left(\eta_2 - \eta_1 \right) + \eta_1 \,. \end{aligned}$$

Hence, using $\xi_i = \widetilde{\beta}\ \widetilde{\lambda}_1/\widetilde{\lambda}_i$ and $\eta_i = 1 - \widetilde{\lambda}_i$, $i = 2, m$, we have

$$\begin{aligned} \frac{1 + \widetilde{\beta}}{\widetilde{\lambda}_m + \widetilde{\lambda}_1 \widetilde{\beta}} \le \ \delta^{-1} \ &\le \ \frac{1 + \widetilde{\beta}}{\widetilde{\lambda}_2 + \widetilde{\lambda}_1 \widetilde{\beta}} \,, \\ 1 - \frac{\widetilde{\lambda}_m^2 + \widetilde{\lambda}_1^2 \widetilde{\beta}}{\widetilde{\lambda}_m + \widetilde{\lambda}_1 \widetilde{\beta}} \le \ \rho_{TG} \ &\le 1 - \frac{\widetilde{\lambda}_2^2 + \widetilde{\lambda}_1^2 \widetilde{\beta}}{\widetilde{\lambda}_2 + \widetilde{\lambda}_1 \widetilde{\beta}} \,. \end{aligned}$$

The inequalities (4.17) and (4.18) (except the second term in the minimum) readily follow. To conclude the proof of (4.18), let $\mathbf{e}_k$ be the k-th unit vector and let

$$\mathbf{v} = \begin{cases} \mathbf{e}_2 & \text{if } \widetilde{\mathbf{p}}^H \mathbf{e}_2 = 0 \\ \mathbf{e}_1 - \mathbf{e}_2 \left(\frac{\widetilde{\lambda}_1^{\ 1/2}\, \widetilde{\mathbf{p}}^H \mathbf{e}_1}{\widetilde{\lambda}_2^{\ 1/2}\, \widetilde{\mathbf{p}}^H \mathbf{e}_2} \right) & \text{otherwise .} \end{cases}$$

Note that $\widetilde{\pi}\ \widetilde{\Lambda}^{-1/2}\, \mathbf{v} = \mathbf{0}$ and $\widetilde{\Lambda}\ \widetilde{\pi} = \widetilde{\pi}^H\ \widetilde{\Lambda}$. Hence,

$$\begin{aligned} \rho_{TG} &= \rho \Big(\ (I_m - \widetilde{\pi})^2 \ \left(I_m - \widetilde{\Lambda} \right) \Big) \\ &= \rho \Big(\ (I_m - \widetilde{\pi}) \ \left(I_m - \widetilde{\Lambda} \right) \ (I_m - \widetilde{\pi}) \ \Big) \\ &= \rho \Big(\ \widetilde{\Lambda}^{1/2} \ (I_m - \widetilde{\pi}) \ \left(I_m - \widetilde{\Lambda} \right) \ (I_m - \widetilde{\pi}) \ \widetilde{\Lambda}^{-1/2} \ \Big) \\ &= \rho \Big(\ \widetilde{\Lambda}^{-1/2} \ (I_m - \widetilde{\pi})^H \ \widetilde{\Lambda}^{1/2} \ \left(I_m - \widetilde{\Lambda} \right) \ \widetilde{\Lambda}^{1/2} \ (I_m - \widetilde{\pi}) \ \widetilde{\Lambda}^{-1/2} \ \Big) \end{aligned}$$

$$
\begin{aligned}
&\geq \frac{\mathbf{v}^H\widetilde{\Lambda}^{-1/2}\ (I_m-\widetilde{\pi})^H\ \widetilde{\Lambda}^{1/2}\ \left(I_m-\widetilde{\Lambda}\right)\ \widetilde{\Lambda}^{1/2}\ (I_m-\widetilde{\pi})\ \widetilde{\Lambda}^{-1/2}\mathbf{v}}{\mathbf{v}^H\mathbf{v}} \\
&= \frac{\mathbf{v}^H\left(I_m-\widetilde{\Lambda}\right)\mathbf{v}}{\mathbf{v}^H\mathbf{v}} \\
&\geq 1-\widetilde{\lambda}_2\,.
\end{aligned}
$$

We next prove the left inequality (4.15). First observe that

$$\widetilde{\lambda}_1|\widetilde{p}_1|^2\widetilde{\lambda}_c^{-1} = \frac{\widetilde{\lambda}_1|\widetilde{p}_1|^2}{\sum_{i=1}^m\widetilde{\lambda}_i|\widetilde{p}_i|^2} = 1-\frac{\sum_{i=2}^m\widetilde{\lambda}_i|\widetilde{p}_i|^2}{\sum_{i=1}^m\widetilde{\lambda}_i|\widetilde{p}_i|^2} \geq 1-\frac{\sum_{i=2}^m\widetilde{\lambda}_i|\widetilde{p}_i|^2}{\widetilde{\lambda}_1\sum_{i=1}^m|\widetilde{p}_i|^2} \geq 1-\frac{\widetilde{\lambda}_1\widetilde{\alpha}}{\widetilde{\lambda}_2}\,,$$

and hence, with (4.20), there holds

$$\delta^{-1} \leq \left(1-\frac{\widetilde{\lambda}_1\widetilde{\alpha}}{\widetilde{\lambda}_2}\right)\widetilde{\lambda}_2^{-1}+\frac{\widetilde{\lambda}_1\widetilde{\alpha}}{\widetilde{\lambda}_2}\widetilde{\lambda}_1^{-1} = \widetilde{\lambda}_2^{-1}\left(1+\widetilde{\alpha}(1-\frac{\widetilde{\lambda}_1}{\widetilde{\lambda}_2})\right)\,.$$

It remains to prove the right inequality (4.15). Note that, according to Theorem 3.2,

$$\delta \leq \min\left(\,1-\rho_{TG}\ ,\ \|\widetilde{\pi}\|^{-2}\,\right)\,.$$

Hence, provided that

$$\widetilde{\alpha} \leq \frac{4}{(1-\rho_{TG})^2}\|\widetilde{\pi}\|^2 \tag{4.25}$$

holds (we prove it below), we have

$$\delta \leq \min\left(\ 1-\rho_{TG}\ ,\ \frac{1}{\widetilde{\alpha}}\frac{4}{(1-\rho_{TG})^2}\ \right) \leq \max_{x>0}\ \min\left(\ x\ ,\ \frac{1}{\widetilde{\alpha}}\frac{4}{x^2}\ \right) \leq \left(\frac{4}{\widetilde{\alpha}}\right)^{1/3}\,.$$

We are thus left with the proof of (4.25), for which we use

$$\|\widetilde{\pi}\|^2 \ = \ \rho\left(\widetilde{\mathbf{p}}\ \widetilde{\lambda}_c^{-1}\ \widetilde{\mathbf{p}}^H\widetilde{\Lambda}^2\widetilde{\mathbf{p}}\ \widetilde{\lambda}_c^{-1}\ \widetilde{\mathbf{p}}^H\right) \ = \ \frac{\|\widetilde{\mathbf{p}}\|^2\ \widetilde{\mathbf{p}}^H\widetilde{\Lambda}^2\widetilde{\mathbf{p}}}{\widetilde{\lambda}_c^2} \ = \ \|\widetilde{\mathbf{p}}\|^2\ \frac{\sum_{i=1}^m\widetilde{\lambda}_i^2|\widetilde{p}_i|^2}{\widetilde{\lambda}_c^2}\,. \tag{4.26}$$

According to (4.21), we have

$$\rho_{TG} \geq 1-\widetilde{\lambda}_1-\widetilde{\lambda}_1|\widetilde{p}_1|^2\widetilde{\lambda}_c^{-1}\ (\widetilde{\lambda}_m-\widetilde{\lambda}_1)\,.$$

Hence, considering first the case where $\widetilde{\lambda}_1 \leq \frac{1-\rho_{TG}}{2}$ and $\widetilde{\lambda}_1 \neq \widetilde{\lambda}_m$, there holds

$$\widetilde{\lambda}_1|\widetilde{p}_1|^2\widetilde{\lambda}_c^{-1} = \frac{|\widetilde{p}_1|^2\widetilde{\lambda}_1}{\sum_{i=1}^m\widetilde{\lambda}_i|\widetilde{p}_i|^2} \geq \frac{1-\widetilde{\lambda}_1-\rho_{TG}}{\widetilde{\lambda}_m-\widetilde{\lambda}_1} \geq \frac{1-\rho_{TG}}{2(\widetilde{\lambda}_m-\widetilde{\lambda}_1)} \geq \frac{1-\rho_{TG}}{2}\,, \tag{4.27}$$

the last inequality following from $0 \leq \widetilde{\lambda}_1 < \widetilde{\lambda}_m \leq 1$. Note that (4.27) implies $|\widetilde{p}_1|^2 > 0$. The right inequality (4.25) follows then from

$$\widetilde{\alpha} = \sum_{i=2}^{m} \frac{\widetilde{\lambda}_i^2 |\widetilde{p}_i|^2}{\widetilde{\lambda}_1^2 \|\widetilde{\mathbf{p}}\|^2} \leq \|\widetilde{\mathbf{p}}\|^2 \frac{\sum_{i=1}^{m} \widetilde{\lambda}_i^2 |\widetilde{p}_i|^2}{(\widetilde{\lambda}_1 |\widetilde{p}_1|^2)^2} \leq \frac{4}{(1-\rho_{TG})^2} \|\widetilde{\mathbf{p}}\|^2 \frac{\sum_{i=1}^{m} \widetilde{\lambda}_i^2 |\widetilde{p}_i|^2}{\widetilde{\lambda}_c^2} \,,$$

together with (4.26). If $\widetilde{\lambda}_1 = \widetilde{\lambda}_m$, we have

$$\widetilde{\alpha} = \sum_{i=2}^{m} \frac{|\widetilde{p}_i|^2}{\|\widetilde{\mathbf{p}}\|^2} \leq 1 \leq \frac{4}{(1-\rho_{TG})^2} \|\widetilde{\pi}\|^2 \,,$$

the last inequality coming from the fact that $\widetilde{\pi}$ is a projector, and hence $\|\widetilde{\pi}\| \geq 1$. On the other hand, when $\widetilde{\lambda}_1 \geq \frac{1-\rho_{TG}}{2}$ one has (since $\widetilde{\lambda}_i \leq 1$)

$$\widetilde{\alpha} = \sum_{i=2}^{m} \frac{\widetilde{\lambda}_i^2 |\widetilde{p}_i|^2}{\widetilde{\lambda}_1^2 \|\widetilde{\mathbf{p}}\|^2} < \frac{4}{(1-\rho_{TG})^2} \sum_{i=2}^{m} \frac{\widetilde{\lambda}_i^2 |\widetilde{p}_i|^2}{\|\widetilde{\mathbf{p}}\|^2} \leq \frac{4}{(1-\rho_{TG})^2} \leq \frac{4}{(1-\rho_{TG})^2} \|\widetilde{\pi}\|^2 \,,$$

the last inequality coming from $\|\widetilde{\pi}\|^2 \geq 1$. ■

This theorem can be applied in the context of Fourier analysis, setting $\widetilde{\Lambda} = \widetilde{\Lambda}^{(k,j)}$ and $\widetilde{\mathbf{p}} = \widetilde{\mathbf{p}}_j^{(k-1)}$, where $\widetilde{\Lambda}^{(k,j)}$ and $\widetilde{\mathbf{p}}_j^{(k-1)}$ come from the block representation of $\widetilde{A}_k = N_k^{(2\nu)\,1/2} A_k \, N_k^{(2\nu)\,1/2}$ and $\widetilde{P}_k = N_k^{(2\nu)\,-1/2} P_k$. Hence, the main constants for block $j < l_k$ at level k are

$$\widetilde{\alpha}_\nu^{(k,j)} = \sum_{i=2}^{m_j^{(k)}} \frac{\widetilde{\lambda}_i^{(k,j)\,2} \, |(\widetilde{\mathbf{p}}_j^{(k-1)})_i|^2}{\widetilde{\lambda}_1^{(k,j)\,2} \, \|\widetilde{\mathbf{p}}_j^{(k-1)}\|^2} = \frac{\sum_{i=2}^{m_j^{(k)}} \sigma_i^{(k,j)} \, \lambda_i^{(k,j)\,2} \, |(\mathbf{p}_j^{(k-1)})_i|^2}{\left(\sigma_1^{(k,j)} \lambda_1^{(k,j)}\right)^2 \sum_{i=1}^{m_j^{(k)}} \sigma_i^{(k,j)\,-1} \, |(\mathbf{p}_j^{(k-1)})_i|^2} \tag{4.28}$$

and

$$\widetilde{\beta}_\nu^{(k,j)} = \sum_{i=2}^{m_j^{(k)}} \frac{\widetilde{\lambda}_i^{(k,j)\,2} \, |(\widetilde{\mathbf{p}}_j^{(k-1)})_i|^2}{\widetilde{\lambda}_1^{(k,j)\,2} \, |(\widetilde{\mathbf{p}}_j^{(k-1)})_1|^2} = \sum_{i=2}^{m_j^{(k)}} \frac{\sigma_i^{(k,j)} \, \lambda_i^{(k,j)\,2} \, |(\mathbf{p}_j^{(k-1)})_i|^2}{\sigma_1^{(k,j)} \, \lambda_1^{(k,j)\,2} \, |(\mathbf{p}_j^{(k-1)})_1|^2} \,, \tag{4.29}$$

where we use subscript ν to recall that these quantities inherit the dependence of $\sigma_i^{(k,j)}$ on the number of smoothing steps. Taking all blocks other than l_k into account, we set

$$\widetilde{\alpha}_\nu^{(k)} = \max_{j=0,\ldots,l_k-1} \widetilde{\alpha}_\nu^{(k,j)} \qquad \text{and} \qquad \widetilde{\beta}_\nu^{(k)} = \max_{j=0,\ldots,l_k-1} \widetilde{\beta}_\nu^{(k,j)} \,. \tag{4.30}$$

Considering the contribution of block l_k to both $1 - \delta_k^{(\nu)\,-1}$, $\rho(E_{MG}^{(k)})$ and $\rho(E_{TG}^{(k)})$, it is given by $1 - \widetilde{\lambda}_1^{(k,l_k)}$. It is not surprising that the contribution is the same since no coarse-grid correction is performed on the corresponding modes, which therefore undergo only the action of the smoother.

Our definition (4.9) of the smoothing factor entails

$$\min\left(\widetilde{\lambda}_1^{(k,l_k)}, \min_{j=1,\ldots,l_k-1} \widetilde{\lambda}_2^{(k,j)}\right) = 1 - \left(\mu^{(k)}\right)^{2\nu} . \tag{4.31}$$

Hence, using successively the right inequality (4.18) (with second term in the minimum), the results in [27, Section 7.2] (for the proof of $\rho(E_{TG}^{(J)}) \leq \rho(E_{MG}^{(J)})$), the inequality (4.10) and the left inequality (4.15) (with $1 - \lambda_1/\lambda_2$ bounded above by 1), one obtains the following cascade of inequalities

$$(\mu^{(J)})^{2\nu} \;\leq\; \rho(E_{TG}^{(J)}) \;\leq\; \rho(E_{MG}^{(J)}) \;\leq\; 1 - \min_{1\leq k\leq J} \delta_k^{(\nu)} \;\leq\; \max_{1\leq k\leq J} \frac{\left(\mu^{(k)}\right)^{2\nu} + \widetilde{\alpha}_\nu^{(k)}}{1 + \widetilde{\alpha}_\nu^{(k)}} . \tag{4.32}$$

Observe that if $\max_{k=1,\ldots,J} \mu^{(k)} \approx \mu^{(J)}$ (which often holds in practice) and if $\widetilde{\alpha}_\nu^{(k)}$ is nicely bounded at each level, these inequalities define a narrow interval containing both the two-grid and V-cycle multigrid convergence factors. On the other hand, if $\widetilde{\alpha}_\nu^{(k)}$ is large at some levels, the right inequality (4.32) becomes ineffective, and the right inequality (4.15) further shows that $1 - \min_{1\leq k\leq J} \delta_k^{(\nu)}$ will be indeed close to 1. As observed in Chapter 3, the actual convergence of the V-cycle may then scale poorly with the number of levels.

Now, the smoothing factor $\mu^{(k)}$ can be directly assessed from $\lambda_i^{(k,j)}$ and $\gamma_i^{(k,j)}$, $i = 1, \ldots, m_j^{(k)}, j = 1, \ldots, l_k$. However, $\widetilde{\alpha}_\nu^{(k)}$ and $\widetilde{\beta}_\nu^{(k)}$ are related to the $\sigma_i^{(k,j)}$, which are known only via the relation

$$1 - \sigma_i^{(k,j)} \lambda_i^{(k,j)} = (1 - \gamma_i^{(k,j)\,-1} \lambda_i^{(k,j)})^{2\nu} .$$

Hence, $\widetilde{\alpha}_\nu^{(k)}$ and $\widetilde{\beta}_\nu^{(k)}$ may be difficult to assess, and their dependence on ν is also unclear. It is therefore not obvious to predict how the previous cascade of inequalities (4.32) evolves with respect to this parameter. The easiest way to overcome this difficulty it to use (4.10) combined with (4.13). The cascade of inequalities then becomes

$$\begin{aligned}(\mu^{(J)})^{2\nu} \leq \rho(E_{TG}^{(J)}) \leq \rho(E_{MG}^{(J)}) \leq 1 - \min_{1\leq k\leq J} \delta_k^{(\nu)} \\ \leq \max_{1\leq k\leq J} \frac{\delta_k^{(1)\,-1} - 1}{\delta_k^{(1)\,-1} - 1 + \nu} \leq \max_{1\leq k\leq J} \frac{\left(\mu^{(k)}\right)^2 + \widetilde{\alpha}_1^{(k)}}{\left(\mu^{(k)}\right)^2 + \widetilde{\alpha}_1^{(k)} + \nu(1 - \left(\mu^{(k)}\right)^2)} .\end{aligned} \tag{4.33}$$

Note that the two rightmost bounds behave like $\mathcal{O}(\nu^{-1})$, whereas the smoothing factor alone suggests an exponential dependence via the left inequality (4.33). For the typical example considered in Section 4.6, the actual behavior of both $\rho(E_{MG}^{(J)})$ and $\rho(E_{TG}^{(J)})$ is close to $\mathcal{O}(\nu^{-1})$ (see Figure 4.1), indicating that the upper bounds provide a more realistic estimate, at least in the considered case.

Now, in Theorem 4.3 below, we show that $\widetilde{\alpha}_\nu^{(k)}$ and $\widetilde{\beta}_\nu^{(k)}$ cannot increase with ν and we further relate these constants to

$$\alpha^{(k)} = \max_{j=1,\dots,l_k-1} \frac{\sum_{i=2}^{m_j^{(k)}} \gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)\,2}\ |(\mathbf{p}_j^{(k-1)})_i|^2}{\left(\gamma_1^{(k,j)\,-1}\ \lambda_1^{(k,j)}\right)^2 \sum_{i=1}^{m_j^{(k)}} \gamma_i^{(k,j)} |(\mathbf{p}_j^{(k-1)})_i|^2}\,, \tag{4.34}$$

$$\beta^{(k)} = \max_{j=1,\dots,l_k-1} \sum_{i=2}^{m_j^{(k)}} \frac{\gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)\,2}\ |(\mathbf{p}_j^{(k-1)})_i|^2}{\gamma_1^{(k,j)\,-1}\ \lambda_1^{(k,j)\,2}\ |(\mathbf{p}_j^{(k-1)})_1|^2}\,. \tag{4.35}$$

Note, however, that these expressions make sense only if, similarly to $\sigma_1^{(k,j)}\lambda_1^{(k,j)} = \min_{1\le s\le m_j^{(k)}} \sigma_s^{(k,j)}\lambda_s^{(k,j)}$, one also has

$$\gamma_1^{(k,j)\,-1}\ \lambda_1^{(k,j)} = \min_{1\le s\le m_j^{(k)}} \gamma_s^{(k,j)\,-1}\ \lambda_s^{(k,j)}. \tag{4.36}$$

This, in turn, holds if,

$$\gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)} \le 2 - \min_{1\le s\le m_j^{(k)}} \gamma_s^{(k,j)\,-1}\ \lambda_s^{(k,j)} \quad , \qquad i = 1,\dots,m_j^{(k)}\,. \tag{4.37}$$

Indeed, (4.37) implies in particular $\min_{1\le i\le m_j^{(k)}} \gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)} \le 1$, and further

$$\begin{aligned}|1 - \min_{1\le s\le m_j^{(k)}} \gamma_s^{(k,j)\,-1}\ \lambda_s^{(k,j)}| &= 1 - \min_{1\le s\le m_j^{(k)}} \gamma_s^{(k,j)\,-1}\ \lambda_s^{(k,j)} \\ &\ge \max(1 - \gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)}, \gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)} - 1)\,,\end{aligned}$$

hence (4.36) by virtue of the ordering (4.8). Observe that (4.37) holds when the smoother is scaled in such a way that the eigenvalues of $R_k^{-1}A_k$ do not exceed 1. On the other hand, if the smoother is related to some damping factor, the condition (4.37) is in fact a constraint on the latter: assuming

$$\gamma_i^{(k,j)\,-1} = \bar{\omega}\ \bar{\gamma}_i^{(k,j)\,-1}\ ,$$

it amounts to, taking all blocks other than l_k into account,

$$\bar{\omega} \ \le \ \min_{1\le j\le l_k-1} \frac{2}{\max_{1\le s\le m_j^{(k)}} \gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)} + \min_{1\le s\le m_j^{(k)}} \gamma_i^{(k,j)\,-1}\ \lambda_i^{(k,j)}}\,.$$

Theorem 4.3. *Let $\lambda_i > 0$ and $\gamma_i > 0$, $i = 1, ..., m$ satisfy $|1 - \lambda_1\gamma_1^{-1}| \geq |1 - \lambda_2\gamma_2^{-1}| \geq \cdots \geq |1 - \lambda_m\gamma_m^{-1}|$ and set*

$$\sigma_i^{(\nu)} = \lambda_i^{-1}\left(1 - (1 - \lambda_i\gamma_i^{-1})^{2\nu}\right) \quad , \qquad i = 1, ..., m\,,$$

for some integer $\nu > 0$. Let

$$\widetilde{\alpha}_\nu = \frac{\sum_{i=2}^{m} \sigma_i^{(\nu)}\lambda_i^2|p_i|^2}{\left(\sigma_1^{(\nu)}\lambda_1\right)^2 \sum_{i=1}^{m} \sigma_i^{(\nu)\,-1}\,|p_i|^2}\,,$$

and, if $|p_1| > 0$,

$$\widetilde{\beta}_\nu = \sum_{i=2}^{m} \frac{\sigma_i^{(\nu)}\lambda_i^2|p_i|^2}{\sigma_1^{(\nu)}\lambda_1^2|p_1|^2}\,.$$

One has

$$\left(\frac{1}{\nu}\right)^2 \widetilde{\alpha}_1 \leq \widetilde{\alpha}_\nu \leq \widetilde{\alpha}_1\,, \tag{4.38}$$

and, if $|p_1| > 0$,

$$\frac{1}{\nu}\widetilde{\beta}_1 \leq \widetilde{\beta}_\nu \leq \widetilde{\beta}_1\,. \tag{4.39}$$

Moreover, if

$$\max_{1\leq i\leq m} \gamma_i^{-1}\lambda_i \leq 2 - \min_{1\leq i\leq m} \gamma_i^{-1}\lambda_i \tag{4.40}$$

let

$$\alpha = \frac{\sum_{i=2}^{m} \gamma_i^{-1}\lambda_i^2|p_i|^2}{\left(\gamma_1^{-1}\lambda_1\right)^2 \sum_{i=1}^{m} \gamma_i|p_i|^2}\,,$$

and, if $|p_1| > 0$,

$$\beta = \sum_{i=2}^{m} \frac{\gamma_i^{-1}\;\lambda_i^2|p_i|^2}{\gamma_1^{-1}\;\lambda_1^2|p_1|^2}\,.$$

One has

$$\left(\frac{2-\omega}{2\nu}\right)^2 \alpha \leq \widetilde{\alpha}_\nu \leq \alpha\,, \tag{4.41}$$

and, if $|p_1| > 0$,

$$\frac{2-\omega}{2\nu}\beta \leq \widetilde{\beta}_\nu \leq \beta\,. \tag{4.42}$$

where $\omega = \max_{1\leq i\leq m}\gamma_i^{-1}\lambda_i$.

Further, letting

$$\phi = \arccos\left(\frac{|p_1|}{\sqrt{\sum_{i=1}^{m}|p_i|^2}}\right)\,,$$

one has, for any μ such that $|1 - \gamma_i^{-1}\lambda_i| \leq \mu$, $i = 2, ..., m$

$$(1-\mu)^2\frac{\min_{2\leq i\leq m}\gamma_i}{\max_{1\leq i\leq m}\gamma_i}\left(\frac{\sin\phi}{\gamma_1^{-1}\lambda_1}\right)^2 \leq \alpha \leq \omega^2\;\frac{\max_{2\leq i\leq m}\gamma_i}{\min_{1\leq i\leq m}\gamma_i}\left(\frac{\sin\phi}{\gamma_1^{-1}\lambda_1}\right)^2\,, \tag{4.43}$$

and, if $|p_1| > 0$,

$$(1-\mu)^2 \frac{\min_{2\le i\le m}\gamma_i}{\gamma_1}\left(\frac{\tan\phi}{\gamma_1^{-1}\lambda_1}\right)^2 \le \beta \le \omega^2\ \frac{\max_{2\le i\le m}\gamma_i}{\gamma_1}\left(\frac{\tan\phi}{\gamma_1^{-1}\lambda_1}\right)^2 . \tag{4.44}$$

Proof. First, observe that we have

$$1-\lambda_i\sigma_i^{(1)} = (1-\gamma_i^{-1}\lambda_i)^2\,,\ \ i=1,...,m\,,$$

and, hence, the assumed ordering is equivalent to $\lambda_1\sigma_1^{(1)} \le \lambda_2\sigma_2^{(1)} \le \cdots \le \lambda_m\sigma_m^{(1)}$. Moreover, when (4.40) holds, one also has $\gamma_1^{-1}\lambda_1 = \min_{1\le i\le m}\gamma_i^{-1}\lambda_i$. We also note that it is sufficient to prove inequalities (4.41) and (4.42) for $\nu=1$, the general case following from (4.38) and (4.39), respectively.

Next, observe that

$$\frac{\sigma_i^{(\nu)}}{\sigma_i^{(1)}} = \frac{1-(1-\gamma_i^{-1}\lambda_i)^{2\nu}}{1-(1-\gamma_i^{-1}\lambda_i)^2} = \frac{1-(1-s)^\nu}{s} = \sum_{k=0}^{\nu}(1-s)^k \tag{4.45}$$

with $s = 1-(1-\gamma_i^{-1}\lambda_i)^2 = \lambda_i\sigma_i^{(1)} \in [0,1]$, is a decreasing function of $\lambda_i\sigma_i^{(1)}$. Hence, since $\lambda_1\sigma_1^{(1)} \le \lambda_2\sigma_2^{(1)} \le \cdots \le \lambda_m\sigma_m^{(1)}$, one has

$$\sigma_i^{(\nu)} \le \frac{\sigma_1^{(\nu)}\sigma_i^{(1)}}{\sigma_1^{(1)}}\,,\quad i=1,...,m\,.$$

The right inequalities (4.38) and (4.39) straightforwardly follow.

Similarly, since

$$\frac{\sigma_i^{(1)}}{\gamma_i^{-1}} = \frac{1-(1-\gamma_i^{-1}\lambda_i)^2}{\gamma_i^{-1}\lambda_i} = (2-\gamma_i^{-1}\lambda_i)\,, \tag{4.46}$$

the equality $\gamma_1^{-1}\lambda_1 = \min_{1\le i\le m}\gamma_i^{-1}\lambda_i$ implies

$$\sigma_i^{(1)} \le \frac{\sigma_1^{(1)}\gamma_i^{-1}}{\gamma_1^{-1}}\,,\quad i=1,...,m\,;$$

hence the right inequalities (4.41) and (4.42).

Next, from

$$\frac{\sigma_i^{(\nu)}}{\sigma_i^{(1)}} = \sum_{k=0}^{\nu-1}(1-\gamma_i^{-1}\lambda_i)^{2k}$$

we conclude

$$1 \le \frac{\sigma_i^{(\nu)}}{\sigma_i^{(1)}} \le \nu\,;$$

hence the left inequalities (4.38) and (4.39). Similarly, from (4.46) we have

$$2 - \omega \leq \frac{\sigma_i^{(1)}}{\gamma_i^{-1}} \leq 2$$

which in turn implies the left inequalities (4.41) and (4.42).

Finally, for the proof of (4.43) and (4.44) we first note that $|1 - \gamma_i^{-1}\lambda_i| \leq \mu$, $i = 2, ..., m$ and the definition $\omega = \max_{1 \leq i \leq m} \gamma_i^{-1}\lambda_i$ imply $1 - \mu \leq \gamma_i^{-1}\lambda_i \leq \omega$. Hence

$$\left(\frac{1-\mu}{\gamma_1^{-1}\lambda_1}\right)^2 \frac{\sum_{i=2}^m \gamma_i |p_i|^2}{\sum_{i=1}^m \gamma_i |p_i|^2} \leq \alpha \leq \left(\frac{\omega}{\gamma_1^{-1}\lambda_1}\right)^2 \frac{\sum_{i=2}^m \gamma_i |p_i|^2}{\sum_{i=1}^m \gamma_i |p_i|^2},$$

and, if $|p_1| > 0$,

$$\left(\frac{1-\mu}{\gamma_1^{-1}\lambda_1}\right)^2 \sum_{i=2}^m \frac{\gamma_i |p_i|^2}{\gamma_1 |p_1|^2} \leq \beta \leq \left(\frac{\omega}{\gamma_1^{-1}\lambda_1}\right)^2 \sum_{i=2}^m \frac{\gamma_i |p_i|^2}{\gamma_1 |p_1|^2}.$$

The conclusion follows since

$$\sin^2\phi = \frac{\sum_{i=2}^m |p_i|^2}{\sum_{i=1}^m |p_i|^2} \qquad \text{and} \qquad \tan^2\phi = \frac{\sum_{i=2}^m |p_i|^2}{|p_1|^2}. \qquad \blacksquare$$

This theorem shows that, from a qualitative viewpoint, it is sufficient to analyze $\alpha^{(k,j)}$ and $\beta^{(k,j)}$, which involve only $\gamma_i^{(k,j)}, \lambda_i^{(k,j)}$ and $\mathbf{p}_j^{(k-1)}$. Further, if, as often arises, the smoother is well conditioned, all $\gamma_i^{(k,j)}$ are approximately equal (they are all equal for damped Jacobi smoothing). Then $\widetilde{\alpha}_\nu^{(k,j)}$ and $\widetilde{\beta}_\nu^{(k,j)}$, $j < l_k$, behave essentially like $\left(\frac{\sin\phi^{(k,j)}}{\gamma_1^{(k,j)}{}^{-1}\lambda_1^{(k,j)}}\right)^2$ and $\left(\frac{\tan\phi^{(k,j)}}{\gamma^{(k,j)_1}{}^{-1}\lambda_1^{(k,j)}}\right)^2$, respectively, where $\phi^{(k,j)}$ is the angle between the eigenvector associated to $\lambda_1^{(k,j)}$ and the range of the prolongation.

This allows to discuss the condition for having satisfactory two-grid convergence and a satisfactory V-cycle convergence estimate via McCormick's bound (4.10). Considering (4.17) and (4.18), only blocks for which $\widetilde{\lambda}_1^{(k,j)} = \sigma_1^{(k,j)}\lambda_1^{(k,j)}$ is small have to by analyzed carefully. This sounds logical: if all modes inside a block are efficiently relaxed by the smoother, it does not matter that much how the restriction and prolongation operate on these modes. Now, provided that the smoothing factor is bounded away from 1, one will have nice two-grid convergence if and only if, for each block with small $\gamma_1^{(k,j)}{}^{-1}\lambda_1^{(k,j)}$, the quantity $\widetilde{\lambda}_1^{(k,j)}\widetilde{\beta}_\nu^{(k,j)}$ is reasonably bounded above; that is, if $\tan\phi^{(k,j)} \leq c \cdot \left(\gamma_1^{(k,j)}{}^{-1}\lambda_1^{(k,j)}\right)^{1/2}$. On the other hand, the condition is stronger for having a nice V-cycle convergence estimate: this requires $\widetilde{\beta}_\nu^{(k,j)}$ to be bounded above; that is, $\tan\phi^{(k,j)} \leq c \cdot \gamma_1^{(k,j)}{}^{-1}\lambda_1^{(k,j)}$.

Some heuristics present in the multigrid literature [18, p. 1573](see also [21, p. 4]) state: "Interpolation must be able to approximate an eigenvector with error bound proportional to the size of the associated eigenvalue". Our results give a more precise interpretation of such statements. For mere two-grid convergence the tangent of the angle between the eigenvector and the range of the prolongation should be proportional to the square root of the eigenvalue, whereas guaranteed V-cycle convergence requires it be proportional to the eigenvalue.

4.5 Semi-positive definite problems and local Fourier analysis

In this section we consider Fourier analysis for symmetric semi-positive definite linear systems. Such extension is motivated by local Fourier analysis (also called local mode analysis) that has a wider scope than (rigorous) Fourier analysis. The main idea is the assessment of the two-grid convergence or of the smoothing factor without taking boundary conditions into account. In practice, such approach is often equivalent to the use of periodic boundary conditions and therefore leads to linear systems with non-trivial null space.

Another way to interpret local Fourier analysis is to consider it as a limit case of (rigorous) Fourier analysis for SPD problems on grids of increasing size (with, thus, decreasing influence of boundary conditions on estimated parameters). Now, we previously observed for the SPD case that the angle between the range of prolongation and an eigenvector of A should be proportional to the size of the eigenvalue. In the limit case of local Fourier analysis, the modes belonging to the null space $\mathcal{N}(A)$ of A should therefore be interpolated exactly, which also corresponds to a common practice. It then follows that null space components seemingly play no role in the convergence, and hence that Fourier analysis may be carried out ignoring these modes. This, indeed, is the common practice when assessing the two-grid convergence factor (see, for instance, [61, p.109], [68, p.107] and the references therein).

In Theorem 4.4 below we give theoretical foundation to this approach with respect to V-cycle multigrid, showing that McCormick's bound on the convergence rate can also be computed ignoring singular modes, or, more precisely, restricting the minimum in (4.12) to vectors belonging to the range of A_k. Since (4.12) is at the root of the further analysis developed in Theorems 4.2 and 4.3, the application of the results in Section 4.4 to local mode analysis is then straightforward.

Now, to state our theorem, we need to extend our definition of the V-cycle multigrid algorithm in Section 4.2 to A_k possibly singular. The only potential difficulty comes in fact with the bottom level matrix A_0 whose inverse is needed. In Theorem 4.4, we assume that instead one uses any matrix B_0 such that $A_0B_0A_0 = A_0$. Such matrices are called $\{1\}$-inverse in [6], and one may check that if $\mathbf{r}_0 \in \mathcal{R}(A_0)$, then $A_0B_0\mathbf{r}_0 = \mathbf{r}_0$. On the other hand, to generalize (4.12), we need the inverse of the restriction of A_k to its range. The most convenient way to express it is to use the Moore-Penrose inverse A_k^+ of A_k, since, if $A_k = X\mathrm{diag}(\lambda_i)X^T$, then $A_k^+ = X\mathrm{diag}(\lambda_i^+)X^T$ with

$$\lambda_i^+ = \begin{cases} \lambda_i^{-1} & \text{if } \lambda_i \neq 0\,, \\ 0 & \text{otherwise.} \end{cases}$$

The expression of A_k^+ is thus particularly simple when using the Fourier basis which makes A_k diagonal.

Theorem 4.4. *Let $\mathbf{x}_{n+1} = MG(\mathbf{b}, A, \mathbf{x}_n, J)$ be the vector resulting from the application of the multigrid algorithm with V-cycle at level $J > 1$, where A is symmetric semi-positive definite, and, in case A_0 is singular, where A_0^{-1} is exchanged for any matrix $A_0^{(1)}$ such that $A_0A_0^{(1)}A_0 = A_0$. Assume that P_k , $k = 0, \dots, J-1$, A_k , $k = 0, \dots, J$, and R_k , $k = 1, \dots, J$ satisfy the general assumptions stated in Section 4.2 with $\rho(I - R_k^{-1}A_k) < 1$ being replaced by $\rho(I - R_k^{-1}A_k) \leq 1$, with $(I - R_k^{-1}A_k)\mathbf{z} = \lambda\mathbf{z}$ for $|\lambda| = 1$ if and only if $\mathbf{z} \in \mathcal{N}(A_k)$. Let $\mathcal{P}_{\mathcal{R}(A),\mathcal{N}(A)}$ be the orthogonal projector onto the range of A.*

If $\mathbf{b} \in \mathcal{R}(A)$, then, for any solution $\tilde{\mathbf{x}}$ to (4.1),

$$\mathcal{P}_{\mathcal{R}(A),\mathcal{N}(A)}\,(\tilde{\mathbf{x}} - \mathbf{x}_{n+1}) = E_{MG}^{(J)}\,(\tilde{\mathbf{x}} - \mathbf{x}_n) = E_{MG}^{(J)}\;\mathcal{P}_{\mathcal{R}(A),\mathcal{N}(A)}\,(\tilde{\mathbf{x}} - \mathbf{x}_n)$$

for some matrix $E_{MG}^{(J)}$ satisfying

$$\rho(E_{MG}^{(J)}) \;\leq\; 1 - \min_{1\leq k\leq J} \widetilde{\delta}_k^{(\nu)} \quad ,$$

with

$$\widetilde{\delta}_k^{(\nu)} \;=\; \min_{\mathbf{v}_k\in\mathcal{R}(A_k)} \frac{\mathbf{v}_k^T\; N_k^{(2\nu)}\;\mathbf{v}_k}{\mathbf{v}_k^T(A_k^+ - P_{k-1}A_{k-1}^+P_{k-1}^T)\mathbf{v}_k}\,. \tag{4.47}$$

Proof. Let $q_k = \dim(\mathcal{N}(A_k))$. Observe that, A_k being non-negative definite, $A_{k-1} = P_{k-1}^TA_kP_{k-1}$ is non-negative definite with $q_{k-1} \leq q_k$. Without loss of generality, we can express all the matrices using bases of $\mathbb{R}^{n_k}$, $k = J, ..., 0$ such that, when $q_k > 0$, the first q_k canonical vectors span $\mathcal{N}(A_k)$. Hence, A_k admits a block representation

$$A_k = \begin{pmatrix} O_{q_k,q_k} & \\ & A_k^{\mathcal{R}\mathcal{R}} \end{pmatrix}\,, \tag{4.48}$$

with all but lower right blocks being empty if $q_k = 0$.

Similarly, we partition

$$R_k = \begin{pmatrix} R_k^{\mathcal{NN}} & R_k^{\mathcal{NR}} \\ R_k^{\mathcal{RN}} & R_k^{\mathcal{RR}} \end{pmatrix}, \; N_k^{(\nu)} = \begin{pmatrix} N_k^{(\nu)\,\mathcal{NN}} & N_k^{(\nu)\,\mathcal{NR}} \\ N_k^{(\nu)\,\mathcal{RN}} & N_k^{(\nu)\,\mathcal{RR}} \end{pmatrix}, \; P_{k-1} = \begin{pmatrix} P_{k-1}^{\mathcal{NN}} & P_{k-1}^{\mathcal{NR}} \\ P_{k-1}^{\mathcal{RN}} & P_{k-1}^{\mathcal{RR}} \end{pmatrix},$$

where all but lower right blocks of R_k, $N_k^{(\nu)}$ and P_{k-1} become empty when $q_k = 0$, and where $P_{k-1}^{\mathcal{NN}}$ and $P_{k-1}^{\mathcal{RN}}$ are empty when $q_k > 0$ with $q_{k-1} = 0$. If $q_k > 0$, there holds

$$A_{k-1} = P_{k-1}^T A_k P_{k-1} = \begin{pmatrix} P_{k-1}^{\mathcal{RN}\,T} A_k^{\mathcal{RR}} P_{k-1}^{\mathcal{RN}} & P_{k-1}^{\mathcal{RN}\,T} A_k^{\mathcal{RR}} P_{k-1}^{\mathcal{RR}} \\ P_{k-1}^{\mathcal{RR}\,T} A_k^{\mathcal{RR}} P_{k-1}^{\mathcal{RN}} & P_{k-1}^{\mathcal{RR}\,T} A_k^{\mathcal{RR}} P_{k-1}^{\mathcal{RR}} \end{pmatrix} .$$

Hence, in view of the form (4.48) and the fact that $A_k^{\mathcal{RR}}$ is SPD, one must have $P_{k-1}^{\mathcal{RN}} = O$. It then follows that, for any $n_k \times n_k$ matrix

$$B_{k-1} = \begin{pmatrix} * & * \\ * & B_{k-1}^{\mathcal{RR}} \end{pmatrix} ,$$

one has

$$P_{k-1} B_{k-1} P_{k-1}^T = \begin{pmatrix} * & * \\ * & P_{k-1}^{\mathcal{RR}} B_k^{\mathcal{RR}} \, P_{k-1}^{\mathcal{RR}\,T} \end{pmatrix} , \tag{4.49}$$

This latter relation also holds for $q_k = 0$, the blocks denoted by a star $*$ being then empty.

Now, $\mathbf{x}_{n+1} = \mathrm{MG}(\mathbf{b}, A_k, \mathbf{x}_n, k)$ may be expressed as

$$\mathbf{x}_{n+1} = \mathbf{x}_n + B_J(\mathbf{b} - A\mathbf{x}_n), \tag{4.50}$$

where the matrix B_J is defined from the recursion

$$\begin{aligned} &B_0 = A_0^{(1)} \\ &B_k = N_k^{(2\nu)} - (I - N_k^{(\nu)} \, A_k) P_{k-1} B_{k-1} P_{k-1}^T (I - A_k \, N_k^{(\nu)}) , \qquad k = 1, \dots, J \end{aligned}$$

(see, e.g., [65, Section 5.1] ; B_k^{-1} in this reference corresponds to B_k here).

Since there holds

$$I - N_k^{(2\nu)} \, A_k = \begin{pmatrix} I_{q_k} & * \\ & I - N_k^{(2\nu)\,\mathcal{RR}} \, A_k^{\mathcal{RR}} \end{pmatrix} \tag{4.51}$$

with all but lower right blocks being empty if $q_k = 0$, letting

$$B_J = \begin{pmatrix} * & * \\ * & B_J^{\mathcal{RR}} \end{pmatrix} ,$$

it follows from (4.49) that $B_J^{\mathcal{RR}}$ may be computed from the recursion

$$B_k^{\mathcal{RR}} = N_k^{(2\nu)\,\mathcal{RR}} - (I - N_k^{(\nu)\,\mathcal{RR}} A_k^{\mathcal{RR}}) P_{k-1}^{\mathcal{RR}} B_{k-1}^{\mathcal{RR}} \, P_{k-1}^{\mathcal{RR}\,T} (I - A_k^{\mathcal{RR}} \, N_k^{(\nu)\,\mathcal{RR}}) \, , \qquad k = 1, \ldots, J \; .$$

On the other hand, when A_0 is singular, $A_0 B_0 A_0 = A_0$ holds for A_0 of the form (4.48) if and only if $B_0^{\mathcal{RR}} = A_0^{\mathcal{RR}\,-1}$, whereas from (4.51) we deduce

$$I - N_k^{(2\nu)\,\mathcal{RR}} \, A_k^{\mathcal{RR}} = (I - N_k^{(\nu)\,\mathcal{RR}} \, A_k^{\mathcal{RR}})^2 \, .$$

Hence $E_k = I - B_k^{\mathcal{RR}} A_k^{\mathcal{RR}}$ obeys the recursion

$$\begin{aligned} E_0 \; &= O \\ E_k \; &= (I - N_k^{(\nu)\,\mathcal{RR}} \, A_k^{\mathcal{RR}}) \left(I - P_{k-1}^{\mathcal{RR}} (I - E_{k-1}) \, A_{k-1}^{\mathcal{RR}\,-1} \, P_{k-1}^{\mathcal{RR}\,T} \, A_k^{\mathcal{RR}} \right) (I - N_k^{(\nu)\,\mathcal{RR}} \, A_k^{\mathcal{RR}}) \, , \\ & \quad k = 1, 2, \ldots, J \; . \end{aligned}$$

similar to (4.4); that is, corresponding to a multigrid scheme satisfying all assumptions of Theorem 4.1, which therefore implies

$$\rho(E_J) \leq 1 - \max_{1 \leq k \leq J} \delta_k^{(\nu)} \tag{4.52}$$

with

$$\delta_k^{(\nu)} \; = \; \min_{\mathbf{v} \in \mathbb{R}^{n_k - q_k}} \; \frac{\mathbf{v}^T N_k^{\mathcal{RR}} \mathbf{v}}{\mathbf{v}^T (\, A_k^{\mathcal{RR}\,-1} - P_{k-1}^{\mathcal{RR}} \, A_{k-1}^{\mathcal{RR}\,-1} \, P_{k-1}^{\mathcal{RR}\,T}) \mathbf{v}} \, .$$

Moreover, $\widetilde{\delta}_k^{(\nu)} \; = \delta_k^{(\nu)}$ since

$$A_k^+ = \begin{pmatrix} O & \\ & A_k^{\mathcal{RR}\,-1} \end{pmatrix}$$

with all but lower right block being empty when $q_k = 0$.

Finally, using (4.50) and the fact that $\mathbf{b} \in \mathcal{R}(A)$, there holds

$$\tilde{\mathbf{x}} - \mathbf{x}_{n+1} = (I - B_J A)(\tilde{\mathbf{x}} - \mathbf{x}_n),$$

and hence

$$\mathcal{P}_{\mathcal{R}(A),\mathcal{N}(A)} \, (\tilde{\mathbf{x}} - \mathbf{x}_{n+1}) = \begin{pmatrix} O & O \\ O & I \end{pmatrix} \begin{pmatrix} I & * \\ & I - B_k^{\mathcal{RR}} A_k^{\mathcal{RR}} \end{pmatrix} (\tilde{\mathbf{x}} - \mathbf{x}_n)$$

$$= \begin{pmatrix} O & O \\ O & E_J \end{pmatrix} (\tilde{\mathbf{x}} - \mathbf{x}_n)$$

$$= \begin{pmatrix} O & O \\ O & E_J \end{pmatrix} \mathcal{P}_{\mathcal{R}(A),\mathcal{N}(A)} \ (\tilde{\mathbf{x}} - \mathbf{x}_n) \ ,$$

which, together with (4.52), concludes the proof. ■

4.6 Examples

4.6.1 Usual prolongations in 2D

In this subsection we show how the conclusions of Theorems 4.2 and 4.3 can be used to analyze usual prolongation operators presented in [68]. More precisely, we assess the parameter[2]

$$\tan^2 \phi^{(j)} = \frac{\sum_{i=2}^{m} |(\mathbf{p}^{(j)})_i|^2}{|(\mathbf{p}^{(j)})_1|^2} \ , \tag{4.53}$$

that characterizes the quality of a prolongation, according to (4.44). Indeed, see Section 4.4, if $\tan \phi^{(j)} \leq c \cdot \left({\gamma_1^{(j)}}^{-1} \lambda_1^{(j)} \right)^{1/2}$, $j = 1, ..., l$, holds for not too large c, then a smoothing factor bounded away from one guarantees optimal two-grid convergence, whereas the condition $\tan \phi^{(j)} \leq c \cdot {\gamma_1^{(j)}}^{-1} \lambda_1^{(j)}$ leads to optimal V-cycle multigrid convergence. Since $\gamma_1^{(j)}$ is often close to 1, it is thus critical to check the behavior of $\tan \phi^{(j)}$ when $\lambda_1^{(j)}$ becomes close to 0.

Such a discussion presumes that the prolongation P (which fixes $\tan \phi^{(j)}$) and the coarse-grid matrix A (which determines $\lambda_i^{(j)}$) are both known. Here, we develop a slightly different approach and, for various prolongation operators (taken from [68]), we indicate conditions that the eigenvalues of a potential system matrix A should satisfy, when used with such prolongations, in order to lead to optimal two- and multigrid algorithms. Instead of index j, we use a couple (θ_1,θ_2) of angles (for two-dimensional problems), adopting the notation of [68]. Although in the context of (rigorous) Fourier analysis θ_1 and θ_2 can take only a finite number of values (depending on the mesh-size assumed), we follow here the common practice and perform computations allowing all values inside a fixed interval $(0, \pi)$.

Before we characterize in Table 4.1 various prolongations taken from Table 6.2 in [68], we illustrate with bilinear prolongation how the corresponding results can be derived. From Table 6.2 in [68] we learn that the Fourier symbol for bilinear prolongation is $(1 + \cos(\theta_1))(1 + \cos(\theta_2))$. This means that for a block characterized by a couple (θ_1, θ_2)

[2] Here and in what follows we omit the grid number k, the discussion of this subsection does not depend on the choice of a particular grid.

we have (with elements ordered not necessarily according to (4.8))

$$\begin{aligned}\mathbf{p}^{(\theta_1,\theta_2)} &= \begin{pmatrix} (1+\cos(\theta_1))(1+\cos(\theta_2)) \\ (1+\cos(\theta_1+\pi))(1+\cos(\theta_2)) \\ (1+\cos(\theta_1))(1+\cos(\theta_2+\pi)) \\ (1+\cos(\theta_1+\pi))(1+\cos(\theta_2+\pi)) \end{pmatrix} \\ &= 4\begin{pmatrix} (1-\sin^2(\theta_1/2))(1-\sin^2(\theta_2/2)) \\ \sin^2(\theta_1/2)\sin^2(\theta_2/2) \\ (1-\sin^2(\theta_1/2))\sin^2(\theta_2/2) \\ \sin^2(\theta_1/2)(1-\sin^2(\theta_2/2)) \end{pmatrix}. \end{aligned} \tag{4.54}$$

There, $\mathbf{p}^{(\theta_1,\theta_2)}$ is one of the blocks (4.6) of P in the relevant Fourier basis. Now, small values of $\tan^2\phi^{(\theta_1,\theta_2)}$ are possible when all but one entry of $\mathbf{p}^{(\theta_1,\theta_2)}$ are small; that is, when both θ_1 and θ_2 are close to 0 or π. Since vectors $\mathbf{p}^{(\theta_1,\theta_2)}$, $\mathbf{p}^{(\theta_1+\pi,\theta_2)}$, $\mathbf{p}^{(\theta_1,\theta_2+\pi)}$ and $\mathbf{p}^{(\theta_1+\pi,\theta_2+\pi)}$ have same entries (ordered differently), in what follows we consider only the situation when $(\theta_1,\theta_2)\to(0,0)$ (the same comment holds for other prolongations in Table 4.1). Hence, for small θ_1 and θ_2,

$$\mathbf{p}^{(\theta_1,\theta_2)} = \begin{pmatrix} \mathcal{O}(1) \\ \mathcal{O}\left(\theta_1^2\right) \\ \mathcal{O}\left(\theta_2^2\right) \\ \mathcal{O}\left(\theta_1^2\theta_2^2\right) \end{pmatrix},$$

which, together with (4.53), gives, for the ordering satisfying (4.8) (when all γ_i are bounded below), $\tan^2\phi^{(\theta_1,\theta_2)} = \mathcal{O}\left(\theta_1^2+\theta_2^2\right)$. The same data is given in Table 4.1 for other prolongations coming from Table 6.2 in [68]. If we indicate $\mathcal{O}\left(\theta_1^\eta+\theta_2^\eta\right)$ in the third column it means that optimal two-grid convergence based on an optimal smoothing factor occurs if and only if the eigenvalues of the matrix tend to zero only for θ_1 and θ_2 approaching 0 or π, and not faster than $\mathcal{O}\left((\sin^\eta\theta_1+\sin^\eta\theta_2)^{1/2}\right)$, whereas guaranteed optimal V-cycle convergence requires the eigenvalue going to 0 not faster than $\mathcal{O}\left(\sin^\eta\theta_1+\sin^\eta\theta_2\right)$.

For example, consider any of the usual discretizations of an isotropic laplace operator on a uniform grid. The eigenvalues $\lambda^{(\theta_1,\theta_2)}$ satisfy the already observed symmetry $\lambda^{(\theta_1,\theta_2)} = \lambda^{(\theta_1+\pi,\theta_2)} = \lambda^{(\theta_1,\theta_2+\pi)} = \lambda^{(\theta_1+\pi,\theta_2+\pi)}$. Moreover, for $0\le\theta_1,\theta_2\le\pi/2$, the eigenvalue $\lambda^{(\theta_1,\theta_2)}$ becomes small only when (θ_1,θ_2) is close to $(0,0)$, behaving in it neighborhood as $\mathcal{O}\left({\theta_1}^2+{\theta_2}^2\right)$. In view of the results given in Table 4.1, and provided that the smoother with bounded away from one smoothing factor is used, all considered prolongations lead to an optimal two-grid cycle, and all but constant upwind prolongation are guaranteed optimal with V-cycle. Note that numerical experiments confirm the

prolongation	Fourier symbol	$\tan^2\phi^{(\theta_1,\theta_2)}$
bilinear	$(1+\cos(\theta_1))(1+\cos(\theta_1))$	$\mathcal{O}\left(\theta_1^2+\theta_2^2\right)$
bicubic	$(8+9\cos(\theta_1)-\cos(3\theta_1))$	$\mathcal{O}\left(\theta_1^4+\theta_2^4\right)$
	$\times(8+9\cos(\theta_2)-\cos(3\theta_2))$	
biquintic	$(128-150\cos(\theta_1)+25\cos(3\theta_1)-3\cos(5\theta_1))$	$\mathcal{O}\left(\theta_1^6+\theta_2^6\right)$
	$\times(128-150\cos(\theta_2)+25\cos(3\theta_2)-3\cos(5\theta_2))$	
constant upwind	$(1+\exp(i\theta_1))(1+\exp(\theta_2))$	$\mathcal{O}\left(\theta_1+\theta_2\right)$
seven point	$1+\cos(\theta_1)+\cos(\theta_2)+\cos(\theta_1-\theta_2)$	$\mathcal{O}\left(\theta_1^2+\theta_2^2\right)$

TABLE 4.1: Various prolongations coming from Table 6.2 in [68], with asymptotical behavior of $\tan^2\phi^{(j)}$ for small values of θ_1 and θ_2.

suboptimal behavior of the V-cycle in this case [41].

4.6.2 2D Poisson

In this subsection we illustrate quantitative aspects of the cascades of inequalities (4.32) and (4.33). We consider the linear system resulting from the bilinear finite element discretization of the two-dimensional Poisson problem

$$\begin{aligned} -\Delta\, u = f \quad & \text{in } \ \Omega = (0,1)\times(0,1) \\ u = 0 \quad & \text{in } \ \partial\Omega \end{aligned}$$

on a uniform grid of mesh size $h = 1/M_J$ in both directions. The matrix corresponds then to the following nine point stencil

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}. \tag{4.55}$$

Up to some scaling factor, this is also the stencil obtained with 9-point finite difference discretization. We assume $M_J = 2^J M_0$ for some integer M_0 , allowing J steps of regular geometric coarsening. We consider bilinear prolongation

$$P_k = \begin{pmatrix} J_k \\ I_{n_k} \end{pmatrix}$$

where J_k corresponds to the natural interpolation associated with bilinear finite element basis functions. The restriction P_k^T corresponds then to "full weighting", as defined in, e.g. [61] [3]. We consider damped Jacobi smoothing: $R_k = \omega_{\text{Jac}}^{-1}\text{diag}(A_k)$, with $\omega_{\text{Jac}} = 2/3$; that is, such that $\omega = \max_{\lambda\in\sigma(R_k^{-1}A_k)}\lambda = 1$. Since the stencil is preserved on all levels,

[3] up to some scaling factor; the scalings of the prolongation and restriction are unimportant when using coarse-grid matrices of the Galerkin type.

it is sufficient to consider only two successive grids; to alleviate notation, we therefore let $N = N_k^{(\nu)}$, $A = A_k$, $P = P_{k-1}$, $A_c = A_{k-1} = P^T AP$ and $\pi_A = \pi_{A_k} = PA_c^{-1}P^T A$.

Now, we asses μ and $\widetilde{\alpha}_\nu$ using (rigorous) Fourier analysis. The eigenvectors of A are, for $m,\, l = 1, \dots, M \;-\; 1$, the functions

$$u_{m,l}^{(M)} \;=\; \sin(m\pi x)\,\sin(l\pi y) \tag{4.56}$$

evaluated at the grid points. The eigenvalue corresponding to $u_{m,l}^{(M)}$ is

$$\lambda_{m,l}^{(M)} = 4(3s_m + 3s_l - 4s_m s_l) \tag{4.57}$$

where

$$s_m = \sin^2(\theta^{(m)}) \quad , \quad s_l = \sin^2(\theta^{(l)}) \ . \tag{4.58}$$

with

$$\theta^{(m)} = \frac{m\pi}{2M} \, .$$

The prolongation P satisfies (see, e.g., [61, p. 87])

$$P^T \left\{ \begin{array}{c} u_{m,l}^{(M)} \\ u_{M-m,M-l}^{(M)} \\ -u_{M-m,l}^{(M)} \\ -u_{m,M-l}^{(M)} \end{array} \right\} \;=\; 4 \left\{ \begin{array}{c} (1-s_m)(1-s_l) \\ s_m s_l \\ s_m(1-s_l) \\ (1-s_m)s_l \end{array} \right\} u_{m,l}^{(M/2)}$$

for $1 \le m, l \le M/2 - 1$, with $P^T u_{m,l}^{(M)} = 0$ for $m = M/2$ or $l = M/2$. Hence, P has a block structure (4.6) with blocks

$$\mathbf{p}_{m,l}^T = 4 \Big(\; (1-s_m)(1-s_l) \quad s_m s_l \quad s_m(1-s_l) \quad (1-s_m)s_l \; \Big) \ ,$$

which also correspond to the bilinear prolongation (4.54) considered previously. A and $N^{(2\nu)}$ are block diagonal like in (4.7) with diagonal blocks given by, respectively,

$$\begin{aligned} \Lambda_{m,l} &= \operatorname{diag}\left(\lambda_{m,l}^{(M)} \, , \, \lambda_{M-m,M-l}^{(M)} \, , \lambda_{m,M-l}^{(M)} \, , \, \lambda_{M-m,l}^{(M)}\right) \ , \\ \Sigma_{m,l}^{(\nu)} &= 64 \ \operatorname{diag}\left(\left\{ \frac{1-(1-\lambda_{c,s}^{(M)}/12)^\nu}{\lambda_{c,s}^{(M)}} \right\}_{(c,s)=(m,l),(M-m,M-l),(m,M-l),(M-m,l)} \right) \ . \end{aligned}$$

To assess $\widetilde{\alpha}_\nu$ via (4.41) and to evaluate the smoothing factor μ, we have to determine the smallest eigenvalue of the block (m,l). We restrict ourself to $l, m \le M/2$, for which it is given by $\lambda_{m,l} = \Lambda_{m,l}(1,1)$. The results are extended to other couples (m,l), noting that the blocks (l,m), $(M-l,m)$, $(l,M-m)$ and $(M-l,M-m)$ lead to the same set of eigenvalues for $\Lambda_{m,l}$ with same corresponding entries of the prolongation vector

ν	$\mu^{2\nu}$	$\rho(E_{TG}^{(J)})$	$\rho(E_{MG}^{(J)})$	$1-\delta^{(\nu)}$	$\frac{{\delta^{(1)}}^{-1}-1}{{\delta^{(1)}}^{-1}-1+\nu}$	$\frac{\mu^2+\alpha}{\mu^2+\alpha+\nu(1-\mu^2)}$
1	0.25	0.25	0.271	0.333	0.333	0.625
2	0.0625	0.083	0.121	0.2	0.2	0.455

Table 4.2: The estimates of different terms involved in inequalities (4.33) for $\nu = 1, 2$. Two-grid and V-cycle convergence factors are assessed considering $J = 7$ and $M_0 = 2$.

$\mathbf{p}_{m,l}$, and hence with same contributions to μ and $\widetilde{\alpha}_1$. Transcribing the definitions (4.9), (4.28) and (4.30) for $l, m \le M/2$, we have

$$\mu = \max_{1\le m,l\le M/2} \max_{\substack{i=1,\dots,4 \text{ if } l=M/2 \text{ or } m=M/2 \\ i=2,\dots,4 \text{ if } l,m<M/2}} \left|1-\frac{\Lambda_{m,l}(i,i)}{12}\right| , \tag{4.59}$$

$$\widetilde{\alpha}_\nu = \max_{1\le m,l\le M/2-1} \frac{\sum_{i=2}^4(\mathbf{p}_{m,l})_i^2\Lambda_{m,l}(i,i)^2\Sigma_{m,l}^{(\nu)}(i,i)}{(\Lambda_{m,l}(1,1)\Sigma_{m,l}^{(\nu)}(1,1))^2\sum_{i=1}^4(\mathbf{p}_{m,l})_i^2(\Sigma_{m,l}^{(\nu)}(i,i))^{-1}} , \tag{4.60}$$

$$\tag{4.61}$$

From (4.59) one finds

$$\mu \le \frac{1}{2} .$$

On the other hand, Theorem 4.3 applies, yielding

$$\widetilde{\alpha}_\nu \;\le\; \alpha , \tag{4.62}$$

with

$$\alpha = \max_{m,l\le M/2-1} \frac{\sum_{i=2}^4(\mathbf{p}_{m,l})_i^2\Lambda_{m,l}(i,i)^2}{\Lambda_{m,l}(1,1)^2\sum_{i=1}^4(\mathbf{p}_{m,l})_i^2} , \tag{4.63}$$

We show in Appendix A that

$$\alpha \;\le 1 , \tag{4.64}$$

whereas we show in Chapter 2 that, for this model problem,

$$ {\delta^{(\nu)}}^{-1} \le 1+\frac{1}{2\nu} , \qquad \nu = 1, 2.$$

We are then able to report in Table 4.2 all quantities involved in inequalities (4.33), using numerically computed values for $\rho(E_{TG}^{(J)})$ and $\rho(E_{MG}^{(J)})$, and approximating $\widetilde{\alpha}_\nu$ with its upper bound 1 (from (4.62), (4.64)); numerical investigation reveal that the latter is relatively accurate since one has effectively $\widetilde{\alpha}_1 = 1$ whereas, for $\nu > 1$, $\widetilde{\alpha}_\nu$ is apparently bounded below by 0.75. As a consequence, the rightmost upper bound (4.33) is in this example sharper than the rightmost upper bound (4.32), and we display only the former.

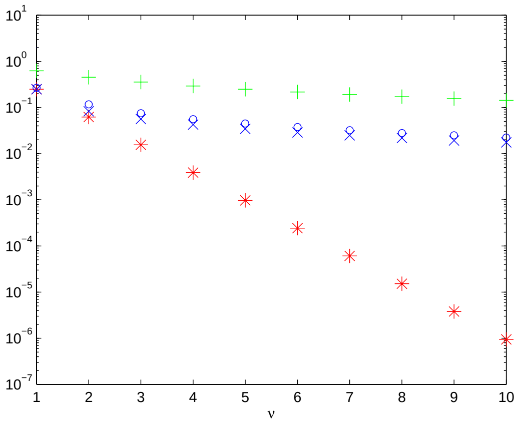


FIGURE 4.1: The dependence of $\rho(E_{MG}^{(J)})$ ($\circ$) and $\rho(E_{TG})$ ($\times$), as well as leftmost lower bound (4.33) ($*$) and rightmost upper bound (4.33) ($+$) , on the number ν of smoothing steps.

The dependence of these quantities with respect to ν is further investigated on Figure 4.1. One sees that the $\mathcal{O}(\nu^{-1})$ behavior of the upper bound (4.33) provides a more realistic estimate than the lower bound $\mu^{2\nu}$ based on the smoothing factor only.

4.7 Conclusion

We have presented two cascades of inequalities (4.32) and (4.33) witch determine intervals containing simultaneously the bound of McCormick, the V-cycle multigrid convergence factor and the two-grid convergence factor on the finest level. The intervals' limits depend on $\mu^{(k)}$, which usually coincides with the smoothing factor on level k, and on the additional quantity $\widetilde{\alpha}_{\nu}^{(k)}$, which can be further bounded by a more simple parameter $\alpha^{(k)}$. This latter parameter depends essentially on the quotient of the sine of the angle ϕ between prolongation and the corresponding smooth eigenvalue of A_k; if it is small, the aforementioned intervals are narrow, indicating that the two-grid and V-cycle convergence factors are both well reflected by the smoothing factor.

Assuming $\mu^{(k)}$ reasonably small, we have further shown that the two-grid convergence is optimal *if and only if* the cosine of ϕ has a bound proportional to the square root of the corresponding eigenvalue of A_k, whereas McCormick bound predicts an optimal V-cycle convergence *if and only if* cosine of ϕ has a bound proportional to the eigenvalue itself. Using this observation, we have clarified the heuristics in the multigrid literature which use such proportionality as a guideline in the design of multigrid solvers.

Finally, we have extended our analysis to positive semi-definite systems, as can arise when using local Fourier analysis. In particular, if the system is compatible, we have shown that the kernel modes of the problem can be ignored in the multilevel setting.

Appendix A

Here we sketch the proof of (4.64), or, equivalently, of

$$\sum_{i=2}^{4} (\mathbf{p}_{m,l})_i^2 \, \Lambda_{m,l}(i,i)^2 - \Lambda_{m,l}(1,1)^2 \sum_{i=1}^{4} (\mathbf{p}_{m,l})_i^2 \leq 0 \,, \qquad 1 \leq l, m \leq M/2-1 \,. \tag{4.65}$$

Expressing (4.65) with respect to s_l and s_m one may check (using, for instance, computer algebra tools) that

$$\begin{aligned}
&\frac{1}{256}\left(\sum_{i=2}^{4} (\mathbf{p}_{m,l})_i^2 \, \Lambda_{m,l}(i,i)^2 - \Lambda_{m,l}(1,1)^2 \sum_{i=1}^{4} (\mathbf{p}_{m,l})_i^2\right) \\
&\quad = -18s_ls_m - 172s_l^2s_m^2 + 54(s_l^2s_m + s_ls_m^2) - 9(s_l^4 + s_m^4) + 194(s_l^3s_m^2 + s_l^2s_m^3) - 250s_l^3s_m^3 \\
&\qquad -89(s_l^4s_m^2 + s_l^2s_m^4) + 88(s_l^4s_m^3 + s_l^3s_m^4) - 16s_l^4s_m^4 - 54(s_l^3s_m + s_m^3s_l) + 42(s_l^4s_m + s_m^4s_l) \\
&\quad = -s_l^4\left((3-7s_m)^2 + 36s_m^2(1-2s_m) + 4s_m^2(1-2s_m)^2\right) \\
&\qquad -s_m^4\left((3-7s_l)^2 + 36s_l^2(1-2s_l) + v4s_l^2(1-2s_l)^2\right) \\
&\qquad -2s_ls_m(9 + 86s_ls_m - 27(s_l+s_m) + 27(s_l^2+s_m^2) - 97(s_l^2s_m + s_ls_m^2) + 125s_l^2s_m^2 - 8s_l^3s_m^3)
\end{aligned}$$

and the proof is done if we show that, for $0 \leq s_m, s_l \leq 1/2$, one has

$$g(s_l, s_m) = 9 + 86s_ls_m - 27(s_l+s_m) + 27(s_l^2+s_m^2) - 97(s_l^2s_m + s_ls_m^2) + 125s_l^2s_m^2 - 8s_l^3s_m^3 \geq 0 \,.$$

This inequality follows from

$$g(s_l, s_m) = f(s_l, s_m) + \frac{1}{2}s_ls_m + \frac{97}{2}s_ls_m(1-2s_l)(1-2s_m) + 2s_l^2s_m^2(1-4s_ls_m)$$

provided that

$$f(s_l, s_m) = (27 - 72s_l^2)s_m^2 + (37s_l - 27)s_m + 9 + 27s_l^2 - 27s_l \geq 0 \,. \tag{4.66}$$

Since the discriminant of this quadratic equation, namely

$$D(s_l) = 7668s_l^4 - 7668s_l^3 + 1009s_l^2 + 918s_l - 243 \,,$$

is negative for $0 \leq s_l \leq 1/2$, and since in the same equation the factor $27 - 72s_l^2$ before s_m^2 is positive for the same interval, the inequality (4.66) and, hence, the result (4.65), follow.

Chapter 5

Algebraic analysis of aggregation-based multigrid

Summary

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 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.

5.1 Introduction

We consider multigrid methods [61, 27, 67] for solving large sparse $n \times n$ linear systems

$$A\mathbf{x} = \mathbf{b} \tag{5.1}$$

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 these 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.

Aggregation schemes are not new and trace back to [11, 20]. They did not receive much attention till recently because of the difficulty to obtain grid independent convergence on their basis [59, p.p. 522–524], see also [69, p. 663], where an accurate three grid analysis is presented for the model Poisson problem. This may be related to the fact that piecewise constant prolongation does not correspond to an interpolation which is at least first order accurate, as required by the standard multigrid theory [27, 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 [63, 64]. This allows to develop an appropriate convergence theory, but, at the same time, some of the attractive features of pure (unsmoothed) aggregation are lost, since the coarse grid matrices are less sparse and more costly to compute.

In this chapter, we investigate such pure aggregation schemes based on piecewise constant prolongation. They may indeed lead to two-grid methods with grid independent convergence properties, as recently shown in [41] for model constant coefficient discrete 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 [61]. 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 [41, 48]. However, this may be cured using more sophisticated K-cycles, in which Krylov subspace acceleration is used at each level [49]. It is also possible to improve scalability by increasing the number of smoothing steps on coarser levels [32].

Now, the (Fourier) analysis developed in [41] only addresses constant coefficient problems with artificial (periodic) boundary conditions. Although there are numerical evidences that aggregation based methods can be robust in presence of varying or discontinuous coefficients [48], this remains yet to be proved. On the other hand, it is also lacking an analysis which 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 classical AMG methods may be derived from the objective to keep reasonably bounded some convergence measure of the resulting two-grid scheme [15, 53, 58, 59].

In this chapter, we fill these gaps by developing a convergence analysis which 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. Further, 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 do not only develop a qualitative analysis, but also a quantitative one, complementary to 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 5.4 do not need the assumption that the underlying domain is convex, and, in fact, allow re-entering corners.

The presented results share some features with the analysis of element-based algebraic multigrid (AMGe) approaches, as developed in [18, 21, 31, 33]. 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.

The remainder of this chapter is organized as follows. The general framework of aggregation-based two-grid methods is introduced in Section 5.2. The algebraic analysis is developed in Section 5.3, and illustrated in Sections 5.4 and 5.5 on PDE problems with, respectively, continuous and discontinuous coefficients. Concluding remarks are given in Section 5.6.

5.2 Aggregation-based two-grid schemes

The coarsening procedure is based on the agglomeration of the unknowns of the system (5.1) into n_c non-empty disjoint sets called *aggregates*. The size of the k-th aggregate is denoted by $n^{(k)} > 0$. Note that some aggregation procedures (e.g., [48]) 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 theoretical support to this approach. Therefore, besides the n_c regular aggregates we define the (pseudo) 0-th 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 k-th aggregate, $k = 0, ..., n_c - 1$.

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

$$(P)_{ij} = \begin{cases} 1 & \text{if } i \text{ belongs to } j\text{-th aggregate} \ , \quad j = 1, ..., n_c \\ 0 & \text{otherwise} \, . \end{cases} \tag{5.2}$$

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

$$P = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n^{(1)}} & & & \\ & \mathbf{1}_{n^{(2)}} & & \\ & & \ddots & \\ & & & \mathbf{1}_{n^{(n_c)}} \end{pmatrix} . \tag{5.3}$$

In what follows, we assume a slightly more general form of (5.3)

$$P = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{p}^{(1)} & & & \\ & \mathbf{p}^{(2)} & & \\ & & \ddots & \\ & & & \mathbf{p}^{(n_c)} \end{pmatrix} \tag{5.4}$$

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 often the best (or even the only reasonable) choice.

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 $A_c = P^T A P$. In order to complete the definition of a two-grid scheme, one also needs to specify the pre- and post-smoother matrices R_1, R_2, as well as the number 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

$$E_{TG} = (I - R_2^{-1} A)^{\nu_2} (I - P^T A_c^{-1} P A)(I - R_1^{-1} A)^{\nu_1} . \tag{5.5}$$

The main objective of this chapter is the analysis of its spectral radius $\rho\,(E_{TG})$ (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

$$I - X^{-1}A = (I - R_1^{-1}A)^{\nu_1}(I - R_2^{-1}A)^{\nu_2} . \tag{5.6}$$

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, which does not necessarily requires the symmetry of R_1 and R_2.

5.3 Algebraic analysis

The starting point of our analysis is a notorious identity for the two-grid convergence rate introduced in [23, Theorem 4.3]. We recall it up to a slight generalization in Theorem 5.1 below. The generalization, that is based on the results in [44], allows for nonsymmetric smoothing scheme, e.g., $\nu_1 = 1$ and $\nu_2 = 0$. It is somehow important because the parameter μ_D for $D = \text{diag}(A)$, which is investigated in the remainder of this chapter, appears then directly connected to the convergence factor of a simplified two-grid scheme with only 1 pre- or post-smoothing step.

Theorem 5.1. *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 R_1, ν_1 and R_2, ν_2 be such that X, defined by* (5.6)*, is an $n \times n$ SPD matrix and let E_{TG} be the two-grid iteration matrix defined by* (5.5)*.*

Then, setting $\pi_X = P(P^TXP)^{-1}P^TX$, we have

$$\rho(E_{TG}) = \max\left(\lambda_{max}(X^{-1}A) - 1, 1 - \frac{1}{\mu_X}\right) , \tag{5.7}$$

where

$$\mu_X = \max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^TX(I - \pi_X)\mathbf{v}}{\mathbf{v}^TA\mathbf{v}} .$$

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

$$\mu_D = \max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^TD(I - \pi_D)\mathbf{v}}{\mathbf{v}^TA\mathbf{v}} ,$$

there holds

$$\mu_X \le \left(\max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^TX\mathbf{v}}{\mathbf{v}^TD\mathbf{v}}\right) \mu_D . \tag{5.8}$$

In particular, if $R_1 = R_2 = \omega^{-1}D$ with $\omega^{-1} \ge \lambda_{\max}(D^{-1}A)$, one has

$$\rho(E_{TG}) = 1 - \frac{1}{\mu_X} , \tag{5.9}$$

with

$$\mu_X \le \omega^{-1}\,\mu_D\ , \tag{5.10}$$

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

Proof. The equality (5.7) is a direct consequence of [44, Theorem 2.1 and Corollary 2.1], combined with the assumptions that A and X are SPD, which implies

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

The inequality (5.8) follows from Corollary 2.2 in [44], setting in this latter $Y = D$, $L_Y = D^{1/2}$ and $Q = \pi_D$.

To prove (5.9), observe that $\omega^{-1} \ge \lambda_{\max}(D^{-1}A)$ implies, together with (5.6), $\lambda_{\max}(X^{-1}A) \le 1$. Hence (5.7) gives (5.9) since it is known by [44, Theorem 2.1] that $\mu_X \ge 1$.

The inequality (5.10) follows from (5.8) combined with

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

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 (5.10) is an equality. ■

When D is chosen independently of P, the first factor in the right hand side of (5.8) depends only on the smoothing scheme. If $R_1 = R_2^T = R$ and $\nu_1 = \nu_2 = \nu$, setting $S = \left(I - R^{-1}A\right)^{\nu}$, one has further

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

Hence, when $D = \operatorname{diag}(A)$ (the choice that is privileged in the rest of this work) this quantity is nothing but the inverse of the *smoothing factor* in Ruge-Stüben analysis [59]. On the other hand, the second factor in the right hand side of (5.8) depends on P but not on X, and keeping it bounded amounts to satisfying an *approximation property.*

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

$$A = A_b + A_r\,, \tag{5.11}$$

where A_b and A_r are both symmetric nonnegative definite and A_b is block diagonal:

$$A_b = \begin{pmatrix} A^{(0)} & & & \\ & A^{(1)} & & \\ & & \ddots & \\ & & & A^{(n_c)} \end{pmatrix} , \tag{5.12}$$

where $A^{(k)}$, $k = 0, ..., n_c$, 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 nonpositive, the matrix is an M-matrix). The matrices $A^{(k)}$, $k = 0, ..., n_c$ can be constructed by restricting the matrix A to the unknowns belonging to the k-th aggregate and then by subtracting the corresponding contribution $C^{(k)} = \text{diag}(c_i)$ from its diagonal, in order to keep

$$A_r = \begin{pmatrix} C^{(0)} & * & \cdots & * \\ * & C^{(1)} & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & C^{(n_c)} \end{pmatrix} \tag{5.13}$$

diagonally dominant, and, hence, nonnegative definite. Since A is diagonally dominant, the contribution subtracted from the diagonal of each $A^{(k)}$ can be such that either each row of A_b is weakly diagonally dominant; that is

$$(A_b)_{jj} - \sum_{\substack{i=1 \\ i \neq j}}^{n} |(A_b)_{ij}| = 0, \;\; j = 1, ..., n\,; \tag{5.14}$$

or such that each row of A_r is weakly diagonally dominant; that is

$$(A_r)_{jj} - \sum_{\substack{i=1 \\ i \neq j}}^{n} |(A_r)_{ij}| = 0, \;\; j = 1, ..., n\,; \tag{5.15}$$

or something in between; that is

$$(|(A)_{jj}| - \sum_{\substack{i=1 \\ i \neq j}}^{n} |(A)_{ij}|) + \sum_{\substack{i=1 \\ i \neq j}}^{n} |(A_b)_{ij}| \geq (A_b)_{jj} \geq \sum_{\substack{i=1 \\ i \neq j}}^{n} |(A_b)_{ij}|, \;\; j = 1, ..., n\,. \tag{5.16}$$

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

Theorem 5.2. *Let $A = A_b + A_r$ be an $n \times n$ SPD matrix, with A_b and A_r symmetric nonnegative definite and A_b having the block-diagonal form* (5.12)*. Let P be an $n \times n_c$ matrix of rank $n_c < n$ and of the form* (5.4)*. Let*

$$D = \begin{pmatrix} D^{(0)} & & & \\ & D^{(1)} & & \\ & & \ddots & \\ & & & D^{(n_c)} \end{pmatrix} \tag{5.17}$$

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

$$\mu_D = \max_{\mathbf{v} \in \mathbb{R}^n \backslash \{\mathbf{0}\}} \frac{\mathbf{v}^T D (I - \pi_D) \mathbf{v}}{\mathbf{v}^T A \mathbf{v}} \,. \tag{5.18}$$

Letting

$$\mu_D^{(0)} = \begin{cases} 0 & \textit{if } n^{(0)} = 0 \\ \displaystyle\sup_{\mathbf{v^{(0)}} \in \mathbb{R}^{n^{(0)}} \backslash \mathcal{N}(A^{(0)})} \frac{{\mathbf{v}^{(0)}}^T D^{(0)} \mathbf{v}^{(0)}}{{\mathbf{v}^{(0)}}^T A^{(0)} \mathbf{v}^{(0)}} & \textit{if } n^{(0)} > 0 \end{cases}$$

and, for $k = 1, ..., n_c$,

$$\mu_D^{(k)} = \begin{cases} 0 & \textit{if } n^{(k)} = 1 \\ \displaystyle\sup_{\mathbf{v^{(k)}} \in \mathbb{R}^{n^{(k)}} \backslash \mathcal{N}(A^{(k)})} \frac{{\mathbf{v}^{(k)}}^T D^{(k)} (I - \pi_D^{(k)}) \mathbf{v}^{(k)}}{{\mathbf{v}^{(k)}}^T A^{(k)} \mathbf{v}^{(k)}} & \textit{if } n^{(k)} > 1 \,, \end{cases} \tag{5.19}$$

where

$$\pi_D^{(k)} = \mathbf{p}^{(k)} ({\mathbf{p}^{(k)}}^T D^{(k)} \mathbf{p}^{(k)})^{-1} \, {\mathbf{p}^{(k)}}^T D^{(k)} \,, \tag{5.20}$$

there holds

$$\mu_D \le \max_{k=0,...,n_c} \mu_D^{(k)} \,. \tag{5.21}$$

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

$$\mu_D^{(k)} = \begin{cases} 0 & \textit{if } n^{(k)} = 1 \\ \displaystyle\max_{\mathbf{v^{(k)}} \in \mathcal{R}(A^{(k)}) \backslash \{\mathbf{0}\}} \frac{{\mathbf{v}^{(k)}}^T D^{(k)} (I - \pi_D^{(k)}) \mathbf{v}^{(k)}}{{\mathbf{v}^{(k)}}^T A^{(k)} \mathbf{v}^{(k)}} & \textit{if } n^{(k)} > 1 \,. \end{cases} \tag{5.22}$$

Proof. We first prove the *if and only if* result for $k = 1, ..., n_c$, the case $k = 0$ being trivial. The *if* statement assumes $\mathcal{N}(A^{(k)}) \subset \operatorname{span}\{\mathbf{p^{(k)}}\}$ 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 (5.19) becomes a maximum over $\mathbb{R}^{n^{(k)}} \backslash \{\mathbf{0}\} = \mathcal{R}(A^{(k)}) \backslash \{\mathbf{0}\}$, hence, (5.22) 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 $D^{(k)}(I-\pi_D^{(k)})\mathbf{p}^{(k)} = (D^{(k)}(I-\pi_D^{(k)}))^T\mathbf{p}^{(k)} = \mathbf{0}$, we have

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

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

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

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

We now prove (5.21). 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 $k = 0, ..., n_c$. Moreover, since $n_c < n$, there holds $\mu_D > 0$.

Now, observe that

$$D(I-\pi_D) = \begin{pmatrix} D^{(0)} & & & \\ & D^{(1)}(I-\pi_D^{(1)}) & & \\ & & \ddots & \\ & & & D^{(n_c)}(I-\pi_D^{(n_c)}) \end{pmatrix} \tag{5.23}$$

and, hence,

$$\begin{aligned} \mu_D &= \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^T D(I-\pi_D)\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \\ &= \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^T D(I-\pi_D)\mathbf{v}}{\mathbf{v}^T A_b\mathbf{v} + \mathbf{v}^T A_r\mathbf{v}} \\ &= \max_{\mathbf{v}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\sum_{k=1,\dots,n_c} {\mathbf{v}^{(k)}}^T D^{(k)}(I-\pi_D^{(k)})\mathbf{v}^{(k)} + {\mathbf{v}^{(0)}}^T D^{(0)}\mathbf{v}^{(0)}}{\sum_{k=0,\dots,n_c} {\mathbf{v}^{(k)}}^T A^{(k)}\mathbf{v}^{(k)} + \mathbf{v}^T A_r\mathbf{v}}\,. \end{aligned} \tag{5.24}$$

Let $\mathbf{v}_* = ({\mathbf{v}_*^{(0)}}^T\ {\mathbf{v}_*^{(1)}}^T \cdots {\mathbf{v}_*^{(n_c)}}^T)^T$ be the vector that realizes this maximum. Notice that $\sum_{k=0,\dots,n_c} {\mathbf{v}_*^{(k)}}^T A^{(k)}\mathbf{v}_*^{(k)} > 0$. Indeed, because of the boundness of $\mu_D^{(k)}$, $k = 0, ..., n_c$, the equality $\sum_{k=0,\dots,n_c} {\mathbf{v}_*^{(k)}}^T A^{(k)}\mathbf{v}_*^{(k)} = 0$ would imply a zero numerator in the right hand side of (5.24), whereas, since A is SPD, $\mathbf{v}_*^T A_r\mathbf{v}_* > 0$, which would further lead to $\mu_D = 0$. This latter contradicts our assumption $n_c < n$.

Next, since μ_D is finite, as shown above, $\mathbf{v}_*^{(k)} \in \mathcal{N}(A^{(k)})$ implies

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

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

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

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) 0-th aggregate. The proposition below presents a simple estimate of the pseudo aggregate's quality based on diagonal dominance excess of corresponding rows.

Proposition 5.1. *Assume that A is diagonally dominant, that the splitting $A = A_b + A_r$ satisfies* (5.15) *for $j = 1, ..., n_c$ and that $D^{(0)} = \operatorname{diag}\{(A)_{ii} | i = 1, ..., n^{(0)}\}$. If $n^{(0)} > 0$, one has*

$$\mu_D^{(0)} = \max_{\mathbf{v} \in \mathbb{R}^{n^{(0)}}} \frac{\mathbf{v}^T D^{(0)} \mathbf{v}}{\mathbf{v}^T A^{(0)} \mathbf{v}} \leq \max_{i=1,\dots,n_c} \frac{(A)_{ii}}{2(A)_{ii} - \sum_{j=1}^n |(A)_{ij}|} . \tag{5.25}$$

Proof. Set $\eta_i = 2(A)_{ii} - \sum_{j=1}^n |(A)_{ij}|$ and note that if $\eta_i = 0$ at least for one $i \leq n_c$, the inequality is trivially satisfied. Otherwise, observing that $A^{(0)} \geq \operatorname{diag}(\eta_i)$, the inequality (5.25) follows. $\blacksquare$

Regarding aggregates $1, ..., n_c$, 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 if $A^{(k)}$ and $D^{(k)}$ are given, and determine the $\mathbf{p}^{(k)}$ that leads to the best quality.

Theorem 5.3. *Let $A^{(k)}$ and $D^{(k)}$ be, respectively, an $n^{(k)} \times n^{(k)}$ non-zero symmetric nonnegative definite matrix and an $n^{(k)} \times n^{(k)}$ SPD matrix, with $n^{(k)} > 1$. Let $\mathbf{p}^{(k)}$ be a non-zero vector of size $n^{(k)}$. Let*

$$\mu_D^{(k)} = \sup_{\mathbf{v} \in \mathbb{R}^{n^{(k)}} \setminus \mathcal{N}(A^{(k)})} \frac{\mathbf{v}^T D^{(k)} (I - \pi_D^{(k)}) \mathbf{v}}{\mathbf{v}^T A^{(k)} \mathbf{v}} , \tag{5.26}$$

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

$$\lambda_2^{-1} \leq \mu_D^{(k)} \leq \lambda_1^{-1} . \tag{5.27}$$

Moreover, if

$$D^{(k)\,-1}\,A^{(k)}\mathbf{p} = \lambda_1\,\mathbf{p}\,, \tag{5.28}$$

then

$$\mu_D^{(k)} = \frac{1}{\lambda_2}\,, \tag{5.29}$$

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

$$\mathbf{v}^T D^{(k)}(I-\pi_D^{(k)})\mathbf{v} = \mu_D^{(k)}\mathbf{v}^T A^{(k)}\mathbf{v} \qquad \mathbf{v}\in\mathcal{R}(A^{(k)})$$

if and only if

$$D^{(k)\,-1}\,A^{(k)}\mathbf{v} = \lambda_2\mathbf{v} \quad \textit{with} \quad \mathbf{v}^T D^{(k)}\mathbf{p} = 0\,. \tag{5.30}$$

Proof. Note that the case $\mu_D^{(k)} = \infty$ implies nonempty $\mathcal{N}(A^{(k)})$ and, hence, $\lambda_1 = 0$. The inequalities (5.27) are then trivially satisfied. Moreover, according to Theorem 5.2 we have then $\mathcal{N}(A^{(k)}) \not\subseteq \operatorname{span}\{\mathbf{p}\}$. Hence, if (5.28) holds, $\dim(\mathcal{N}(A^{(k)})) \geq 2$, which in turn implies $\lambda_2 = 0$ and, therefore, (5.29).

Now, consider $\mu_D^{(k)} < \infty$ which, according to Theorem 5.2, implies $\mathcal{N}(A^{(k)}) \subset \operatorname{span}\{\mathbf{p}\}$, and, hence, $\lambda_2 > 0$. If $\mathcal{N}(A^{(k)})$ is nonempty, then $\mathcal{N}(A^{(k)}) = \operatorname{span}\{\mathbf{p}\}$ and $\lambda_1 = 0$, which in turn implies (5.28). Therefore, for all $\mathbf{v}\in\mathcal{R}(A^{(k)})$, $\mathbf{p}^T D^{(k)}\mathbf{v} = 0$, and, hence, $\pi_D^{(k)}\mathbf{v} = \mathbf{0}$. Then, according to the Theorem 5.2, we further have

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

In addition, a vector $\mathbf{v}$ reaches the maximum if and only if (5.30) 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 (5.27) we set

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

and note that $\pi_D^{(k)}\mathbf{v} = \mathbf{0}$. Injecting such $\mathbf{v}\neq\mathbf{0}$ into (5.26) we find

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

The right inequality (5.27) follows from

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

Moreover, if $\mathbf{p} = \mathbf{x}_1$, then $\mathbf{x}_i$, $i = 1, ..., n^{(k)}$ 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 ${A^{(k)}}^{-1} D^{(k)}(I - \pi_D^{(k)})$, (5.29) follows. Moreover, (5.30) 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 $\mathbf{v}^T D^{(k)}(I - \pi_D^{(k)})\mathbf{v} = \mu_D^{(k)} \mathbf{v}^T A^{(k)} \mathbf{v}$. ■

By way of illustration, consider a symmetric diagonally dominant M-matrix and assume that the splitting $A = A_b + A_r$ is based on the rule (5.14). Then, each $A^{(k)}$ is singular with its null space equal to span$\{\mathbf{1}_{n^{(k)}}\}$. Theorem 5.2 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 5.3, $\mu_D^{(k)} = \lambda_2 ({D^{(k)}}^{-1} A^{(k)})^{-1}$. When the diagonal dominance is strict, the two side inequality (5.16) 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 best. The following remarks give some clues in this respect.

Remark 5.3.1 When $A^{(k)}$ is irreducible and diagonally dominant with nonpositive 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 all positive (e.g., $\mathbf{1}_{n^{(k)}}$) is necessarily the eigenvector associated with the smallest eigenvalue, which is unique.

Remark 5.3.2 Consider a diagonally dominant M-matrix for which the splitting $A = A_b + A_r$ is based on (5.16). 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 as large as possible while satisfying (5.16) with the additional constraint that $\mathbf{1}_{n^{(k)}}$ is an eigenvector of ${D^{(k)}}^{-1} A^{(k)}$, so that the condition ensuring (5.29) holds. In particular, when $D^{(k)}$ is a diagonal matrix, it amounts to using $A^{(k)} = A_0^{(k)} + \eta D^{(k)}$ with $A_0^{(k)}$ satisfying (5.14) and with η being the largest constant such that (5.16) still holds.

5.4 Discrete PDEs with constant and smoothly varying coefficients

5.4.1 Preliminaries

We start considering matrices associated with the five point stencil

$$\begin{bmatrix} & -\alpha_y & \\ -\alpha_x & \alpha_d & -\alpha_x \\ & -\alpha_y & \end{bmatrix} \quad \text{with } \alpha_x, \alpha_y > 0 \text{ and } \alpha_d \geq 2(\alpha_x + \alpha_y) \tag{5.31}$$

on a rectangular grid of arbitrary shape. For such matrices we want to assess boxwise aggregates with four nodes per aggregate (as on Figure 5.1(a)) and linewise aggregates

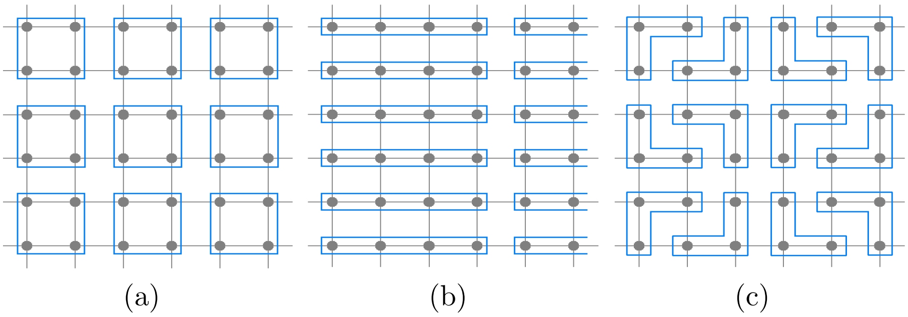


FIGURE 5.1: Examples of (a) boxwise, (b) linewise and (c) L-shaped aggregation patterns.

with two, three and four nodes (as on Figure 5.1(b)). We select the splitting $A = A_b + A_r$ satisfying (5.15). The prolongation vector is $\mathbf{p}^{(k)} = \mathbf{1}_{n^{(k)}}$, $k = 1, ..., n_c$ and, as can be checked from (5.32) and (5.34) below, it is an eigenvector of $D^{(k)^{-1}} A^{(k)}$ associated with the smallest eigenvalue $\delta_d \alpha_d^{-1}$, where $\delta_d = \alpha_d - 2(\alpha_x + \alpha_y) \geq 0$. Theorem 5.3 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

$$A^{(k)} = \begin{pmatrix} \alpha_x + \alpha_y & -\alpha_x & -\alpha_y & 0 \\ -\alpha_x & \alpha_x + \alpha_y & 0 & -\alpha_y \\ -\alpha_y & 0 & \alpha_x + \alpha_y & -\alpha_x \\ 0 & -\alpha_y & -\alpha_x & \alpha_x + \alpha_y \end{pmatrix} + \delta_d I \ , \tag{5.32}$$

and, hence,

$$\mu_D^{(k)} = \frac{2\alpha_x + 2\alpha_y + \delta_d}{2\min(\alpha_x, \alpha_y) + \delta_d} \, , \tag{5.33}$$

whereas for linewise aggregation of size m in x direction

$$A^{(k)} = \begin{pmatrix} \alpha_x & -\alpha_x & & \\ -\alpha_x & 2\alpha_x & \ddots & \\ & \ddots & \ddots & -\alpha_x \\ & & -\alpha_x & \alpha_x \end{pmatrix} + \delta_d I \tag{5.34}$$

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

$$\mu_D^{(k)} = \frac{2\alpha_x + 2\alpha_y + \delta_d}{(2 - \sqrt{m-2})\alpha_x + \delta_d} \, . \tag{5.35}$$

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

$$\mu_D^{(k)} \leq 3 + \sqrt{2} \, . \tag{5.36}$$

Since linewise aggregates of size 3 and 2 have better quality than linewise aggregates of size 4, as can be concluded from (5.35), this upper bound holds for them as well.

5.4.2 Constant coefficients

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

$$\frac{\partial}{\partial x}\left(\alpha_x \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\alpha_y \frac{\partial u}{\partial y}\right) + \beta u = f \quad \text{on } \Omega\,, \tag{5.37}$$

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 (5.36) 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 $\alpha_x = \alpha_y$, (5.35) with $m = 3$ also applies to L-shaped aggregates as illustrated on Figure 5.1(c).

Regarding Neumann boundary conditions, only the quality of aggregates that contain boundary nodes is not covered by the above analysis. Again, however, isotropic coefficients and linewise aggregates aligned with strong coupling yield bounds similar to (5.33) and (5.35). 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 Lemma 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 lemma, one has then $\mu_D^{(k)} \leq 2$ in the two former cases and $\mu_D^{(k)} \leq 2.23$ 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 [48] 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.

5.4.3 Smoothly varying coefficients

Consider now the same discrete PDE (5.37) 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

	$\alpha_x = \alpha_y\,,\ \ \delta_d = 0$						$\alpha_x = 10\alpha_y\,,\ \ \delta_d = 0$					
	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

TABLE 5.1: The value of μ_D and of its upper bound (5.21) for different grid sizes.

because $\beta = 0$ and, hence, $\mathcal{N}(A^{(k)}) = \text{span}\{\mathbf{1}_{n^{(k)}}\}$, or by using the trick suggested at the end of Section 5.3 in Remark 5.3.2 (see also Remark 5.3.1). Hence, as shown in Theorem 5.3, $\mu_D^{(k)}$ is the inverse of the second smallest eigenvalue of ${D^{(k)}}^{-1} A^{(k)}$. Since the eigenvalues of a matrix are continuous functions 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.

5.4.4 Numerical example

We consider the linear system resulting from the five point finite difference discretization of (5.37) 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 (5.14), the matrices $A^{(k)}$ and $D^{(k)}$ are the same for all aggregates. As a consequence, the quality estimate $\mu_D^{(k)}$ is the same as well.

We consider first an isotropic situation ($\alpha_x = \alpha_y$). The columns from 2 to 7 of Table 5.1 then give the values of μ_D and of its upper bound $\mu_D^{(k)}$ for three types of aggregation pattern, presented on Figure 5.1. Observe that when mode 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.

5.4.5 Sharpness of the estimate

Numerical results in Table 5.1 indicate that the bound (5.21) on μ_D can be asymptotically sharp for N large enough. Moreover, as shown in Theorem 5.1, if only one Jacobi smoothing iteration is performed, we further have $\rho(E_{TG}) = 1 - \omega\mu_D^{-1}$. 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. Without loss of generality, we assume in addition that $\alpha_x \geq \alpha_y$. 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 $\mathbf{v}_b = (1\,1\, -1\, -1)^T \in \mathcal{R}(A^{(k)})$ 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 5.3, the supremum in definition (5.19) of $\mu_D^{(k)}$. Therefore, setting

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

we locally reproduce the maximizing vectors for every aggregate. Moreover, setting $\gamma_1 = \gamma_2 = \cdots = \gamma_N = -\gamma_{N+1} = \cdots = -\gamma_{2N} = \gamma_{2N+1} = \cdots = 1$ 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 (5.13) 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

$$\begin{aligned}
\widetilde{\mathbf{v}} A_r \widetilde{\mathbf{v}} &= -\sum_{i,j=1}^{n} \frac{1}{2}\,(A_r)_{ij}\,((\widetilde{\mathbf{v}})_i - (\widetilde{\mathbf{v}})_j)^2 + \sum_{i=1}^{n} \sigma_i\, (\widetilde{\mathbf{v}})_i{}^2 && (5.38)\\
&= \sum_{i=1}^{n} \sigma_i\, (\widetilde{\mathbf{v}})_i{}^2 \\
&= 2N(\alpha_x + \alpha_y) \\
&= 2N^{-1}(\alpha_x + \alpha_y)\alpha_d^{-1} \widetilde{\mathbf{v}}^T D \widetilde{\mathbf{v}}\,. && (5.39)
\end{aligned}$$

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

$$\begin{aligned}
\mu_D &\geq \frac{\widetilde{\mathbf{v}}^T D (I - \pi_D)\widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^T A_b \widetilde{\mathbf{v}} + \widetilde{\mathbf{v}}^T A_r \widetilde{\mathbf{v}}} \\
&= \frac{\widetilde{\mathbf{v}}^T D \widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^T A_b \widetilde{\mathbf{v}} + \widetilde{\mathbf{v}}^T A_r \widetilde{\mathbf{v}}} \\
&= \frac{\widetilde{\mathbf{v}}^T D \widetilde{\mathbf{v}}}{\mu_D^{(k)}{}^{-1}\, \widetilde{\mathbf{v}}^T D \widetilde{\mathbf{v}} + \widetilde{\mathbf{v}}^T A_r \widetilde{\mathbf{v}}} \\
&= \frac{\mu_D^{(k)}}{1 + \mu_D^{(k)} \frac{\widetilde{\mathbf{v}}^T A_r \widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^T D \widetilde{\mathbf{v}}}}\,. && (5.40)
\end{aligned}$$

It then follows from (5.39) 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 5.4. *Let $A = A_b + A_r$, P, D, μ_D and $\mu_D^{(k)}$, $k = 0, ..., n_c$, be defined as in Theorem 5.2. Assume $\mu_D^{(k)}$ finite for $k = 0, ..., n_c$ and let, for $n^{(0)} > 0$ and $n^{(k)} > 1$ $k = 1, ..., n_c$,*

$$\begin{aligned} \widetilde{\mathbf{v}}_0 &\in \underset{\mathbf{v^{(0)}} \in \mathbb{R}^{n^{(0)}} \backslash \{\mathbf{0}\}}{\arg\max} \left(\frac{\mathbf{v}^{(0)\,T} D^{(0)} \mathbf{v}^{(0)}}{\mathbf{v}^{(0)\,T} A^{(0)} \mathbf{v}^{(0)}} \right) , \\ \widetilde{\mathbf{v}}_k &\in \underset{\mathbf{v^{(k)}} \in \mathcal{R}(A^{(k)}) \backslash \{\mathbf{0}\}}{\arg\max} \left(\frac{\mathbf{v}^{(k)\,T} D^{(k)} (I - \pi_D^{(k)}) \mathbf{v}^{(k)}}{\mathbf{v}^{(k)\,T} A^{(k)} \mathbf{v}^{(k)}} \right) , \end{aligned} \tag{5.41}$$

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

$$\widetilde{\mathbf{v}} = \left(\gamma_0 \theta_0^{-1} \, \widetilde{\mathbf{v}}_0{}^T \; \gamma_1 \theta_1^{-1} \, \widetilde{\mathbf{v}}_1{}^T \; \cdots \; \gamma_{n_c} \theta_{n_c}^{-1} \, \widetilde{\mathbf{v}}_{n_c}{}^T \right)^T , \tag{5.42}$$

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

$$\widetilde{\mathbf{v}}^T A_r \widetilde{\mathbf{v}} \leq \varepsilon \widetilde{\mathbf{v}}^T A_b \widetilde{\mathbf{v}} , \tag{5.43}$$

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

$$(\mathbf{c} + \widetilde{\mathbf{v}})^T A_r (\mathbf{c} + \widetilde{\mathbf{v}}) \leq \varepsilon \left(\max_{k=1,\dots,n_c} \mu_D^{(k)} \right)^{-1} \widetilde{\mathbf{v}}^T D \widetilde{\mathbf{v}} \tag{5.44}$$

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

Then

$$\mu_D \geq \frac{1}{1+\varepsilon} \frac{\sum_{k=0}^{n_c} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0}^{n_c} \gamma_k^2} . \tag{5.45}$$

Proof. We first prove the lower bound (5.45) based on the assumption (5.43). Starting with the equality (5.24) in the proof of Theorem 5.2 and setting $\mathbf{v}^{(k)} = \gamma_k \theta_k^{-1} \widetilde{\mathbf{v}}_k$ together with $\mathbf{v} = \widetilde{\mathbf{v}}$, we have

$$\begin{aligned} \mu_D &\geq \frac{\sum_{k=1,\dots,n_c} \mathbf{v}^{(k)\,T} D^{(k)} (I - \pi_D^{(k)}) \mathbf{v}^{(k)} + \mathbf{v}^{(0)\,T} D^{(0)} \mathbf{v}^{(0)}}{\sum_{k=0,\dots,n_c} \mathbf{v}^{(k)\,T} A^{(k)} \mathbf{v}^{(k)} + \mathbf{v}^T A_r \mathbf{v}} \\ &\geq \frac{1}{1+\varepsilon} \frac{\sum_{k=1,\dots,n_c} \gamma_k^2 \theta_k^{-2} \, \widetilde{\mathbf{v}}_k{}^T D^{(k)} (I - \pi_D^{(k)}) \widetilde{\mathbf{v}}_k + \gamma_0^2 \theta_0^{-2} \, \widetilde{\mathbf{v}}_0{}^T D^{(0)} \widetilde{\mathbf{v}}_0}{\sum_{k=0,\dots,n_c} \gamma_k^2 \theta_k^{-2} \, \widetilde{\mathbf{v}}_k{}^T A^{(k)} \widetilde{\mathbf{v}}_k} \\ &= \frac{1}{1+\varepsilon} \frac{\sum_{k=1,\dots,n_c} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0,\dots,n_c} \gamma_k^2} , \end{aligned} \tag{5.46}$$

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

Now, we prove the lower bound (5.45) based on the assumptions related to (5.44). We may then assume that

$$(\mathbf{c}+\widetilde{\mathbf{v}})^T A_r(\mathbf{c}+\widetilde{\mathbf{v}}) \leq \varepsilon \widetilde{\mathbf{v}}^T A_b \widetilde{\mathbf{v}} \tag{5.47}$$

holds, this inequality being proved later. Since $\mu_D^{(k)}$ is finite, Theorem 5.2 implies $\mathcal{N}(A_r^{(k)}) \subset \operatorname{span}\{\mathbf{p}^{(k)}\}$, $k = 1, ..., n_c$. From the singularity of $A^{(k)}$, $k = 1, ..., n_c$, we further conclude that $\mathcal{N}(A_r^{(k)}) = \operatorname{span}\{\mathbf{p}^{(k)}\}$ and, hence,

$$\left(\xi_k \mathbf{p}^{(k)} + \widetilde{\mathbf{v}}_k\right)^T A^{(k)} \left(\xi_k \mathbf{p}^{(k)} + \widetilde{\mathbf{v}}_k\right) = \widetilde{\mathbf{v}}_k^T A^{(k)} \widetilde{\mathbf{v}}_k\,. \tag{5.48}$$

Moreover, using the definition (5.20) of $\pi_D^{(k)}$, we also have

$$\left(\xi_k \mathbf{p}^{(k)} + \widetilde{\mathbf{v}}_k\right)^T D^{(k)}(I-\pi_D^{(k)}) \left(\xi_k \mathbf{p}^{(k)} + \widetilde{\mathbf{v}}_k\right) = \widetilde{\mathbf{v}}_k^T D^{(k)}(I-\pi_D^{(k)}) \widetilde{\mathbf{v}}_k\,. \tag{5.49}$$

Therefore, injecting $\mathbf{v} = \mathbf{c} + \widetilde{\mathbf{v}}$ and $\mathbf{v}^{(k)} = \xi_k \mathbf{p}^{(k)} + \gamma_k \theta_k^{-1} \widetilde{\mathbf{v}}_k$ into (5.46) and using (5.48) and (5.49), the proof is finished as in the previous case.

We are thus left with the proof of (5.47). From Theorem 5.3 we conclude that $D^{(k)}{}^{-1} A^{(k)} \widetilde{\mathbf{v}}_k = \lambda_2(D^{(k)}{}^{-1} A^{(k)}) \widetilde{\mathbf{v}}_k$ and $\widetilde{\mathbf{p}}^{(k)}{}^T D^{(k)} \widetilde{\mathbf{v}}_k = 0$. Therefore,

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

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

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

which, together with (5.44) implies (5.47). ■

Now, we return to the previous example and prove the asymptotical sharpness for linewise aggregates of size $m \leq 4$. As in the boxwise case, the vector (5.41) is the second eigenvector of $\alpha_d^{-1} A^{(k)}$ given by

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

Hence, choosing

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

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 (5.38) with first term in the right hand side vanishing, we have

$$\begin{aligned}\widetilde{\mathbf{v}}A_r\widetilde{\mathbf{v}} &= 2Nm^{-1}\alpha_y\|\mathbf{v}_b\|^2+\alpha_x N(\|(\mathbf{v}_b)_1\|^2+\|(\mathbf{v}_b)_m\|^2)\\ &\le 2N(\alpha_x+\alpha_y)\|\mathbf{v}_b\|^2\\ &= 2N^{-1}(\alpha_x+\alpha_y)\alpha_d^{-1}\mu_D^{(k)}\widetilde{\mathbf{v}}^TA_b\widetilde{\mathbf{v}}\,, \qquad (5.50)\end{aligned}$$

and, hence, (5.45) holds with $\varepsilon = N^{-1}(\alpha_x+\alpha_y)\alpha_d^{-1}\mu_D^{(k)}$. The asymptotical sharpness follows then from Theorem 5.4.

Considering a general situation, we note that a lower bound close to the upper bound (5.21) can be proved via (5.45) if there exists a vector $\widetilde{\mathbf{v}}$ of the form (5.42), 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 $\max_{k=0,\dots,n_c}\mu_D^{(k)}$;

(b) ε, defined via (5.43) or (5.44), 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

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

independently of the choice of γ_k. As illustrated in Section 5.5, the use of γ_k^2 with variable magnitude allows to prove 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 (5.38) and look for a vector $\widetilde{\mathbf{v}}$ of the form (5.42) such that $(A_r)_{ij}\left((\widetilde{\mathbf{v}})_i-(\widetilde{\mathbf{v}})_j\right)=0$ for all i and j. If such a vector exists, the first term in (5.38) is zero. Then, let $\Omega_h=\{1,\dots,n_c\}$ be the set of all coarse unknowns and set $\partial\Omega_h=\left\{i\,|\,\sigma_i=\sum_{j=1}^n(A_r)_{ij}\neq 0\right\}$. 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 $(\widetilde{\mathbf{v}})_i$, $i=1,\dots,n_c$ reasonably bounded, we have

$$\widetilde{\mathbf{v}}^TA_r\widetilde{\mathbf{v}}=\sum_{i=1}^n\sigma_i(\widetilde{\mathbf{v}})_i^2=\mathcal{O}(|\partial\Omega_h|)\,,$$

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

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

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

Further, the lower bound (5.45) can be obtained using only a (given) set of aggregates (numbered from 1 to $\bar{n}_c$ for convenience), setting

$$\widetilde{\mathbf{v}} = (\gamma_1\theta_1^{-1}\,\widetilde{\mathbf{v}}_1{}^T\,\cdots\,\gamma_{\bar{n}_c}\theta_{\bar{n}_c}^{-1}\,\widetilde{\mathbf{v}}_{\bar{n}_c}{}^T\quad \mathbf{0}^T\,\cdots\,\mathbf{0}^T)^T\,. \tag{5.51}$$

Then, (5.38) 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

$$\mu_D \geq \frac{1}{1+\bar{\varepsilon}}\frac{\sum_{k=0}^{\bar{n}_c}\gamma_k^2\mu_D^{(k)}}{\sum_{k=0}^{\bar{n}_c}\gamma_k^2}\,,$$

with $\bar{\varepsilon}=\mathcal{O}(|\partial\bar{\Omega}_h|/|\bar{\Omega}_h|)$. In practice, it means that the upper bound (5.21) 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 (5.31). Then, assuming that the whole square is covered with box aggregates as on the Figure 5.1(a), the relations (5.33), (5.40) and (5.39) can be used (with $\bar{N}$ instead of N) to show that

$$\mu_D \geq \frac{1}{1+\bar{\varepsilon}}\bar{\mu}_D \tag{5.52}$$

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

5.5 Discrete PDEs with discontinuous coefficients

5.5.1 Preliminaries

As in the previous section, our analysis is based on the aggregates' quality, which in turn involves the computation of the second smallest eigenvalue of small matrices. The following lemma is helpful in this respect.

Lemma 5.1. *Let*

$$A_d=\frac{1}{2}\begin{pmatrix} 4\alpha_1 & -2\alpha_1 & -2\alpha_1 & \\ -2\alpha_1 & 3\alpha_1+\alpha_2 & & -\alpha_1-\alpha_2 \\ -2\alpha_1 & & 3\alpha_1+\alpha_3 & -\alpha_1-\alpha_3 \\ & -\alpha_1-\alpha_2 & -\alpha_1-\alpha_3 & 2\alpha_1+\alpha_2+\alpha_3 \end{pmatrix} \tag{5.53}$$

and

$$D_d = \operatorname{diag}\left(4\alpha_1\ \ 2(\alpha_1{+}\alpha_2)\ \ 2(\alpha_1{+}\alpha_3)\ \ (\alpha_1{+}\alpha_2{+}\alpha_3{+}\alpha_4)\right)\,, \tag{5.54}$$

where $\alpha_1 > 0$ *and* $\alpha_2, \alpha_3, \alpha_4 > 0$. A_d *is positive semi-definite, and let* $\lambda_2(D_d^{-1}A_d)$ *be the smallest nonzero eigenvalue of* $D_d^{-1}A_d$.

Then,

$$\lambda_2(D_d^{-1}A_d) \geq \frac{5-\sqrt{17}}{8}\,, \tag{5.55}$$

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

$$\lambda_2(D_d^{-1}A_d) = \min\left(\frac{1}{2}\,, \frac{3\alpha_1+\alpha_3}{4(\alpha_1+\alpha_3)}\right)\,. \tag{5.56}$$

Moreover, if $\alpha_1 \geq \alpha_2, \alpha_3, \alpha_4$, *one has*

$$\lambda_2(D_d^{-1}A_d) \geq \beta \tag{5.57}$$

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,

$$\lambda_2(D_d^{-1}A_d) \geq \frac{1}{2} \quad \textit{if} \quad \begin{cases} \quad \alpha_1 \geq \alpha_2 = \alpha_3 = \alpha_4 \\ \textit{or}\ \alpha_1 = \alpha_2 \geq \alpha_3 = \alpha_4 \\ \textit{or}\ \alpha_1 = \alpha_2 = \alpha_3 \geq \alpha_4 \quad . \end{cases} \tag{5.58}$$

Proof. We first prove (5.55). 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

$$\lambda_2(D_d^{-1}A_d) \geq \lim_{\alpha_4\to\infty} \lambda_2(D_d^{-1}A_d) = \lambda_2(C_d)\,, \tag{5.59}$$

where

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

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 $\alpha_3 \geq \alpha_2$ (one may see that they play a symmetric role in the definition of A_d and D_d). Then, setting

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

we have,

$$\begin{aligned}\lambda_{\min}(D_r^{-1}A_r) = \min_{\mathbf{v}\in\mathbb{R}^3\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^T A_r \mathbf{v}}{\mathbf{v}^T D_r \mathbf{v}} &= \min_{\mathbf{v}\in\mathbb{R}^3\backslash\{\mathbf{0}\}} \frac{\mathbf{v}^T \widetilde{A}_r \mathbf{v} + \frac{1}{2}(\alpha_3-\alpha_2)\,(\mathbf{v})_3{}^2}{\mathbf{v}^T \widetilde{D}_r \mathbf{v} + 2(\alpha_3-\alpha_2)\,(\mathbf{v})_3{}^2} \\ &\geq \min(\lambda_{\min}(\widetilde{D}_r^{-1}\widetilde{A}_r)\,,\, \frac{1}{4})\,. \qquad (5.60)\end{aligned}$$

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[1]). Since $2\alpha_1 - \sqrt{17\alpha_1^2+14\alpha_1\alpha_2+\alpha_2^2} \geq (2-\sqrt{17})(\alpha_1+\alpha_2)$ holds for $\alpha_1, \alpha_2 > 0$, the inequality (5.55) 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 (5.56).

To prove (5.57) 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 $\alpha_2 \geq \alpha_3$ (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 (5.53) 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

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

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

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

Similarly, let $\widetilde{D}_{0,0}$, $\widetilde{D}_{1,0}$ and $\widetilde{D}_{1,1}$ be the matrices defined via (5.54) 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., [3, Lemma 3.13]), we have

$$\begin{aligned} \lambda_2(D_d^{-1}A_d) &= \max_{\mathbf{v}\in\mathbb{R}^4\backslash\{\mathbf{0}\}} \min_{\mathbf{w}\perp\mathbf{v}} \frac{\mathbf{w}^T D_d^{-1/2} A_d D_d^{-1/2}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \\ &= \max_{\mathbf{v}\in\mathbb{R}^4\backslash\{\mathbf{0}\}} \min_{\mathbf{z}\perp\mathbf{v}} \frac{\mathbf{z}^T A_d\mathbf{z}}{\mathbf{z}^T D_d\mathbf{z}} \\ &= \max_{\mathbf{v}\in\mathbb{R}^4\backslash\{\mathbf{0}\}} \min_{\mathbf{z}\perp\mathbf{v}} \frac{(\alpha_1-\alpha_2)\mathbf{z}^T\widetilde{A}_{0,0}\mathbf{z} + (\alpha_2-\alpha_3)\mathbf{z}^T\widetilde{A}_{1,0}\mathbf{z} + \alpha_3\mathbf{z}^T\widetilde{A}_{1,1}\mathbf{z}}{(\alpha_1-\alpha_2)\mathbf{z}^T\widetilde{D}_{0,0}\mathbf{z} + (\alpha_2-\alpha_3)\mathbf{z}^T\widetilde{D}_{1,0} + \alpha_3\mathbf{z}^T\widetilde{D}_{1,1}\mathbf{z}} \\ &\geq \max_{\mathbf{v}\in\mathbb{R}^4\backslash\{\mathbf{0}\}} \min\left(\min_{\mathbf{z}\perp\mathbf{v}} \frac{\mathbf{z}^T\widetilde{A}_{0,0}\mathbf{z}}{\mathbf{z}^T\widetilde{D}_{0,0}\mathbf{z}}, \min_{\mathbf{z}\perp\mathbf{v}} \frac{\mathbf{z}^T\widetilde{A}_{1,0}\mathbf{z}}{\mathbf{z}^T\widetilde{D}_{1,0}\mathbf{z}}, \min_{\mathbf{z}\perp\mathbf{v}} \frac{\mathbf{z}^T\widetilde{A}_{1,1}\mathbf{z}}{\mathbf{z}^T\widetilde{D}_{1,1}\mathbf{z}}\right) \end{aligned}$$

Hence,

$$\lambda_2(D_d^{-1}A_d) \geq \min\left(\min_{\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4} \frac{\mathbf{z}^T\widetilde{A}_{0,0}\mathbf{z}}{\mathbf{z}^T\widetilde{D}_{0,0}\mathbf{z}}, \min_{\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4} \frac{\mathbf{z}^T\widetilde{A}_{1,0}\mathbf{z}}{\mathbf{z}^T\widetilde{D}_{1,0}\mathbf{z}}, \min_{\mathbf{z}\perp\widetilde{D}_{1,0}\mathbf{1}_4} \frac{\mathbf{z}^T\widetilde{A}_{1,1}\mathbf{z}}{\mathbf{z}^T\widetilde{D}_{1,1}\mathbf{z}}\right), \qquad (5.61)$$

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

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

Therefore, the proof of (5.57) is done if we show that the second term in (5.61) is the smallest. For this, we note that the vector $z = (64\ -34\ 33\ -62)^T$ is orthogonal to $\widetilde{D}_{1,0}\mathbf{1}_4 = (4\ 4\ 2\ 3)^T$ and that $\mathbf{z}^T\widetilde{A}_{1,0}\mathbf{z} = 15861.5$ with $\mathbf{z}^T\widetilde{D}_{1,0}\mathbf{z} = 34718$. Hence, the second term is smaller than 0.46. Further, the first term in (5.61) is larger than 0.46, as can be concluded from positive definiteness of

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

which implies $\mathbf{z}^T\widetilde{A}_{0,0}\mathbf{z} - 0.46\mathbf{z}^T\widetilde{D}_{0,0}\mathbf{z} \geq 0$ for any $\mathbf{z} \perp \widetilde{D}_{1,0}\mathbf{1}_4$. Similarly, the third term in (5.61) is larger than 0.46 since

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

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.

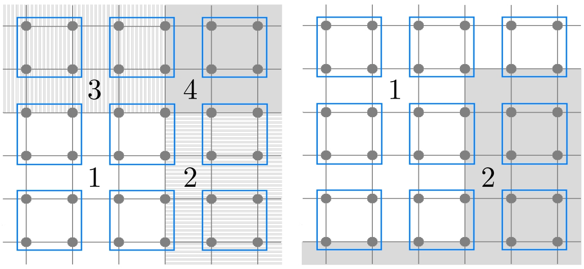
Eventually, we prove (5.58). If $\alpha_1 = \alpha_2$ and $\alpha_3 = \alpha_4$, the inequality is already proved in (5.56). If $\alpha_2 = \alpha_3 = \alpha_4$, taking $\widetilde{D}_d = \text{diag}(4\alpha_1\ \ 2(\alpha_1+\alpha_2)\ \ 2(\alpha_1+\alpha_2)\ \ 2(\alpha_1+\alpha_2))$ we have $\lambda_k(D_d^{-1}A_d) \geq \lambda_k(\widetilde{D}_d^{-1}A_d) \in \left\{0\,,\,\frac{1}{2}\,,\,\frac{3\alpha_1+\alpha_2}{4(\alpha_1+\alpha_2)}\,,\,\frac{1}{2}+\frac{3\alpha_1+\alpha_2}{4(\alpha_1+\alpha_2)}\right\}$ 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 $\alpha_1 \geq \alpha_4$ implies $10\alpha_1^2 + 4\alpha_1\alpha_4 + 2\alpha_4^2 \leq 16\alpha_1^2$, the inequality (5.58) follows. ■

5.5.2 Analysis

We consider the PDE (5.37) 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 directions (mesh box integration scheme [42]), 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 on Figure 5.1(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 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 reproduces the geometry of the problem.

The most general situation corresponding to this setting is then schematized on Figure 5.2(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 (5.14), the corresponding aggregate's matrices $A^{(k)}$ and $D^{(k)}$ are given by



(a) (b)

FIGURE 5.2: (a) general box aggregate situation with respect to discontinuities and (b) discontinuity nodes aggregated with white point nodes.

(5.53) and (5.54), respectively, with α_i, $i = 1, ..., 4$, being the PDE coefficient in the subdomain Ω_i.

Because of the assumption (5.14) and of Theorem 5.3, aggregate's quality $\mu_D^{(k)}$ is the inverse of the second smallest eigenvalue of $D^{(k)}{}^{-1} A^{(k)}$. Lemma 5.1 then shows us the following.

- The approach is robust in all cases, since, by (5.55), $\mu_D^{(k)}$ is always bounded above independently of the relation between the coefficients α_i.

- Nevertheless, from a practical viewpoint, (5.55) allows a significant decrease of aggregate's quality compared with the constant coefficient case. However, according to (5.57), which implies $\mu_D^{(k)} \leq 2.23$ (compared with 2 in constant coefficient case), a major deterioration is avoided when $\alpha_1 \geq \alpha_2, \alpha_3, \alpha_4$. 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 (5.2)(b). Then, if the rule above is applied; that is, if α_1 is in addition the largest coefficient, (5.58) applies and shows that there is no deterioration at all compared with the constant coefficient case.

5.5.3 Numerical example

Consider the PDE (5.37) on a square domain $\Omega = [0, 1] \times [0, 1]$ with $\beta = 0$,

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

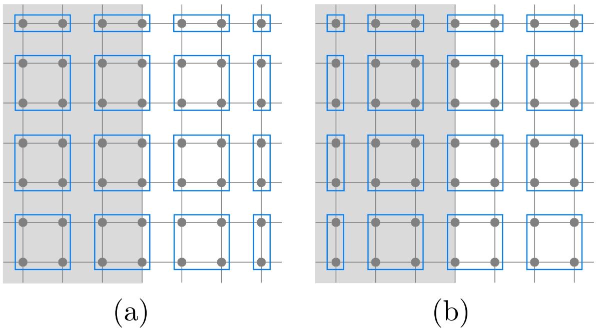


FIGURE 5.3: Two potential aggregation strategies for the numerical example.

and with Dirichlet boundary conditions. Consider the linear system (5.1) resulting from its five point finite difference discretization (mesh box integration scheme [42]) on the regular grid of mesh size $h = N^{-1}$. Since discontinuities 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 5.3(a)) or from the right bottom corner (as on Figure 5.3(b)).

	strategy (a)		strategy (b)	
N	$\max\limits_{k=0,\dots,n_c} \mu_D^{(k)}$	μ_D	$\max\limits_{k=0,\dots,n_c} \mu_D^{(k)}$	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

TABLE 5.2: The value of μ_D and of its upper bound (5.21) for different aggregation strategies and for $d = 10$.

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 (5.33) (for box aggregates) or from (5.35) 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 5.2. Observe that for the second strategy the convergence estimate is exactly the same as in the constant coefficient case. For box aggregates, 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

$$\mu_D^{(k)} = \lambda_2({D^{(k)}}^{-1} A^{(k)})^{-1} = \frac{4(1+d)}{3+d}\,, \tag{5.62}$$

using (5.56) 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 (5.62) implies $\mu_D^{(k)} = 3.38$ for $d = 10$ and $\mu_D^{(k)} \to 4$ for $d \to \infty$.

5.5.4 Sharpness of the estimate

Table 5.2 indicates that, once again, the upper bound (5.21) is seemingly asymptotically exact. In fact, the reasoning developed at the end of Section 5.4 shows that, asymptotically, μ_D cannot be smaller than 2 for an isotropic ($\alpha_x = \alpha_y$) PDE (5.37) 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 strong coefficient region, since then $\mu_D^{(k)} \le 2.23$. If, in addition, $\mu_D^{(k)} \le 2$, 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 5.2, 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 numerical example above with the first aggregation strategy (depicted on Figure 5.3(a)), which does not follow the "strong coefficient" rule. Note that, using the same trick as explained at the end of Section 5.4, 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 5.4, we need to construct two vectors $\widetilde{\mathbf{v}}$ and $\widetilde{\mathbf{c}}$ such that

$$\frac{\sum_{k=0}^{n_c} \gamma_k^2 \mu_D^{(k)}}{\sum_{k=0}^{n_c} \gamma_k^2} \to \max_{k=1,\dots,n_c} \mu_D^{(k)} \text{ for } N \to \infty , \tag{5.63}$$

whereas ε, defined by (5.44), goes to 0 as N becomes large. In the example under investigation, there are some pair and singleton aggregates (see Figure 5.3), but we limit the support of both vectors to the $(2\ell+1)\times(2\ell+1)$ box aggregates, where $\ell = N/4-1$. We identify each such aggregate k with a couple $(i_x^{(k)}, i_y^{(k)})$ of indices, $1 \le i_x^{(k)}, i_y^{(k)} \le 2\ell + 1$, 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 5.4 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 $\mathbf{v}_d = (\tau\ 1\ \tau\ 1)^T$ for discontinuity aggregate, with $\tau = -(d+1)/2$, and by $\mathbf{v}_o = (-1\ 1\ -1\ 1)^T$ for the ordinary ones. The corresponding local energy (semi-) norms are given by $\theta_d^2 = \mathbf{v}_d^T A^{(k)} \mathbf{v}_d = (3+d)^2/2$ for discontinuity aggregates, by $\theta_o^2 = \mathbf{v}_o^T A^{(k)} \mathbf{v}_o = 8$ 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 $(\widetilde{\mathbf{v}}^{(1)\,T}, \widetilde{\mathbf{v}}^{(2)\,T}, \cdots \widetilde{\mathbf{v}}^{(n_c)\,T})^T$ with

$$\widetilde{\mathbf{v}}^{(k)} = \ell^{-1}(\ell - |\ell+1-i_y^{(k)}|) \times \begin{cases} \frac{\tau}{2}\ell^{-1}\mathbf{v}_o & \text{if } 1 \le i_x^{(k)} < \ell+1 \\ \mathbf{v}_d & \text{if } i_x^{(k)} = \ell+1 \\ \frac{1}{2}\ell^{-1}\mathbf{v}_o & \text{if } \ell+1 < i_x^{(k)} \le 2\ell+1 \\ \mathbf{0} & \text{otherwise}, \end{cases} \tag{5.64}$$

and the vector $\widetilde{\mathbf{c}}$ corresponds to $(\widetilde{\mathbf{c}}^{(1)\,T}, \widetilde{\mathbf{c}}^{(2)\,T}, \cdots \widetilde{\mathbf{c}}^{(n_c)\,T})^T$ with

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

From (5.64) we conclude that

$$\gamma_k^2 = \ell^{-2}(\ell - |\ell+1-i_y^{(k)}|)^2 \times \begin{cases} \frac{\tau^2}{4}\ell^{-2}\,\theta_o^2 & \text{if } 1 \le i_x^{(k)} < \ell+1 \\ \theta_d^2 & \text{if } i_x^{(k)} = \ell+1 \\ \frac{1}{4}\ell^{-2}\,d\theta_o^2 & \text{if } \ell+1 < i_x^{(k)} \le 2\ell+1 \\ \mathbf{0} & \text{otherwise}, \end{cases} \tag{5.65}$$

and, setting

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

there holds

$$\begin{aligned} \sum_{k:\, i_x^{(k)}=\ell+1} \gamma_k^2 &= \theta_d^2 \sum_{1<i_y^{(k)}<2\ell+1} \ell^{-2}(\ell - |\ell+1-i_y^{(k)}|)^2 \;=\; \theta_d^2\ell^{-2}\mathrm{s}(\ell), \\ \sum_{k:\, i_x^{(k)}\neq\ell+1} \gamma_k^2 &= \theta_o^2\,\frac{d+\tau^2}{4} \sum_{\substack{1<i_y^{(k)}<2\ell+1 \\ 1\le i_x^{(k)}<\ell+1}} \ell^{-4}(\ell - |\ell+1-i_y^{(k)}|)^2 \;=\; \theta_o^2\,\frac{d+\tau^2}{4}\ell^{-3}\mathrm{s}(\ell). \end{aligned}$$

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 (5.63) 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 $(\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})^T A_r (\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}})$ as expressed by (5.38) 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 $i_y^{(k)} \le 2\ell$. 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

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

Therefore, using (5.38) with, this time, the first term being nonzero and the second one vanishing because of the limited scope of $\widetilde{\mathbf{c}}+\widetilde{\mathbf{v}}$, we have

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

whereas

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

Hence, $\widetilde{\mathbf{v}}^T D\widetilde{\mathbf{v}} = \mathcal{O}(\ell)$ whereas $(\widetilde{\mathbf{c}} + \widetilde{\mathbf{v}})^T A_{rest}(\widetilde{\mathbf{c}} + \widetilde{\mathbf{v}}) = \mathcal{O}(1)$, showing with (5.55) that (5.44) holds with $\varepsilon = \mathcal{O}(\ell^{-1})$, and therefore, together with (5.63), proving the asymptotical sharpness of the estimate.

5.6 Conclusion

We have developed an analysis of an 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 5.1).

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.

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 whether 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.

Chapter 6

Fourier Analysis of aggregation-based two-grid method for edge element

Summary

We consider Reitzinger and Schöberl multigrid method for curl-curl problems discretized with edge finite elements. We perform a Fourier analysis of its two-grid variant and show that the corresponding convergence rate can be bounded independently of the problem size. This result is also compared with the actual two-grid convergence, indicating that the analysis is accurate. Some numerical experiments are further performed in multigrid setting with various cycling strategies, showing that an optimal implementation of the method may be obtained when using the K-cycle.

6.1 Introduction

We consider multigrid methods for linear systems resulting from the discretization with edge elements of

$$\mathrm{curl}(\alpha\,\mathrm{curl}(\mathbf{E}))+\beta\,\mathbf{E}=\mathbf{f}\;\text{ on }\Omega\,, \tag{6.1}$$

where $\alpha,\beta>0$ and $\Omega\subset\mathbb{R}^3$. This problem arises when the vector potential is computed in magnetostatics, when time-harmonic formulation of Maxwell's equations is used, or when eddy current approximation is considered (see [8, 5] and the references therein). Note that, since edge element discretization is performed on (6.1), the degrees of freedom in the linear system are associated with edges.

It is well known that standard multigrid techniques, if applied to such discretized problems, have poor convergence properties. When the multigrid hierarchy is induced by the refinement of an underlying coarse mesh, as in geometric multigrid, it is further proved in [77] that a two-grid method can not be optimal if based on a simple point smoother (like standard Jacobi or Gauss-Seidel). Modifications of standard multigrid by either using special smoothing techniques [29, 2] or by decoupling multilevel hierarchies for edge and node unknowns [30, 35, 4] have been proposed recently to overcome this difficulty.

Here we consider more precisely an algebraic multigrid method based on the coarsening by aggregation of edge unknowns, as introduced by Reitzinger and Schöberl in [52]. The main idea behind this approach is to perform the aggregation of edge unknowns so that it also corresponds (via a given edge-node incidence matrix) to an aggregation of nodes. Doing so, one insures the correct representation of the near null space of the problem on coarser levels.

Note that, alike classical multigrid methods based on coarsening by aggregation [11, 20, 48], Reitzinger and Schöberl (RS) approach has low computational cost per iteration and modest storage requirements. However, similarly to classical aggregation techniques, piecewise constant (up to the edge's orientation) prolongation is used, which in turn results in level-dependent convergence behaviour with V-cycle setting (as already observed in [52]).

Regarding aggregation techniques for elliptic boundary value problems, several approaches have been proposed recently to overcome this lack of optimality. One consists in using, instead of simple V-cycle, a more sophisticated K-cycle, in which Krylov subspace acceleration is performed at each level [49]. It is also possible to improve the scalability by increasing the number of smoothing steps on coarser levels [32]. Such approaches can also be implemented with RS algebraic multigrid, keeping the original advantage of modest resource requirements. As for the second implementation, numerical experiments in the original RS paper [52] seem to indicate that this approach has level-dependent convergence similar to that of V-cycle. Regarding the implementation of RS approach with Krylov-based (K-) cycling strategy, it is however an open question to what extent level-independent convergence properties can be obtained.

Here we investigate this point. We start by assessing the two-grid convergence of the RS approach, since a (truly) multigrid method can not be optimal if the convergence rate of the corresponding two-grid scheme deteriorates with the problem size. We evaluate the convergence properties using Fourier analysis. This technique was adapted only recently in [9] to curl-curl problems and, as far as we know, no such analysis is available for the RS approach.

More precisely, a (local) two-grid Fourier analysis for a two-dimensional model problem based on geometrical (bilinear) prolongation operator is performed in [9] for hybrid [29] and AFW block [2] smoothers. Here we extend presented ideas to a piecewise constant (RS-like) edge prolongation and show that the considered two-grid scheme with hybrid smoother has level independent convergence properties in three dimensions. The use of three-dimensional setting is motivated by its importance in the field of electromagnetical computations.

Note that instead of using the local Fourier analysis framework, we perform an exact Fourier analysis for problems with periodic boundary conditions. Both approaches are quite similar, the latter allowing however to account for the grid size. This further enables to supplement the Fourier analysis results with the assessment of convergence properties of the two-grid scheme for the same problem with Dirichlet boundary conditions. Their comparison indicates that Fourier analysis gives an accurate prediction of the convergence rate.

Once the level-independent convergence is proved for a two-grid scheme, some numerical experiments are performed using the corresponding multigrid ingredients with V-, W- and K-cycles (the two latter approaches have similar operation count per iteration). The results indicate that the convergence speed in the case of the first two cycling strategies deteriorates with the number of levels, whereas the last approach has almost the same iteration count as the two-grid scheme on the finest level.

The reminder of this paper is organized as follows. In Section 6.2 we recall some useful properties of discrete curl-curl problems and present the main ingredients of the RS approach. In Section 6.3 we give the Fourier representation of these ingredients for the considered three-dimensional model problem. The results of the Fourier analysis together with numerical experiments are presented and discussed in Section 6.4.

6.2 Preliminaries

6.2.1 Discretized problem

The use of edge finite elements requires the weak formulation of the problem (6.1) as can be found, for instance, in [29]. More precisely, letting

$$H(\text{curl};\Omega) = \left\{\mathbf{v} \in \mathbf{L}^2(\Omega);\ \text{curl}(\mathbf{v}) \in \mathbf{L}^2(\Omega)\right\} \ ,$$

the "weak" problem consists in determining the vector $\mathbf{E} \in H_*(\text{curl};\Omega) \subset H(\text{curl};\Omega)$ such that

$$\int_\Omega \alpha\, \text{curl}(\mathbf{E}) \cdot \text{curl}(\mathbf{v}) dV + \int_\Omega \beta\, \mathbf{E} \cdot \mathbf{v} dV = \int_\Omega \mathbf{f} \cdot \mathbf{v} dV \qquad \forall \mathbf{v} \in H_*(\text{curl};\Omega)\,. \tag{6.2}$$

This formulation can be recovered from the original problem, assuming that

$$\int_{\partial\Omega} (\text{curl}(\mathbf{E}) \times \mathbf{v}) \cdot \mathbf{n}\, d\sigma = \int_{\partial\Omega} \text{curl}(\mathbf{E}) \cdot (\mathbf{v} \times \mathbf{n})\, d\sigma = 0 \tag{6.3}$$

holds, through the multiplication of (6.1) by a test function $\mathbf{v}$ followed by the application of Green's identity

$$\int_\Omega \mathrm{curl}(\mathbf{w}) \cdot \mathbf{v} dV - \int_\Omega \mathbf{w} \cdot \mathrm{curl}(\mathbf{v}) dV = \int_{\partial\Omega} (\mathbf{w} \times \mathbf{v}) \mathbf{n} d\sigma \,.$$

The condition (6.3) is fulfilled, for instance, when homogeneous Dirichlet boundary conditions

$$\mathbf{v} \times \mathbf{n} = \mathbf{0} \qquad \forall \mathbf{v} \in H_*(\mathrm{curl};\Omega)) = H_0(\mathrm{curl};\Omega)) \tag{6.4}$$

are used.

In edge element discretization the degrees of freedom are associated with edges; that is, to any edge denoted by $k = (j_1, j_2)$ with nodes j_1 and j_2 being, respectively, the starting and the end points, corresponds an unknown given by

$$x_k = \int_{j_1}^{j_2} \mathbf{E} \cdot d\mathbf{s} \,.$$

The resulting system

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

is then such that

$$A = \alpha K_{cc} + \beta h^2 M \,, \tag{6.5}$$

where K_{cc} and M are matrices that correspond, respectively, to the stiffness (curl-curl) and the mass terms in (6.2).

One of the reasons why classical multigrid does not suit for such problems is the large near null space of A induced by the null space $\mathcal{N}(K_{cc})$ of K_{cc}. This latter is a discrete representation of the null space of curl$(\cdot)$ operator, which contains all vectors of the form grad(f). That is, $\mathcal{N}(K_{cc})$ is formed by the vectors $G\mathbf{v}$, where G is a discrete gradient matrix. As a straightforward consequence, we thus have $K_{cc}G = O$. When edge shape functions are properly normalized, it can be proved (see [8] and the references therein) that G coincides with the edge-node incidence matrix; that is, denoting by (j_1, j_2) an edge with j_1 as a starting node and j_2 as the end node, we have

$$(G)_{kj} = \begin{cases} 1 & \text{if } k = (*, j) \,, \\ -1 & \text{if } k = (j, *) \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{6.6}$$

Since G associates a given node with several edges, it can be viewed as a transfer operator from nodal to edge representation, its transpose G^T performing the inverse operation. Note, however, that the number of nodes and edges is generally not the same and G^TG, $GG^T \neq I$.

6.2.2 Reitzinger and Schöberl (RS) multigrid

It is a common practice to base the design of an algebraic multigrid method on the definition of a problem-dependent prolongation matrix P. Once the prolongation is available, the restriction is set to its transpose P^T and the coarse grid matrix is given by the Galerkin formula $A_c = P^T AP$. The same procedure can then be applied to the coarse grid system, and so on, until the coarsest grid which is chosen small enough. The above ideas can be extended to algebraic multigrid for edge element discretizations of (6.2), provided that they are combined with an appropriate smoothing scheme (for instance, hybrid [29] or AFW [2] smoothers).

Now, in the RS approach, one first performs an agglomeration of n nodes into $n_c > 0$ aggregates $\Gamma_k, k = 1, ..., n_c$. The edge prolongation matrix is then defined by

$$(P^{(e)})_{jk} = \begin{cases} \ \ 1 & \text{if } j = (j_1, j_2) \text{ and } k = (k_1, k_2) \text{ with } j_1 \in \Gamma_{k_1},\ j_2 \in \Gamma_{k_2} \\ -1 & \text{if } j = (j_1, j_2) \text{ and } k = (k_2, k_1) \text{ with } j_1 \in \Gamma_{k_1},\ j_2 \in \Gamma_{k_2} \\ \ \ 0 & \text{otherwise}\,. \end{cases} \tag{6.7}$$

In other words, edges are grouped together in a unique “edge” aggregate if they connect nodes belonging to same “node” aggregates.

Note that, setting the auxiliary “nodal” prolongation to

$$(P)^{(n)}_{jk} = \begin{cases} 1 & \text{if } j \in \Gamma_k, \quad k = 1, ..., n_c\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{6.8}$$

one satisfies a seemingly important commutation property (see [52] for the proof)

$$GP^{(n)} = P^{(e)} G_c\,, \tag{6.9}$$

with G and G_c being, respectively, the fine and coarse edge-node incidence matrices. The importance of the property (6.9) mainly resides in the fact that the columns of G span the near null space of A. The commutation property then ensures that the columns of G_c belong to the near null space of $A_c = {P^{(e)}}^T AP^{(e)}$.

We also observe that the range of the prolongation matrix $P^{(e)}$ as defined by (6.7) does not contain the entirety of the near null space of A. Therefore, some near null space components of the error are not reduced appropriately by the coarse-grid correction. On the other hand, a simple pointwise smoother R cannot reduce these components as well, since $(I - R^{-1}A)\mathbf{v} \approx (1 - \mathcal{O}(h^2))\mathbf{v}$ for any $\mathbf{v} \in \mathcal{N}(K_{cc})$. More sophisticated smoothers should therefore be used which treat appropriately the near null space modes that are not in the range of the prolongation.

Here we consider one of such approaches known as the hybrid smoother. Its main idea is to smooth separately the near null space components of the error, so that they can be correctly approximated on the coarser grid. The hybrid smoother involves two additional matrices: an edges $R^{(e)}$ and a nodal $R^{(n)}$ smoother. If these matrices are chosen to be the lower (or upper) triangular part of, respectively, A and $A^{(n)} = G^T AG$, we recover the classical version of the hybrid smoother. In what follows we however also consider diagonal (or Jacobi-like) smoothers. In the smoothing procedure given below, the number of smoothing steps can be integer or half-integer. In the former case an additional binary parameter η is used to determine if the extra half-step should be performed in the beginning ($\eta = \uparrow$) or at the end ($\eta = \downarrow$) of the hybrid smoothing scheme.

```
Hybrid smoother: $\mathbf{x}_{n+1} = \mathrm{HS}(\mathbf{x}_n, \mathbf{b}, \nu, \eta)$
    (1) if $\nu$ is half-integer and $\eta = \uparrow$: perform the steps (b), (d)-(f) below
    (2) repeat $\lfloor \nu \rfloor$ times:
        (a) Edge pre-smoothing: $\mathbf{x}_n \leftarrow \mathbf{x}_n + {R^{(e)}}^{-1} (\mathbf{b} - A\mathbf{x}_n)$
        (b) Restrict to nodal variables: $\mathbf{r} = G^T (\mathbf{b} - A\mathbf{x}_n);\quad \mathbf{e} = \mathbf{0}$
        (c) Forward sweep: $\mathbf{e} \leftarrow \mathbf{e} + {R^{(n)}}^{-1} \left(\mathbf{r} - A^{(n)}\mathbf{e}\right)$
        (d) Backward sweep: $\mathbf{e} \leftarrow \mathbf{e} + {R^{(n)}}^{-T} \left(\mathbf{r} - A^{(n)}\mathbf{e}\right)$
        (e) Transfer back to edge variables: $\mathbf{x}_n \leftarrow \mathbf{x}_n + G\mathbf{e}$
        (f) Edge post-smoothing: $\mathbf{x}_{n+1} \leftarrow \mathbf{x}_n + {R^{(e)}}^{-T} (\mathbf{b} - A\mathbf{x}_n)$
    (3) if $\nu$ is half-integer and $\eta = \downarrow$: perform the steps (a)-(c), (e) above
```

Now, the two-grid version of the RS approach based on the hybrid smoother is presented below. Note that the pre- and post-smoothing setting are treated differently when the number of smoothing steps is half-integer. The reason for doing so is that the resulting two-grid (and the induced multigrid) preconditioner is then symmetric when $\nu_1 = \nu_2$.

```
RS two-grid cycle: $\mathbf{x}_{n+1} = \mathrm{RS}_{\mathrm{TG}}(\mathbf{b}, \mathbf{x}_n, \nu_1, \nu_2)$
    (1) $\nu_1$ steps of pre-smoothing: $\mathbf{x}_n \leftarrow \mathrm{HS}(\mathbf{x}_n, \nu_1, \mathbf{b}, \uparrow)$
    (2) Compute residual: $\mathbf{r} = \mathbf{b} - A\mathbf{x}_n$
    (3) Restrict residual: $\mathbf{r}_c = {P^{(e)}}^T \mathbf{r}$
    (4) Coarse grid correction: $\mathbf{e}_c = A_c^{-1}\mathbf{r}_c$
    (5) Prolongate coarse-grid correction: $\mathbf{x}_n \leftarrow \mathbf{x}_n + P^{(e)}\mathbf{e}_c$
    (6) $\nu_2$ steps of post-smoothing: $\mathbf{x}_{n+1} \leftarrow \mathrm{HS}(\mathbf{x}_n, \nu_2, \mathbf{b}, \downarrow)$
```

When applying this algorithm, the error satisfies

$$A^{-1}\mathbf{b} - \mathbf{x}_{n+1} = E_{TG}\left(A^{-1}\mathbf{b} - \mathbf{x}_n\right) ,$$

where the iteration matrix E_{TG} is given by

$$E_{TG}^{(\nu_1,\nu_2)} = S_{\downarrow}^{(\nu_2)} \left(I - P^{(e)} \left({P^{(e)}}^T AP^{(e)}\right)^{-1} {P^{(e)}}^T A\right) S_{\uparrow}^{(\nu_1)} . \tag{6.10}$$

The pre-smoothing iteration matrix satisfies

$$S_{\uparrow}^{(\nu_1)} = \begin{cases} S^{\nu_1} & \text{if } \nu_1 \text{ is integer} \\ S^{\nu_1}\,(I - {R^{(e)}}^{-T} A)(I - G\,{R^{(n)}}^{-T} G^T A) & \text{if } \nu_1 \text{ is half-integer}\,, \end{cases} \tag{6.11}$$

where

$$S = (I - {R^{(e)}}^{-T} A)(I - G\,{X^{(n)}}^{-1} G^T A)(I - {R^{(e)}}^{-1} A)\,, \tag{6.12}$$

with $X^{(n)}$ defined by

$$I - {X^{(n)}}^{-1} A^{(n)} = (I - {R^{(n)}}^{-T} A^{(n)})(I - {R^{(n)}}^{-1} A^{(n)})\,. \tag{6.13}$$

Similarly, the post-smoothing iteration matrix is given by

$$S_{\downarrow}^{(\nu_2)} = \begin{cases} S^{\nu_2} & \text{if } \nu_2 \text{ is integer} \\ (I - G\,{R^{(n)}}^{-1} G^T A)(I - {R^{(e)}}^{-1} A)\,S^{\nu_2} & \text{if } \nu_2 \text{ is half-integer}\,, \end{cases} \tag{6.14}$$

Our main objective is the analysis of the spectral radius $\rho\,(E_{TG})$ of E_{TG} , which governs convergence of the two-grid method. We note, however, that the RS multigrid method can be used as a preconditioner, with the preconditioner matrix B_{TG} in the two-grid case given by

$$E_{TG} = I - B_{TG}A\,.$$

Since the system matrix A resulting from the edge element discretization of (6.2) is symmetric positive definite (SPD), and since B_{TG} can be checked to be SPD for $\nu_1 = \nu_2$, the linear system can be solved by the preconditioned conjugated gradient method [28]. In this latter case, the relevant convergence parameter is the condition number (see, e.g., [24, Theorem 10.2.6]), given by

$$\kappa\,(B_{TG}A) = \frac{\lambda_{\max}\,(B_{TG}A)}{\lambda_{\min}\,(B_{TG}A)} = \frac{1 - \lambda_{\min}\,(E_{TG})}{1 - \lambda_{\max}\,(E_{TG})}\,.$$

Unless the coarse grid matrix is weighted (as it is sometimes the case below), one can check that $A^{1/2}E_{TG}A^{-1/2} = I - A^{1/2}B_{TG}A^{1/2}$ is semi-positive definite (see Theorem 3.19 in [65] for nonnegative definiteness of ${B_{TG}}^{-1} - A$ and Theorem 2.1 in [44] with $n_c > 0$ for presence of zero eigenvalues) and, hence, $\lambda_{\min}(E_{TG}) = 0$. The condition number in such case can therefore be deduced from

$$\kappa\,(B_{TG}A) = \frac{1}{1 - \rho\,(E_{TG})}\,, \tag{6.15}$$

and will not be reported explicitly.

Note that, since $AS_{\uparrow}^{(\nu)} = {S_{\downarrow}^{(\nu)}}^T A$ and $AC = C^T A$, where C stands for the coarse grid correction, we have

$$\begin{aligned}
\rho\left(E_{TG}^{(\nu_1,\nu_2)}\right) &= \rho\left({E_{TG}^{(\nu_1,\nu_2)}}^T\right) \\
&= \rho\left(A^{-1}{E_{TG}^{(\nu_1,\nu_2)}}^T A\right) \\
&= \rho\left(A^{-1}\,{S_{\uparrow}^{(\nu_1)}}^T\, C^T\, {S_{\downarrow}^{(\nu_2)}}^T A\right) \\
&= \rho\left(S_{\downarrow}^{(\nu_1)} C S_{\uparrow}^{(\nu_2)}\right) \\
&= \rho\left(E_{TG}^{(\nu_2,\nu_1)}\right) \,,
\end{aligned} \tag{6.16}$$

and the number of pre- and post-smoothing iterations can be interchanged without any impact on the asymptotic two-grid convergence. Moreover, if ν_1 and ν_2 are both integers or half-integers, using $A^{(n)} = G^T A G$ we have $S_{\uparrow}^{(1/2)} S_{\downarrow}^{(1/2)} = S$, which further implies

$$S_{\uparrow}^{(\nu_1)} S_{\downarrow}^{(\nu_2)} = S^{\nu_1+\nu_2} = S_{\uparrow}^{(\nu_1+\nu_2)} = S_{\downarrow}^{(\nu_1+\nu_2)} \,,$$

and, hence,

$$\rho\left(E_{TG}^{(\nu_1,\nu_2)}\right) = \rho\left(S_{\downarrow}^{(\nu_2)} C S_{\uparrow}^{(\nu_1)}\right) = \rho\left(C S_{\uparrow}^{(\nu_1)} S_{\downarrow}^{(\nu_2)}\right) = \rho\left(C S_{\uparrow}^{(\nu_1+\nu_2)}\right) = \rho\left(E_{TG}^{(\nu_1+\nu_2,0)}\right) \,.$$

In this case the two-grid convergence factor depends only on the overall number $\nu = \nu_1 + \nu_2$ of smoothing steps.

Now, in what follows we consider the hybrid smoother with $\nu = 1/2$, 1 and 2 smoothing iterations. In the two latter cases both ν_1 and ν_2 are either integer or half-integer; hence, the asymptotic convergence factor then depends only on ν. The case $\nu = 1/2$ corresponds to either $(\nu_1, \nu_2) = (1/2, 0)$ or $(0, 1/2)$. However, it follows from (6.16) that both have the same asymptotic convergence, this latter depending again on ν. We therefore report the results with respect to ν instead of (ν_1, ν_2), at least in the two-grid setting.

6.3 Fourier analysis

6.3.1 Model problem

Consider now $\Omega = (0,1)^3$ with periodic boundary conditions. The vectors in $H_*(\text{curl};\Omega) = H_P(\text{curl};\Omega)$ are therefore also assumed periodic; that is, for any $\mathbf{v} \in H_P(\text{curl};\Omega)$ we have $\mathbf{v}(0,y,z) = \mathbf{v}(1,y,z)$, $\mathbf{v}(x,0,z) = \mathbf{v}(x,1,z)$ and $\mathbf{v}(x,y,0) = \mathbf{v}(x,y,1)$. Note that the constraint (6.3) is then satisfied since the contributions of opposite faces of $\partial\Omega$ are opposite. The weak formulation (6.2) can therefore be considered and the resulting problem is further discretized by trilinear (brick) edge elements[1] (see, e.g., [66, p.54]) on the cubic grid $(N+1)\times(N+1)\times(N+1)$ of grid size $h = N^{-1}$.

Since it is sometimes convenient to refer to an edge via its position on the grid, we also associate a triple $\mathbf{k} = (k_x, k_y, k_z)$ to any node unknown such that $h\mathbf{k}$ gives node's coordinate position, and to any edge unknown such that $h\mathbf{k}$ correspond to coordinate position of the corresponding edge's middle point. Note that k_a, $a = x, y, z$, is a half-integer if the corresponding edge is oriented in the a direction and integer otherwise.

Now, following the notation in [9], we set

$$\mathcal{I}(\Delta_x, \Delta_y, \Delta_z) = \{(k_x + \Delta_x, k_y + \Delta_y, k_z + \Delta_z | 0 \le k_x, k_y, k_z < N)\},$$

and let $\mathcal{E}^{[x]} = \mathcal{I}(1/2, 0, 0)$, $\mathcal{E}^{[y]} = \mathcal{I}(0, 1/2, 0)$, $\mathcal{E}^{[z]} = \mathcal{I}(0, 0, 1/2)$ and $\mathcal{N} = \mathcal{I}(0, 0, 0)$ be the index set of, respectively, edge unknowns in x, y and z directions and node unknowns. We also note that, for any edge $\mathbf{k}$, the set of its neighbours; that is, the set of edges that have a common element with $\mathbf{k}$ is given by $\langle \mathbf{k} + \mathbf{t} \rangle = (\langle k_x + t_x \rangle, \langle k_y + t_y \rangle, \langle k_z + t_z \rangle)^T$, where $\mathbf{t} \in \mathcal{T}$, with

$$\mathcal{T} = \{\mathbf{t} = (t_x, t_y, t_z) \mid t_x, t_y, t_z \in \{1,\ \frac{1}{2},\ 0,\ -\frac{1}{2},\ -1\} \text{ and } t_x + t_y + t_z \in \mathbb{Z}\}$$

and

$$\langle k \rangle = \begin{cases} k & \text{if } k < N\,, \\ k - N & \text{otherwise}\ . \end{cases}$$

Assuming that the matrix A arises from the discretization of (6.2) with coefficients α and β being constant, the entry $(A)_{\mathbf{k}\mathbf{k}'}$ for a given edge orientation depends on the relative edge's position $\mathbf{k} - \mathbf{k}^{'}$, and, hence, satisfies

$$(A\mathbf{v})_{\mathbf{k}} = \begin{cases} \sum_{t\in\mathcal{T}} s_{\mathbf{t}}^{[x]}(\mathbf{v})_{\langle \mathbf{k}+\mathbf{t} \rangle} & \text{if } \mathbf{k} \in \mathcal{E}^{[x]}\,, \\ \sum_{t\in\mathcal{T}} s_{\mathbf{t}}^{[y]}(\mathbf{v})_{\langle \mathbf{k}+\mathbf{t} \rangle} & \text{if } \mathbf{k} \in \mathcal{E}^{[y]}\,, \\ \sum_{t\in\mathcal{T}} s_{\mathbf{t}}^{[z]}(\mathbf{v})_{\langle \mathbf{k}+\mathbf{t} \rangle} & \text{if } \mathbf{k} \in \mathcal{E}^{[z]}\,. \end{cases} \tag{6.17}$$

Similarly to the two-dimensional analysis in [9], we associate a stencil to edges in any of the three directions. For instance, for edges in x direction the stencil can be represented

[1] the elements that belong to the boundary edges being periodically extended to the opposite boundary.

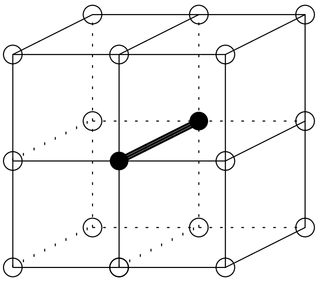


FIGURE 6.1: Edge neighbourhood.

as a triple of two-dimensional stencils

$$\mathcal{ST}^{[x]}(L) = \left[\left[\begin{array}{ccccc} \circ & s^{[x]}_{-\frac{1}{2},-\frac{1}{2},1} & \circ & s^{[x]}_{-\frac{1}{2},\frac{1}{2},1} & \circ \\ s^{[x]}_{-\frac{1}{2},-1,\frac{1}{2}} & & s^{[x]}_{-\frac{1}{2},0,\frac{1}{2}} & & s^{[x]}_{-\frac{1}{2},1,\frac{1}{2}} \\ \circ & s^{[x]}_{-\frac{1}{2},-\frac{1}{2},0} & \bullet & s^{[x]}_{-\frac{1}{2},\frac{1}{2},0} & \circ \\ s^{[x]}_{-\frac{1}{2},-1,-\frac{1}{2}} & & s^{[x]}_{-\frac{1}{2},0,-\frac{1}{2}} & & s^{[x]}_{-\frac{1}{2},1,-\frac{1}{2}} \\ \circ & s^{[x]}_{-\frac{1}{2},-\frac{1}{2},-1,} & \circ & s^{[x]}_{-\frac{1}{2},\frac{1}{2},-1} & \circ \end{array} \right] \right.$$
$$\left. \left[\begin{array}{ccc} s^{[x]}_{0,-1,1} & s^{[x]}_{0,0,1} & s^{[x]}_{0,1,1} \\ s^{[x]}_{0,-1,0} & s^{[x]}_{0,0,0} & s^{[x]}_{0,1,0} \\ s^{[x]}_{0,-1,-1} & s^{[x]}_{0,0,-1} & s^{[x]}_{0,1,-1} \end{array} \right] \left[\begin{array}{ccccc} \circ & s^{[x]}_{\frac{1}{2},-\frac{1}{2},1} & \circ & s^{[x]}_{\frac{1}{2},\frac{1}{2},1} & \circ \\ s^{[x]}_{\frac{1}{2},-1,\frac{1}{2}} & & s^{[x]}_{\frac{1}{2},0,\frac{1}{2}} & & s^{[x]}_{\frac{1}{2},1,\frac{1}{2}} \\ \circ & s^{[x]}_{\frac{1}{2},-\frac{1}{2},0} & \bullet & s^{[x]}_{\frac{1}{2},\frac{1}{2},0} & \circ \\ s^{[x]}_{\frac{1}{2},-1,-\frac{1}{2}} & & s^{[x]}_{\frac{1}{2},0,-\frac{1}{2}} & & s^{[x]}_{\frac{1}{2},1,-\frac{1}{2}} \\ \circ & s^{[x]}_{\frac{1}{2},-\frac{1}{2},-1,} & \circ & s^{[x]}_{\frac{1}{2},\frac{1}{2},-1} & \circ \end{array} \right] \right] ,$$

the edges in this stencil being also represented on Figure 6.1. More particularly, the "bold" segment corresponds to the considered edge and to the entry $s^{[x]}_{0,0,0}$, the other 8 edges in x direction forming the rest of the central 2D stencil; the remaining 24 edges are oriented in y and z direction and belong to two planes, those with smaller x coordinate corresponding to the first 2D stencil, the others being associated to the third one. For these two stencils, the black and white bullets schematize the nodes (as on Figure 6.1) and the value between two bullets in the stencil corresponds to the edge between them on the figure.

Note that the same stencil representation can be used in y and z directions. In these latter cases, to avoid any confusion on the choice of directions perpendicular to the considered edge, we assume that stencil entries $s^{[a]}_{t_x,t_y,t_z}$, $a = y, z$, of every column have the same t_x, and the entries of every line have the same t_z. Note that the stencil in the x direction given above also satisfies this assumption.

Now, assuming edge elements oriented in the positive axis direction, the matrices K_{cc} and M for the considered problem satisfy (6.17) with

$$\mathcal{ST}^{[x]}(K_{cc}) = \frac{1}{6}\left[\left[\begin{array}{ccccc} \circ & 1 & \circ & -1 & \circ \\ -1 & & -4 & & -1 \\ \circ & 4 & \bullet & -4 & \circ \\ 1 & & 4 & & 1 \\ \circ & 1 & \circ & -1 & \circ \end{array}\right]\left[\begin{array}{ccc} -2 & -2 & -2 \\ & & \\ -2 & 16 & -2 \\ & & \\ -2 & -2 & -2 \end{array}\right]\left[\begin{array}{ccccc} \circ & -1 & \circ & 1 & \circ \\ 1 & & 4 & & 1 \\ \circ & -4 & \bullet & 4 & \circ \\ -1 & & -4 & & -1 \\ \circ & -1 & \circ & 1 & \circ \end{array}\right]\right], \tag{6.18}$$

and

$$\mathcal{ST}^{[x]}(M) = \frac{1}{36}\left[\left[\,\bullet\,\right]\left[\begin{array}{ccc} 1 & 4 & 1 \\ & & \\ 4 & 16 & 4 \\ & & \\ 1 & 4 & 1 \end{array}\right]\left[\,\bullet\,\right]\right], \tag{6.19}$$

respectively. The stencils are the same in the y and z directions.

6.3.2 Fourier analysis setting

For edge unknowns, we define the Fourier modes separately in each direction:

$$\left(\mathbf{u}^{[a]}(\theta)\right)_{\mathbf{k}} = \begin{cases} e^{i\,\mathbf{k}\theta} = \frac{1}{\sqrt{N^3}} e^{i(k_x\theta_x + k_y\theta_y + k_z\theta_z)} & \text{if } \mathbf{k} \in \mathcal{E}^{[a]} \\ 0 & \text{otherwise}, \end{cases} \tag{6.20}$$

whereas for the node unknowns the usual definition is adopted

$$(\mathbf{u}(\theta))_{\mathbf{k}} = e^{i\,\mathbf{k}\theta} = \frac{1}{\sqrt{N^3}} e^{i(k_x\theta_x + k_y\theta_y + k_z\theta_z)}\,, \quad \mathbf{k} \in \mathcal{N}\,. \tag{6.21}$$

The following abbreviations are used in the rest of the chapter:

$$\mathrm{c}_a = \cos(\theta_a/2) \qquad \text{and} \qquad \mathrm{s}_a = \sin(\theta_a/2)\,, \qquad a = x, y, z\,. \tag{6.22}$$

The proposition below shows that the subspace spanned by a triple of edge modes $\big(\mathbf{u}^{[x]}(\theta)\ \ \mathbf{u}^{[y]}(\theta)\ \ \mathbf{u}^{[z]}(\theta)\big)$ is invariant with respect to the system matrix A if $\theta \in \Theta$, where

$$\Theta = \left\{ \left(\frac{2\pi\ell_x}{N},\ \frac{2\pi\ell_y}{N},\ \frac{2\pi\ell_z}{N} \right)^T \mid \ell_x, \ell_y, \ell_z \in \mathbb{N} \text{ and } 0 \le \ell_x, \ell_y, \ell_z < N \right\}. \tag{6.23}$$

That is, in Fourier basis we have $A = \operatorname{diag}(\mathcal{A}(\theta))$, with $\mathcal{A}(\theta)$ given by (6.25).

Proposition 6.1. *Let A be defined by* (6.5)*, where K_{cc} and M are edge matrices on a $(N+1) \times (N+1) \times (N+1)$ cubic grid satisfying* (6.17) *with stencils in x, y and*

z directions given by (6.18) and (6.19), respectively. Let $\mathbf{u}^{[a]}(\theta)$, $a = x, y, z$ and Θ be defined by (6.20) and (6.23), respectively.

Then, for any $\theta \in \Theta$ there holds

$$A\left(\mathbf{u}^{[x]}(\theta)\;\; \mathbf{u}^{[y]}(\theta)\;\; \mathbf{u}^{[z]}(\theta)\right) = \left(\mathbf{u}^{[x]}(\theta)\;\; \mathbf{u}^{[y]}(\theta)\;\; \mathbf{u}^{[z]}(\theta)\right)\mathcal{A}(\theta) \tag{6.24}$$

where

$$\mathcal{A}(\theta) = \alpha\mathcal{K}_{cc}(\theta) + \beta h^2\mathcal{M}(\theta) \tag{6.25}$$

with

$$\mathcal{K}_{cc}(\theta) = \frac{4}{3}\begin{pmatrix} 3\mathrm{s}_y^2 + 3\mathrm{s}_z^2 - 4\mathrm{s}_y^2\mathrm{s}_z^2 & -\mathrm{s}_x\mathrm{s}_y(3 - 2\mathrm{s}_z^2) & -\mathrm{s}_x(3 - 2\mathrm{s}_y^2)\mathrm{s}_z \\ -\mathrm{s}_x\mathrm{s}_y(3 - 2\mathrm{s}_z^2) & 3\mathrm{s}_x^2 + 3\mathrm{s}_z^2 - 4\mathrm{s}_x^2\mathrm{s}_z^2 & -(3 - 2\mathrm{s}_x^2)\mathrm{s}_y\mathrm{s}_z \\ -\mathrm{s}_x(3 - 2\mathrm{s}_y^2)\mathrm{s}_z & -(3 - 2\mathrm{s}_x^2)\mathrm{s}_y\mathrm{s}_z & 3\mathrm{s}_x^2 + 3\mathrm{s}_y^2 - 4\mathrm{s}_x^2\mathrm{s}_y^2 \end{pmatrix} \tag{6.26}$$

$$\mathcal{M}(\theta) = \frac{1}{9}\begin{pmatrix} (3 - 2\mathrm{s}_y^2)(3 - 2\mathrm{s}_z^2) & & \\ & (3 - 2\mathrm{s}_x^2)(3 - 2\mathrm{s}_z^2) & \\ & & (3 - 2\mathrm{s}_x^2)(3 - 2\mathrm{s}_y^2) \end{pmatrix} \tag{6.27}$$

and with s_a, $a = x, y, z$, given by (6.22).

Proof. Note that, using (6.17), (6.18) and (6.19), we have

$$\begin{aligned}
A\mathbf{u}^{[z]}(\theta) = {} & \alpha\frac{1}{3}\Big(8 - e^{i\theta_x} - e^{-i\theta_x} - e^{i\theta_y} - e^{-i\theta_y} \\
& - e^{i(\theta_x+\theta_y)} - e^{i(-\theta_x+\theta_y)} - e^{i(\theta_x-\theta_y)} - e^{-i(\theta_x+\theta_y)}\Big)\,\mathbf{u}^{[x]}(\theta) \\
& + \alpha\frac{1}{6}\Big(e^{-i\theta_z/2}(4e^{-i\theta_y/2} - 4e^{i\theta_y/2} - e^{i(\theta_y/2+\theta_x)} + e^{i(-\theta_y/2+\theta_x)} - e^{i(\theta_y/2-\theta_x)} + e^{i(-\theta_y/2-\theta_x)}) \\
& + e^{i\theta_z/2}(4e^{i\theta_y/2} - 4e^{-i\theta_y/2} + e^{i(\theta_y/2+\theta_x)} - e^{i(-\theta_y/2+\theta_x)} + e^{i(\theta_y/2-\theta_x)} - e^{i(-\theta_y/2-\theta_x)})\Big)\,\mathbf{u}^{[y]}(\theta) \\
& + \alpha\frac{1}{6}\Big(e^{-i\theta_z/2}(4e^{-i\theta_x/2} - 4e^{i\theta_x/2} - e^{i(\theta_x/2+\theta_y)} + e^{i(-\theta_x/2+\theta_y)} - e^{i(\theta_x/2-\theta_y)} + e^{i(-\theta_x/2-\theta_y)}) \\
& + e^{i\theta_z/2}(4e^{i\theta_x/2} - 4e^{-i\theta_x/2} + e^{i(\theta_x/2+\theta_y)} - e^{i(-\theta_x/2+\theta_y)} + e^{i(\theta_x/2-\theta_y)} - e^{i(-\theta_x/2-\theta_y)})\Big)\,\mathbf{u}^{[z]}(\theta) \\
& + \beta\frac{1}{36}h^2\Big(16 + 4e^{i\theta_x} + 4e^{-i\theta_x} + 4e^{i\theta_y} + 4e^{-i\theta_y} \\
& + +e^{i(\theta_x+\theta_y)} + e^{i(-\theta_x+\theta_y)} + e^{i(\theta_x-\theta_y)} + e^{-i(\theta_x+\theta_y)}\Big)\,\mathbf{u}^{[x]}(\theta)
\end{aligned}$$

and, after some tedious trigonometry, the last column of (6.26) and (6.27) follows. The other lines are determined similarly. ■

Regarding the edge-node incidence matrix, the general expression (6.6) can be further rewritten for the considered grid as

$$(G)_{\mathbf{k}_e\mathbf{k}_n} = \begin{cases} 1 & \text{if } \mathbf{k}_n = \langle \mathbf{k}_e + (\tfrac{1}{2},0,0)\rangle\,, \langle \mathbf{k}_e + (0,\tfrac{1}{2},0)\rangle \text{ or } \langle \mathbf{k}_e + (0,0,\tfrac{1}{2})\rangle \\ -1 & \text{if } \mathbf{k}_n = \langle \mathbf{k}_e - (\tfrac{1}{2},0,0)\rangle\,, \langle \mathbf{k}_e - (0,\tfrac{1}{2},0)\rangle \text{ or } \langle \mathbf{k}_e - (0,0,\tfrac{1}{2})\rangle \\ 0 & \text{otherwise}\,. \end{cases} \tag{6.28}$$

The following theorem gives the Fourier representation of G.

Proposition 6.2. *Let G be defined by* (6.28) *on a* $(N+1)\times(N+1)\times(N+1)$ *cubic grid. Let $\mathbf{u}^{[a]}(\theta)$, $a = x, y, z$, $\mathbf{u}(\theta)$ and Θ be defined by* (6.20), (6.21) *and* (6.23), *respectively.*

Then, for any $\theta \in \Theta$ there holds

$$G^T \left(\mathbf{u}^{[x]}(\theta) \;\; \mathbf{u}^{[y]}(\theta) \;\; \mathbf{u}^{[z]}(\theta) \right) = 2i\, \mathbf{u}(\theta)(\mathrm{s}_x \;\, \mathrm{s}_y \;\, \mathrm{s}_z) \tag{6.29}$$

and with s_a, $a = x, y, z$, *given by* (6.22).

Proof. Note that for any $k \in \mathcal{N}$ there holds

$$\begin{aligned} (G^T\mathbf{u}^{[x]}(\theta))_{\mathbf{k}} &= \frac{1}{\sqrt{N^3}} \left(e^{i((k_x+1/2)\theta_x + k_y\theta_y + k_z\theta_z)} - e^{i((k_x-1/2)\theta_x + k_y\theta_y + k_z\theta_z)} \right) \\ &= \left(e^{i\theta_x/2} - e^{-i\theta_x/2} \right) (\mathbf{u}(\theta))_{\mathbf{k}} \\ &= 2i \sin(\theta_x/2)\, (\mathbf{u}(\theta))_{\mathbf{k}}\,, \end{aligned}$$

which gives the first entry of (6.29). The proof for the other entries is similar. ■

Now, we assume that N is even and consider two types of aggregation patterns:

(xy) we aggregate nodes into squares in xy-plane, leading to

$$\Gamma^{xy}_{\mathbf{k}} = \{(2k_x, 2k_y, k_z),\ (2k_x{+}1, 2k_y, k_z),\ (2k_x, 2k_y{+}1, k_z),\ (2k_x{+}1, 2k_y{+}1, k_z)\}$$

with k_x, k_y, k_z being integer and such that $0 \le 2k_x, 2k_y, k_z < N$.

(xyz) we aggregate nodes into cubes by grouping the nodes

$$\begin{aligned} \Gamma^{xyz}_{\mathbf{k}} = \{&(2k_x, 2k_y, 2k_z),\ (2k_x{+}1, 2k_y, 2k_z),\ (2k_x, 2k_y{+}1, 2k_z),\ (2k_x{+}1, 2k_y{+}1, 2k_z), \\ &(2k_x, 2k_y, 2k_z{+}1),\ (2k_x{+}1, 2k_y, 2k_z{+}1),\ (2k_x, 2k_y{+}1, 2k_z{+}1),\ (2k_x{+}1, 2k_y{+}1, 2k_z{+}1)\}\ , \end{aligned}$$

with k_x, k_y, k_z being integer and such that $0 \le 2k_x, 2k_y, 2k_z < N$.

We extend our coordinate notation to the coarse grid edge unknowns, letting

$$\mathcal{I}_\ell(\Delta_x,\Delta_y,\Delta_z) = \begin{cases} \{(k_x{+}\Delta_x, k_y{+}\Delta_y, k_z{+}\Delta_z | 0 \leq 2k_x, 2k_y, k_z < N)\} & \text{if } \ell = xy\,, \\ \{(k_x{+}\Delta_x, k_y{+}\Delta_y, k_z{+}\Delta_z | 0 \leq 2k_x, 2k_y, 2k_z < N)\} & \text{if } \ell = xyz\,, \end{cases}$$

and setting their index set to $\mathcal{E}_\ell^{[x]} = \mathcal{I}_\ell(1/2,0,0)$, $\mathcal{E}_\ell^{[y]} = \mathcal{I}_\ell^{[z]}(0,1/2,0)$ and $\mathcal{E}_\ell^{[z]} = \mathcal{I}_\ell(0,0,1/2)$, $\ell = xy, xyz$, for edges oriented in x, y and z direction, respectively.

The edge prolongation defined by (6.7) is then given for any $\mathbf{k} = (k_x, k_y, k_z)$ by

$$(\,P_{xy}^{(e)\,T}\mathbf{w})_\mathbf{k} = \begin{cases} (\mathbf{w})_{\mathbf{k}_1} + (\mathbf{w})_{\mathbf{k}_2}\,, & \\ \quad \mathbf{k}_1 = (2k_x{+}1/2, 2k_y{+}1, k_z), \mathbf{k}_2 = (2k_x{+}1/2, 2k_y, k_z) & \text{if } \mathbf{k} \in \mathcal{E}_{xy}^{[x]}\,, \\ (\mathbf{w})_{\mathbf{k}_1} + (\mathbf{w})_{\mathbf{k}_2}\,, & \\ \quad \mathbf{k}_1 = (2k_x{+}1, 2k_y{+}1/2, k_z),\ \mathbf{k}_2 = (2k_x, 2k_y{+}1/2, k_z) & \text{if } \mathbf{k} \in \mathcal{E}_{xy}^{[y]}\,, \\ (\mathbf{w})_{\mathbf{k}_1} + (\mathbf{w})_{\mathbf{k}_2} + (\mathbf{w})_{\mathbf{k}_3} + (\mathbf{w})_{\mathbf{k}_4}\,, & \\ \quad \mathbf{k}_1 = (2k_x{+}1, 2k_y{+}1, k_z)\,, \mathbf{k}_2 = (2k_x{+}1, 2k_y, k_z)\,, & \\ \quad \mathbf{k}_3 = (2k_x, 2k_y{+}1, k_z)\,, \mathbf{k}_4 = (2k_x, 2k_y, k_z) & \text{if } \mathbf{k} \in \mathcal{E}_{xy}^{[z]}\,. \end{cases} \tag{6.30}$$

in **(xy)** case and by

$$(\,P_{xyz}^{(e)T}\,\mathbf{w})_\mathbf{k} = \begin{cases} (\mathbf{w})_{\mathbf{k}_1} + (\mathbf{w})_{\mathbf{k}_2} + (\mathbf{w})_{\mathbf{k}_3} + (\mathbf{w})_{\mathbf{k}_4}\,, & \\ \quad \mathbf{k}_1 = (2k_x{+}1/2, 2k_y{+}1, 2k_z{+}1),\ \mathbf{k}_2 = (2k_x{+}1/2, 2k_y, 2k_z), & \\ \quad \mathbf{k}_3 = (2k_x{+}1/2, 2k_y, 2k_z{+}1),\ \mathbf{k}_4 = (2k_x{+}1/2, 2k_y{+}1, 2k_z) & \text{if } \mathbf{k} \in \mathcal{E}_{xyz}^{[x]}\,, \\ (\mathbf{w})_{\mathbf{k}_1} + (\mathbf{w})_{\mathbf{k}_2} + (\mathbf{w})_{\mathbf{k}_3} + (\mathbf{w})_{\mathbf{k}_4}\,, & \\ \quad \mathbf{k}_1 = (2k_x{+}1, 2k_y{+}1/2, 2k_z{+}1),\ \mathbf{k}_2 = (2k_x, 2k_y{+}1/2, 2k_z), & \\ \quad \mathbf{k}_3 = (2k_x, 2k_y{+}1/2, 2k_z{+}1),\ \mathbf{k}_4 = (2k_x{+}1, 2k_y{+}1/2, 2k_z) & \text{if } \mathbf{k} \in \mathcal{E}_{xyz}^{[y]}\,, \\ (\mathbf{w})_{\mathbf{k}_1} + (\mathbf{w})_{\mathbf{k}_2} + (\mathbf{w})_{\mathbf{k}_3} + (\mathbf{w})_{\mathbf{k}_4}\,, & \\ \quad \mathbf{k}_1 = (2k_x{+}1, 2k_y{+}1, 2k_z{+}1/2),\ \mathbf{k}_2 = (2k_x, 2k_y, 2k_z{+}1/2), & \\ \quad \mathbf{k}_3 = (2k_x, 2k_y{+}1, 2k_z{+}1/2),\ \mathbf{k}_4 = (2k_x{+}1, 2k_y, 2k_z{+}1/2) & \text{if } \mathbf{k} \in \mathcal{E}_{xyz}^{[z]} \end{cases} \tag{6.31}$$

in **(xyz)** case. Further, we define the coarse grid Fourier modes for $\ell = xy, xyz$, and $a = x, y, z$, as

$$\left(\mathbf{u}_\ell^{[a]}(\theta)\right)_\mathbf{k} = \begin{cases} e^{i\,\mathbf{k}\theta} = \frac{1}{\sqrt{N^3}} e^{i(k_x\theta_x + k_y\theta_y + k_z\theta_z)} & \text{if } \mathbf{k} \in \mathcal{E}_\ell^{[a]} \\ 0 & \text{otherwise}\,. \end{cases} \tag{6.32}$$

As shown in the following proposition, the frequency aliasing is then such that all Fourier modes (6.20) corresponding to the frequencies in $\Theta_\ell(\theta)$, $\ell = xy, xyz$, lead to a unique frequency on the coarse grid, with

$$\Theta_{xy}(\theta) = \left((\theta_x,\,\theta_y,\,\theta_z)^T,\,(\theta_x+\pi,\,\theta_y,\,\theta_z)^T,\,(\theta_x,\,\theta_y+\pi,\,\theta_z)^T,\,(\theta_x+\pi,\,\theta_y+\pi,\,\theta_z)^T\right)\,, \tag{6.33}$$

and

$$\Theta_{xyz}(\theta) = \left((\theta_x,\,\theta_y,\,\theta_z)^T,\,(\theta_x{+}\pi,\,\theta_y,\,\theta_z)^T,\,(\theta_x,\,\theta_y{+}\pi,\,\theta_z)^T,\,(\theta_x{+}\pi,\,\theta_y{+}\pi,\,\theta_z)^T\right.$$

$$(\theta_x,\,\theta_y,\,\theta_z{+}\pi)^T,\,(\theta_x{+}\pi,\,\theta_y,\,\theta_z{+}\pi)^T,\,(\theta_x,\,\theta_y{+}\pi,\,\theta_z{+}\pi)^T,\,(\theta_x{+}\pi,\,\theta_y{+}\pi,\,\theta_z{+}\pi)^T\big)\ . \quad (6.34)$$

Proposition 6.3. *Let $P_{xy}^{(e)}$, Θ_{xy} and $P_{xyz}^{(e)}$, Θ_{xyz} be defined by (6.30), (6.33) and by (6.31), (6.34), respectively. Let $\mathbf{u}^{[a]}(\theta)$ and $\mathbf{u}_\ell^{[a]}(\theta)$, $a = x, y, z$, $\ell = xy, xyz$, be defined by (6.20) and (6.32) and set s_a and c_a as in (6.22).*

Then, for any $(\theta_1, \theta_2, \theta_3, \theta_4) = \Theta_{xy}(\frac{2\pi\ell_x}{N},\ \frac{2\pi\ell_y}{N},\ \frac{2\pi\ell_z}{N})$, $\ell_x, \ell_y, \ell_z \in \mathbb{N}$, and for any $a = x, y, z$, there holds

$$P_{xy}^{(e)}{}^T\left(\mathbf{u}^{[a]}(\theta_1)\ \ \mathbf{u}^{[a]}(\theta_2)\ \ \mathbf{u}^{[a]}(\theta_3)\ \ \mathbf{u}^{[a]}(\theta_4)\right) = \mathbf{u}_{xy}^{[a]}(\theta_c)\ \mathcal{P}_{xy}(\theta_c)^H\ ,$$

where $\theta_c = \left(2\frac{2\pi\ell_x}{N},\ 2\frac{2\pi\ell_y}{N},\ \frac{2\pi\ell_z}{N}\right)^T$ and

$$\mathcal{P}_{xy}(\theta_c)^H = \begin{cases} -2e^{i(\theta_x+\theta_y)/2}\left(-\mathrm{c}_y\ \ i\mathrm{c}_y\ \ i\mathrm{s}_y\ \ \mathrm{s}_y\,\right) & \textit{if } a = x\,, \\ -2e^{i(\theta_x+\theta_y)/2}\left(-\mathrm{c}_x\ \ i\mathrm{s}_x\ \ i\mathrm{c}_x\ \ \mathrm{s}_x\,\right) & \textit{if } a = y\,, \\ -4e^{i(\theta_x+\theta_y)/2}\left(-\mathrm{c}_y\mathrm{c}_x\ \ i\mathrm{c}_y\mathrm{s}_x\ \ i\mathrm{s}_y\mathrm{c}_x\ \ \mathrm{s}_y\mathrm{s}_x\,\right) & \textit{if } a = z\,. \end{cases}$$

Similarly, for any $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8) = \Theta_{xyz}(\frac{2\pi\ell_x}{N},\ \frac{2\pi\ell_y}{N},\ \frac{2\pi\ell_z}{N})$, $\ell_x, \ell_y, \ell_z \in \mathbb{N}$, and for any $a = x, y, z$, there holds

$$P_{xyz}^{(e)}{}^T\left(\mathbf{u}^{[a]}(\theta_1)\ \ \mathbf{u}^{[a]}(\theta_2)\ \ \mathbf{u}^{[a]}(\theta_3)\ \ \mathbf{u}^{[a]}(\theta_4)\ \ \mathbf{u}^{[a]}(\theta_5)\ \ \mathbf{u}^{[a]}(\theta_6)\ \ \mathbf{u}^{[a]}(\theta_7)\ \ \mathbf{u}^{[a]}(\theta_8)\right) = \mathbf{u}_{xyz}^{[a]}(\theta_c)\ \mathcal{P}_{xyz}^{[a]}(\theta_c)^H\ ,$$

where $\theta_c = \left(2\frac{2\pi\ell_x}{N},\ 2\frac{2\pi\ell_y}{N},\ 2\frac{2\pi\ell_z}{N}\right)^T$ and

$$\mathcal{P}_{xyz}^{[a]}(\theta_c)^H = \begin{cases} -4e^{i(\theta_x+\theta_y+\theta_z)/2}\left(-\mathrm{c}_y\mathrm{c}_z\ \ i\mathrm{c}_y\mathrm{c}_z\ \ i\mathrm{s}_y\mathrm{c}_z\ \ \mathrm{s}_y\mathrm{c}_z\ \ i\mathrm{c}_y\mathrm{s}_z\ \ \mathrm{c}_y\mathrm{s}_z\ \ \mathrm{s}_y\mathrm{s}_z\ \ -i\mathrm{s}_y\mathrm{s}_z\,\right) & \textit{if } a = x\,, \\ -4e^{i(\theta_x+\theta_y+\theta_z)/2}\left(-\mathrm{c}_x\mathrm{c}_z\ \ i\mathrm{s}_x\mathrm{c}_z\ \ i\mathrm{c}_x\mathrm{c}_z\ \ \mathrm{s}_x\mathrm{c}_z\ \ i\mathrm{c}_x\mathrm{s}_z\ \ \mathrm{s}_x\mathrm{s}_z\ \ \mathrm{c}_x\mathrm{s}_z\ \ -i\mathrm{s}_x\mathrm{s}_z\,\right) & \textit{if } a = y\,, \\ -4e^{i(\theta_x+\theta_y+\theta_z)/2}\left(-\mathrm{c}_x\mathrm{c}_y\ \ i\mathrm{s}_x\mathrm{c}_y\ \ i\mathrm{c}_x\mathrm{s}_y\ \ \mathrm{s}_x\mathrm{s}_y\ \ i\mathrm{c}_x\mathrm{c}_y\ \ \mathrm{s}_x\mathrm{c}_y\ \ \mathrm{c}_x\mathrm{s}_y\ \ -i\mathrm{s}_x\mathrm{s}_y\,\right) & \textit{if } a = z\,. \end{cases}$$

Proof. We indicate the proof for $P_{xy}^{(e)}$ when $a = x$, the proof is similar in the other cases. For any $\mathbf{k} \in \mathcal{E}_{xy}^{[x]}$, setting $\theta_1 = (\theta_x, \theta_y, \theta_z)$, we have

$$\begin{aligned}
\left(P_{xy}^{(e)}{}^T\,\mathbf{u}^{[x]}(\theta_1)\right)_{\mathbf{k}} &= e^{i((2k_x+1/2)\theta_x+(2k_y+1)\theta_y+k_z\theta_z)} + e^{i((2k_x+1/2)\theta_x+2k_y\theta_y+k_z\theta_z)} \\
&= e^{i\mathbf{k}\theta_c}e^{i\theta_x/2}(1+e^{i\theta_y}) \\
\left(P_{xy}^{(e)}{}^T\,\mathbf{u}^{[x]}(\theta_2)\right)_{\mathbf{k}} &= e^{i((2k_x+1/2)(\theta_x+\pi)+(2k_y+1)\theta_y+k_z\theta_z)} + e^{i((2k_x+1/2)(\theta_x+\pi)+2k_y\theta_y+k_z\theta_z)} \\
&= e^{i\mathbf{k}\theta_c}e^{i2k_x\pi}e^{i(\theta_x+\pi)/2}(1+e^{i\theta_y}) \\
\left(P_{xy}^{(e)}{}^T\,\mathbf{u}^{[x]}(\theta_3)\right)_{\mathbf{k}} &= e^{i((2k_x+1/2)\theta_x+(2k_y+1)(\theta_y+\pi)+k_z\theta_z)} + e^{i((2k_x+1/2)\theta_x+2k_y(\theta_y+\pi)+k_z\theta_z)} \\
&= e^{i\mathbf{k}\theta_c}e^{i\theta_x/2}(1-e^{i\theta_y}) \\
\left(P_{xy}^{(e)}{}^T\,\mathbf{u}^{[x]}(\theta_4)\right)_{\mathbf{k}} &= e^{i((2k_x+1/2)(\theta_x+\pi)+(2k_y+1)(\theta_y+\pi)+k_z\theta_z)} + e^{i((2k_x+1/2)(\theta_x+\pi)+2k_y\theta_y+k_z\theta_z+)} \\
&= e^{i\mathbf{k}\theta_c}e^{i2k_x\pi}e^{i(\theta_x+\pi)/2}(1-e^{i\theta_y})
\end{aligned}$$

and, since $1+e^{i\theta_y} = \mathrm{c}_y e^{i\theta_y/2}$, $1-e^{i\theta_y} = -i\mathrm{s}_y e^{i\theta_y/2}$, $e^{i\pi/2} = i$ and $e^{i2k_x\pi} = -1$, the desired result follows. ■

Now, note that the Propositions 6.1 and 6.3 imply that both the system matrix A and the coarse grid correction matrix $I - P_\ell^{(e)} \left(P_\ell^{(e)^T} A P_\ell^{(e)} \right)^{-1} P_\ell^{(e)^T} A$, $\ell = xy,\, xyz\,,$ with P_ℓ given either by (6.30) or by (6.31), possess invariant subspaces

$$\left\{ \mathbf{u}^{[a]}(\theta) \,|\, a = x, y, z \text{ and } \theta \in \Theta_\ell(\frac{2\pi\ell_x}{N},\ \frac{2\pi\ell_y}{N},\ \frac{2\pi\ell_z}{N}) \right\} \tag{6.35}$$

with ℓ_x, ℓ_y, ℓ_z being integer. Since the union of such subspaces for $0 \le \ell_x, \ell_y, \ell_z < N$ forms an orthogonal basis in $\mathbb{R}^{3N^3}$, the coarse grid correction matrix has a block diagonal structure in such basis with $m \times m$ blocks ($m = 12$ in the **(xy)** case and $m = 24$ if **(xyz)** is considered). If, in addition, the subspace (6.35) is invariant under the smoothing iteration matrix (6.14), the same conclusion on the block structure holds for the two-grid iteration matrix (6.10); that is, in the Fourier basis $E_{TG} = \text{diag}\left(\Xi_{\ell_x,\ell_y,\ell_z}\right)$ with $\Xi_{\ell_x,\ell_y,\ell_z}$ being a $m \times m$ matrix. The invariance requirement on the smoothing iteration matrix is in turn fulfilled if there exist matrices $\mathcal{R}^{(e)}$ and $\mathcal{R}^{(n)}$ such that

$$R^{(e)} \left(\mathbf{u}^{[x]}(\theta)\ \ \mathbf{u}^{[y]}(\theta)\ \ \mathbf{u}^{[z]}(\theta) \right) = \left(\mathbf{u}^{[x]}(\theta)\ \ \mathbf{u}^{[y]}(\theta)\ \ \mathbf{u}^{[z]}(\theta) \right) \mathcal{R}^{(e)}(\theta)\,, \tag{6.36}$$

$$R^{(n)} \mathbf{u}(\theta) = \mathbf{u}(\theta) \mathcal{R}^{(n)}(\theta)\,. \tag{6.37}$$

It is then possible to assess the two-grid convergence factor via the spectral radii of $\Xi_{\ell_x,\ell_y,\ell_z}$, namely

$$\rho\left(E_{TG}\right) = \max_{\ell_x,\ell_y,\ell_z} \rho\left(\Xi_{\ell_x,\ell_y,\ell_z}\right)\,.$$

If both $R^{(e)}$ and $R^{(n)}$ are of Jacobi type; that is, if

$$R^{(e)} = \frac{1}{\omega^{(e)}}\,\text{diag}(A)\,, \tag{6.38}$$

$$R^{(n)} = \frac{1}{\omega^{(n)}}\,\text{diag}(A^{(n)})\,, \tag{6.39}$$

then (6.36) and (6.37) hold with

$$\mathcal{R}^{(e)}(\theta) = \frac{1}{\omega^{(e)}} \left(\alpha \frac{16}{6} + \beta h^2 \frac{16}{36} \right) I_m\,,$$

$$\mathcal{R}^{(n)}(\theta) = \frac{1}{\omega^{(n)}}\,\text{diag}(A^{(n)}) = \frac{1}{\omega^{(n)}}\,\text{diag}(G^T M G) = \frac{1}{\omega^{(n)}}\,\beta h^2 \frac{8}{3}\,,$$

the second equality of the second line coming from $K_{cc}G = O$.

If $R^{(e)}$ and $R^{(n)}$ are of Gauss-Seidel type; that is, if $R^{(e)}$ and $R^{(n)}$ are (up to some reordering of unknowns) upper (or lower) triangular part of A and $G^T A G$, respectively, then the relations (6.36) and (6.37) are not satisfied. However, following the usual practice [61, 68, 8], we can approximate these matrices by $\widetilde{R}^{(e)}$ and $\widetilde{R}^{(n)}$ which satisfy (6.36) and (6.37), respectively.

In particular, since $A^{(n)} = G^T MG$ has the following three-dimensional stencil

$$\left[\begin{bmatrix} -1 & -2 & -1 \\ -2 & 0 & -2 \\ -1 & -2 & -1 \end{bmatrix} \begin{bmatrix} -2 & 0 & -2 \\ 0 & 32 & 0 \\ -2 & 0 & -2 \end{bmatrix} \begin{bmatrix} -1 & -2 & -1 \\ -2 & 0 & -2 \\ -1 & -2 & -1 \end{bmatrix}\right],$$

which also correspond to a trilinear discretization of Poisson equation, the stencil of $\tilde{R}^{(n)}$ can be chosen as

$$\mathcal{ST}(\tilde{R}^{(n)}) = \frac{1}{12}\left[\begin{bmatrix} \cdot \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 32 & 0 \\ -2 & 0 & -2 \end{bmatrix} \begin{bmatrix} -1 & -2 & -1 \\ -2 & 0 & -2 \\ -1 & -2 & -1 \end{bmatrix}\right],$$

if the nodal unknowns are updated in lexicographical order. Then,

$$\mathcal{R}^{(n)}(\theta) = \frac{1}{12}(32 - 4e^{-i\theta_z}\left(3 - 3\mathrm{s}_x^2 - 3\mathrm{s}_y^2 + 4\mathrm{s}_x^2\mathrm{s}_y^2\right) - 4e^{-i\theta_y}(1 - 2\mathrm{s}_x^2))\,,$$

with $\theta = (\theta_x,\, \theta_y,\, \theta_z) \in \Theta$.

For the edge Gauss-Seidel smoother, different strategies can be considered, depending on the order in which the edge unknowns are updated.

direction-based strategy: edges in x direction are updated before those in y direction, which in turn are updated before those in z direction; the ordering inside each direction is lexicographical.

point-based strategy: edges that are associated to a particular node are updated one after another (if not associated to an already updated node; that is, if not already updated) ; the nodes are considered in lexicographical order.

The first strategy can by approximated by the stencil

$$\mathcal{ST}^{[x]}(\widetilde{R}^{(e)}) = \frac{1}{6}\alpha \left[\begin{bmatrix} \bullet \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ -2 & 16 & 0 \\ -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix}\right]$$

$$+ \frac{1}{36}\beta h^2 \left[\begin{bmatrix} \bullet \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 4 & 16 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix}\right],$$

in x direction, by

$$\mathcal{ST}^{[y]}(\widetilde{R}^{(e)}) = \frac{1}{6}\alpha \left[\begin{bmatrix} \circ & 1 & \circ & -1 & \circ \\ 0 & & 0 & & 0 \\ \circ & 4 & \bullet & -4 & \circ \\ 0 & & 0 & & 0 \\ \circ & 1 & \circ & -1 & \circ \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ -2 & 16 & 0 \\ & & \\ -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} \circ & -1 & \circ & 1 & \circ \\ 0 & & 0 & & 0 \\ \circ & -4 & \bullet & 4 & \circ \\ 0 & & 0 & & 0 \\ \circ & -1 & \circ & 1 & \circ \end{bmatrix}\right] + \frac{1}{36}\beta h^2 \left[\begin{bmatrix} \bullet \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ 4 & 16 & 0 \\ & & \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix}\right],$$

in y direction and by

$$\mathcal{ST}^{[z]}(\widetilde{R}^{(e)}) = \frac{1}{6}\alpha \left[\begin{bmatrix} \circ & 1 & \circ & -1 & \circ \\ -1 & & -4 & & -1 \\ \circ & 4 & \bullet & -4 & \circ \\ 1 & & 4 & & 1 \\ \circ & 1 & \circ & -1 & \circ \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ -2 & 16 & 0 \\ & & \\ -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} \circ & -1 & \circ & 1 & \circ \\ 1 & & 4 & & 1 \\ \circ & -4 & \bullet & 4 & \circ \\ -1 & & -4 & & -1 \\ \circ & -1 & \circ & 1 & \circ \end{bmatrix}\right] + \frac{1}{36}\beta h^2 \left[\begin{bmatrix} \bullet \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ 4 & 16 & 0 \\ & & \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix}\right],$$

in z direction. The corresponding Fourier block can be evaluated as in the proof of Proposition 6.1, and is given by

$$\begin{aligned}
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{11} &= \alpha\frac{1}{3}(8 - e^{-i\theta_y} - e^{-i\theta_z} - 2(1-2\mathrm{s}_y^2)e^{-i\theta_z}) + \beta h^2\frac{1}{9}(4 + e^{-i\theta_y} + e^{-i\theta_z} + \frac{1}{2}(1-2\mathrm{s}_y^2)e^{-i\theta_z}) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{12} &= 0 \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{13} &= 0 \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{21} &= -\alpha\frac{4}{3}\mathrm{s}_x\mathrm{s}_y(3 - 2\mathrm{s}_z^2) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{22} &= \alpha\frac{1}{3}(8 - e^{-i\theta_x} - e^{-i\theta_z} - 2(1-2\mathrm{s}_x^2)e^{-i\theta_z}) + \frac{1}{9}\beta h^2(4 + e^{-i\theta_x} + e^{-i\theta_z} + \frac{1}{2}(1-2\mathrm{s}_x^2)e^{-i\theta_z}) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{23} &= 0 \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{31} &= -\alpha\frac{4}{3}\mathrm{s}_x(3 - 2\mathrm{s}_y^2)\mathrm{s}_z \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{32} &= -\alpha\frac{4}{3}(3 - 2\mathrm{s}_x^2)\mathrm{s}_y\mathrm{s}_z \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{33} &= \alpha\frac{1}{3}(8 - e^{-i\theta_x} - e^{-i\theta_y} - 2(1-2\mathrm{s}_x^2)e^{-i\theta_y}) + \beta h^2\frac{1}{9}(4 + e^{-i\theta_x} + e^{-i\theta_y} + \frac{1}{2}(1-2\mathrm{s}_x^2)e^{-i\theta_y})\,.
\end{aligned}$$

The second strategy corresponds to the stencil

$$\mathcal{S}\mathcal{T}^{[x]}(\widetilde{R}^{(e)}) = \frac{1}{6}\alpha \left[\begin{bmatrix} \circ & 0 & \circ & 0 & \circ \\ -1 & & 0 & & 0 \\ \circ & 4 & \bullet & 0 & \circ \\ 1 & & 4 & & 1 \\ \circ & 1 & \circ & -1 & \circ \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ -2 & 16 & 0 \\ & & \\ -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} \circ & 0 & \circ & 0 & \circ \\ 1 & & 0 & & 0 \\ \circ & -4 & \bullet & 0 & \circ \\ -1 & & -4 & & -1 \\ \circ & -1 & \circ & 1 & \circ \end{bmatrix} \right]$$

$$+ \frac{1}{36}\beta h^2 \left[\begin{bmatrix} \bullet \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ 4 & 16 & 0 \\ & & \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix} \right],$$

in x direction, to the stencil

$$\mathcal{S}\mathcal{T}^{[y]}(\widetilde{R}^{(e)}) = \frac{1}{6}\alpha \left[\begin{bmatrix} \circ & 0 & \circ & 0 & \circ \\ -1 & & 0 & & 0 \\ \circ & 4 & \bullet & -4 & \circ \\ 1 & & 4 & & 1 \\ \circ & 1 & \circ & -1 & \circ \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ -2 & 16 & 0 \\ & & \\ -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} \circ & 0 & \circ & 0 & \circ \\ 0 & & 0 & & 0 \\ \circ & 0 & \bullet & 0 & \circ \\ -1 & & -4 & & -1 \\ \circ & -1 & \circ & 1 & \circ \end{bmatrix} \right]$$

$$+ \frac{1}{36}\beta h^2 \left[\begin{bmatrix} \bullet \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ 4 & 16 & 0 \\ & & \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix} \right],$$

in y direction and to the stencil

$$\mathcal{S}\mathcal{T}^{[z]}(\widetilde{R}^{(e)}) = \frac{1}{6}\alpha \left[\begin{bmatrix} \circ & 0 & \circ & 0 & \circ \\ -1 & & -4 & & 0 \\ \circ & 4 & \bullet & -4 & \circ \\ 1 & & 4 & & 1 \\ \circ & 1 & \circ & -1 & \circ \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ -2 & 16 & 0 \\ & & \\ -2 & -2 & -2 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix} \right]$$

$$+ \frac{1}{36}\beta h^2 \left[\begin{bmatrix} \bullet \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ & & \\ 4 & 16 & 0 \\ & & \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} \bullet \end{bmatrix} \right],$$

in z direction. The corresponding Fourier block is given by

$$(\widetilde{\mathcal{R}}^{(e)}(\theta))_{11} = \alpha\frac{1}{3}(8 - e^{-i\theta_y} - e^{-i\theta_z} - 2(1-2\mathrm{s}_y^2)e^{-i\theta_z}) + \beta h^2\frac{1}{9}(4 + e^{-i\theta_y} + e^{-i\theta_z} + \frac{1}{2}(1-2\mathrm{s}_y^2)e^{-i\theta_z})$$

$$(\widetilde{\mathcal{R}}^{(e)}(\theta))_{12} = \alpha\frac{i}{3}\mathrm{s}_x(-4e^{-i\theta_y/2} + 2i\mathrm{s}_y e^{-i\theta_z})$$

N		$\nu = 1/2$	$\nu = 1$	$\nu = 2$
50	xy.	0.852	0.794	0.755
	xyz.	0.852	0.794	0.755
100	xy.	0.853	0.796	0.759
	xyz.	0.853	0.796	0.759
150	xy.	0.853	0.796	0.760
	xyz.	0.853	0.796	0.760

Table 6.1: Convergence factor of a two-grid method with Jacobi hybrid smoother, estimated via Fourier analysis.

		point-based			direction-based		
N		$\nu = 1/2$	$\nu = 1$	$\nu = 2$	$\nu = 1/2$	$\nu = 1$	$\nu = 2$
50	xy.	0.745	0.770	0.723	0.716	0.739	0.716
	xyz.	0.770	0.777	0.726	0.716	0.740	0.716
100	xy.	0.751	0.782	0.748	0.722	0.752	0.741
	xyz.	0.776	0.789	0.751	0.722	0.753	0.741
150	xy.	0.752	0.785	0.753	0.723	0.755	0.746
	xyz.	0.777	0.792	0.756	0.723	0.755	0.746
200	xy.	0.752	0.786	0.755	0.724	0.756	0.748
	xyz.	0.777	0.793	0.758	0.724	0.756	0.748

Table 6.2: Convergence factor of a two-grid method with various Gauss-Seidel variants of hybrid smoother, estimated via Fourier analysis.

$$
\begin{aligned}
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{13} &= \alpha\frac{i}{3}\mathrm{s}_x(-(4+e^{i\theta_y})e^{-i\theta_z/2} + 2i\mathrm{s}_z e^{-i\theta_y}) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{21} &= \alpha\frac{2}{3}\mathrm{s}_x(-\mathrm{s}_y e^{-i\theta_z} - ie^{-i\theta_y/2}) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{22} &= \alpha\frac{1}{3}(8 - e^{-i\theta_x} - e^{-i\theta_z} - 2(1-2\mathrm{s}_x^2)e^{-i\theta_z}) + \beta h^2\frac{1}{9}(4 + e^{-i\theta_x} + e^{-i\theta_z} + \frac{1}{2}(1-2\mathrm{s}_x^2)e^{-i\theta_z}) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{23} &= \alpha\frac{1}{6}\left(-4i\mathrm{s}_y(3-2\mathrm{s}_x^2)e^{-i\theta_z/2} - e^{-i\theta_y/2}e^{-i\theta_x}e^{i\theta_z/2}\right) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{31} &= \alpha\frac{i}{3}\mathrm{s}_x e^{-\theta_z/2}(-4 - e^{-i\theta_y}) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{32} &= \alpha\frac{1}{6}e^{-i\theta_z/2}(e^{i\theta_x}e^{-i\theta_y/2} - 2i\mathrm{s}_y(4 - e^{-i\theta_x})) \\
(\widetilde{\mathcal{R}}^{(e)}(\theta))_{33} &= \alpha\frac{1}{3}(8 - e^{-i\theta_x} - e^{-i\theta_y} - 2(1-2\mathrm{s}_x^2)e^{-i\theta_y}) + \beta h^2\frac{1}{9}(4 + e^{-i\theta_x} + e^{-i\theta_y} + \frac{1}{2}(1-2\mathrm{s}_x^2)e^{-i\theta_y})\,.
\end{aligned}
$$

6.4 Numerical results

6.4.1 Two-grid method

For the numerical investigations that follow, we set $\alpha = 1$ and $\beta = 0.01$. When Jacobi smoothers (6.38) and (6.39) are considered, the weights are chosen to be $\omega^{(e)} = 1/3$ and $\omega^{(n)} = 2/3$. This choice corresponds to the biggest values of weights such that the iteration matrices $I - {R^{(e)}}^{-1}A$ and $I - {R^{(n)}}^{-1}A$ are still positive definite for any N.

		$\nu = 1/2$		$\nu = 1$		$\nu = 2$	
N		FA	D	FA	D	FA	D
20	xy.	0.845	0.825	0.779	0.743	0.725	0.671
	xyz.	0.845	0.844	0.779	0.775	0.725	0.721
30	xy.	0.850	0.835	0.788	0.763	0.744	0.707
	xyz.	0.850	0.849	0.788	0.786	0.744	0.741
40	xy.	0.851	0.840	0.792	0.772	0.752	0.721
	xyz.	0.851	0.851	0.792	0.791	0.752	0.749

TABLE 6.3: Comparison of Fourier analysis and actual two-grid convergence factors in the case of Jacobi hybrid smoother.

		$\nu = 1/2$			$\nu = 1$			$\nu = 2$		
N		FA	P	D	FA	P	D	FA	P	D
20	xy.	0.711	0.640	0.600	0.689	0.683	0.646	0.573	0.587	0.583
	xyz.	0.733	0.680	0.681	0.702	0.698	0.717	0.583	0.600	0.659
30	xy.	0.733	0.675	0.639	0.741	0.746	0.693	0.667	0.678	0.636
	xyz.	0.757	0.714	0.715	0.749	0.746	0.746	0.672	0.683	0.706
40	xy.	0.741	0.687	0.662	0.760	0.750	0.715	0.705	0.711	0.672
	xyz.	0.766	0.727	0.710	0.768	0.763	0.758	0.708	0.715	0.724

TABLE 6.4: Comparison of Fourier analysis and actual two-grid convergence factors for point-based Gauss-Seidel hybrid smoother.

We first consider Fourier analysis for large problem sizes. The corresponding results are given in Table 6.1 for the Jacobi version of the hybrid smoother and in Table 6.2 for the different variants of its Gauss-Seidel version. The asymptotical values of the convergence are (approximately) reached for $N = 100$ in the former case and for $N = 150$ in the latter. In both cases, the (almost) asymptotical values are bounded away from 1, showing that RS approach has h-independent convergence properties in two-grid setting.

Note that periodic boundary conditions are rarely used in practice, their main purpose here is to make the exact Fourier analysis possible. It is therefore instructive to compare previous results with convergence factors of similar problems with realistic boundary conditions. Here, the comparison is made with the problem (6.2) having Dirichlet boundary conditions and discretized on the cubic grid $(N+2)\times(N+2)\times(N+2)$ of mesh size $h = (N+1)^{-1}$. Observe that, assuming the same number of unknowns, this value of h differs slightly from the one defined in periodic case.

Now, the convergence factors for problems with periodic (FA) and Dirichlet (D) boundary conditions are given in Table 6.3 for the Jacobi hybrid smoother. Tables 6.4 and 6.5 present the same information for smoothers of Gauss-Seidel type. In both cases, we have evaluated real convergence factors using ARPACK [36] routines. Note that, when a Gauss-Seidel smoother is considered, the corresponding matrices $R^{(e)}$ and $R^{(n)}$ are approximated by $\widetilde{R}^{(e)}$ and $\widetilde{R}^{(n)}$ in order for Fourier analysis to be applicable. That

N		$\nu = 1/2$			$\nu = 1$			$\nu = 2$		
		FA	P	D	FA	P	D	FA	P	D
20	xy.	0.673	0.662	0.602	0.479	0.642	0.599	0.560	0.538	0.537
	xyz.	0.675	0.665	0.668	0.659	0.655	0.670	0.568	0.554	0.608
30	xy.	0.701	0.698	0.652	0.709	0.702	0.652	0.658	0.651	0.603
	xyz.	0.701	0.698	0.697	0.710	0.702	0.697	0.660	0.657	0.679
40	xy.	0.711	0.709	0.671	0.730	0.726	0.685	0.697	0.694	0.643
	xyz.	0.711	0.709	0.709	0.730	0.726	0.729	0.698	0.697	0.707
50	xy.	0.716	0.715	0.682	0.739	0.739	0.702	0.716	0.714	0.670
	xyz.	0.716	0.715	0.713	0.739	0.741	0.739	0.716	0.716	0.720

TABLE 6.5: Comparison of Fourier analysis and actual two-grid convergence factors for direction-based Gauss-Seidel hybrid smoother.

γ	point-based GS			direction-based GS			Jacobi		
	$\nu = 1/2$	$\nu = 1$	$\nu = 2$	$\nu = 1/2$	$\nu = 1$	$\nu = 2$	$\nu = 1/2$	$\nu = 1$	$\nu = 2$
1	0.777	0.792	0.756	0.723	0.755	0.746	0.853	0.796	0.759
0.8	0.737	0.755	0.701	0.639	0.697	0.684	0.837	0.760	0.704
0.6	0.686	0.706	0.614	0.472	0.605	0.581	0.819	0.712	0.618
0.5	0.659	0.675	0.551	0.447	0.536	0.499	0.809	0.682	0.555
0.4	0.644	0.642	0.376	0.594	0.444	0.376	0.798	0.667	0.474
0.3	0.854	0.809	0.673	0.854	0.809	0.673	1.166	0.667	0.667

TABLE 6.6: Dependence of convergence rate on γ for **(xyz)** prolongation and various smoothers.

is why in this latter case the convergence assessed via Fourier analysis (FA) does not coincide with the actual two-grid convergence for problem with periodic (P) boundary conditions.

Regarding the values in these three tables, we observe that the Fourier analysis seems to give an accurate, although sometimes pessimistic, estimate of the real two-grid convergence. More generally, in view of all results presented so far it appears that Gauss-Seidel implementations are superior to the Jacobi ones; the difference between point-based and direction-based variants of Gauss-Seidel is small, the latter performing globally better. We also observe that the use of more smoothing steps does not necessarily pay off, and in some cases it can even slightly deteriorates the convergence; the use of $\nu = 1/2$ (or $\nu_1, \nu_2 = 1/2$ in case of symmetric multigrid method) seems to be a good choice. Regarding the prolongations considered, the performance of **(xy)** and **(xyz)** variants is similar, the latter being more attractive because of the faster decrease in the size of coarser grid(s).

It is observed in [11] in the context of aggregation-based multigrid for Poisson-like problems that a simple way to improve convergence is to use a weighted coarse grid

γ	point-based GS		direction-based GS		Jacobi	
	$\nu_1, \nu_2 = 1/2$	$\nu_1, \nu_2 = 1$	$\nu_1, \nu_2 = 1/2$	$\nu_1, \nu_2 = 1$	$\nu_1, \nu_2 = 1/2$	$\nu_1, \nu_2 = 1$
1	4.80	4.10	4.08	3.94	4.90	4.15
0.8	4.26	3.36	3.45	3.19	4.43	3.43
0.6	3.82	2.32	2.82	2.43	4.05	2.73
0.5	3.70	2.32	2.56	2.08	3.94	2.39
0.4	3.93	2.38	2.53	2.02	3.92	2.37
0.3	4.58	2.67	2.68	2.04	4.23	2.66

TABLE 6.7: Dependence of condition number on γ for **(xyz)** prolongation and various smoothers.

correction instead of the usual one. The same observations hold in the present edge-based two-grid setting, replacing (6.10) by

$$E_{TG} = S_{\downarrow}^{(\nu_2)} \left(I - \gamma^{-1} P^{(e)} \left({P^{(e)}}^T A P^{(e)} \right)^{-1} {P^{(e)}}^T A \right) S_{\uparrow}^{(\nu_1)} ,$$

as can be seen in **(xyz)** case from Table 6.6. Since the relation (6.15) is not necessary satisfied for $\gamma \neq 1$, we report in Table 6.7 the variation of condition number with the weighting factor. Note that the optimal value $\gamma \approx 0.4$ of the weighting parameter is almost independent of the smoother (except for the condition number in case of point-based Gauss-Seidel), and leads to a substantial decrease in the convergence rate (by a factor of two or more for both Gauss-Seidel variants) and slightly less substantial decrease in the condition number.

6.4.2 Multigrid implementation

We now consider the multigrid implementation of the RS algorithm. The convergence behaviour of the method is investigated for V- and W-cycles [61], as well as for the Krylov-based cycling strategy [49]. This latter is implemented as in Algorithm 3.2 from [48], with flexible conjugated gradient (FCG) acceleration at every level and with $t = 0$ (that is, exactly two FCG iterations are performed). Since the choice of FCG is relevant if the preconditioner is symmetric, we set $\nu_1 = \nu_2$ in what follows. The resulting multigrid method is itself used on the finest grid as a preconditioner for the FCG(1) method from [45] (which amounts to standard conjugate gradient method in the case of V- and W-cycles).

In what follows we consider **(xyz)** prolongation with a direction-based Gauss-Seidel smoother as the most interesting combination. This case is supplemented with the Jacobi hybrid smoother to illustrate the effect of a less efficient smoothing scheme. The iteration counts for the three cycling strategies are given in Table 6.8 for the periodic case and in Table 6.9 for the Dirichlet one. In all cases the iterations counts are obtained using 10 randomly chosen right hand sides (the same for three cycling strategies) and

	Jacobi						direction-based GS					
nbr.	ν_1, $\nu_2 = 1/2$			ν_1, $\nu_2 = 1$			ν_1, $\nu_2 = 1/2$			ν_1, $\nu_2 = 1$		
grids	V	W	K	V	W	K	V	W	K	V	W	K
2	22	22	22	19	19	19	18-19	18-19	18-19	16	16	16
3	35-36	30	24-25	29	24	20-21	25-26	22	19	21	17-18	16
4	43-44	36-37	25	33	27-28	20-21	29	24	19	25	20	16
5	51	41	25-26	38	31	21-22	35	28-29	19	28-29	24	16

TABLE 6.8: Iteration counts for various cycling strategies; periodic boundary conditions are considered; the finest grid corresponds to $N = 33$ (98304 edge unknowns).

	Jacobi						direction-based GS					
nbr.	ν_1, $\nu_2 = 1/2$			ν_1, $\nu_2 = 1$			ν_1, $\nu_2 = 1/2$			ν_1, $\nu_2 = 1$		
grids	V	W	K	V	W	K	V	W	K	V	W	K
2	23	23	23	20-21	20-21	20-21	19	19	19	16	16	16
3	39	32-33	28	32-33	27	23	28	23-24	20	22-23	19	17-18
4	51-52	42	29	40-41	33	23	31	26-27	20	23-24	19-20	17-18
5	57-58	45-47	30	42-43	34-35	23	31-32	27	20	24	20-21	17-18

TABLE 6.9: Iteration counts for various cycling strategies; Dirichlet boundary conditions are considered; the finest grid corresponds to $N = 34$ (101376 edge unknowns).

reducing the residual by a factor of 10^{10}. Regarding the results for the Gauss-Seidel case we conclude that, at least for the considered problem, the K-cycle multigrid converges in almost the same number of iterations as the two-grid cycle implemented on the finest grid. A slight increase is observed in the case of the Jacobi smoother, which is however less pronounced than the one for V- and W-cycles.

6.5 Conclusion

We have performed the Fourier analysis of Reitzinger and Schöberl multigrid approach on 3D curl-curl problems discretized with edge finite elements. We have shown that the approach has level-independent convergence properties for various smoother configurations and aggregates' shapes. We have compared the results of the analysis with the convergence rate of similar model problems and observed that the former give accurate estimates of the later. We have observed that a few iterations of the Gauss-Seidel hybrid smoother combined with the cubwise aggregation coarsening leads to a good compromise between resource requirements and convergence speed. In multi-level setting, we have observed that an almost level-independent convergence can be recovered when using K-cycle.

List of Figures

List of Tables

Bibliography

[1] Antonio Aricò, Marco Donatelli, and Stefano Serra-Capizzano. V-cycle optimal convergence for certain (multilevel) structured linear systems. *SIAM J. Matrix Anal. Appl.*, 26:186–214, 2004.

[2] Douglas N Arnold, Richard S Falk, and Ragnar Winther. Multigrid in $H(\text{div})$ and $H(\text{curl})$. *Numerische Mathematik*, 85(2):197–217, 2000.

[3] O. Axelsson. *Iterative Solution Methods.* Cambridge University Press, Cambridge, 1994.

[4] R. Beck. Algebraic multigrid by components splitting for edge elements on simplicial triangulations. *Preprint SC 99-40, ZIB*, December 1999.

[5] Rudolf Beck, Peter Deuflhard, Ralf Hiptmair, Ronald H.W. Hoppe, and Barbara Wohlmuth. Adaptative multilevel methods for edge element discretizations of Maxwell's equations. *Surveys on Mathematics for Industry*, 8:271–312, 1999.

[6] A. Ben Israel and T. N. E. Greville. *Generalized Inverses : theory and applications.* J. Wiley and Sons, New York, 1974.

[7] M. Benzi. Preconditioning techniques for large linear systems: A survey. *J. Computational Physics*, 182:418–477, 2002.

[8] Tim Boonen. *Multigrid algorithms for electromagnetic field computations.* PhD thesis, Katholieke Universiteit Leuven, Belgium, 2008.

[9] Tim Boonen, Jan Van lent, and Stefan Vandewalle. Local Fourier analysis of multigrid for the curl-curl equation. *SIAM Journal on Scientific Computing*, 30(4):1730–1755, 2008.

[10] D. Braess. The convergence rate of a multigrid method with Gauss-Seidel relaxation for the Poisson equation. In W. Hackbusch and U Trottenberg, editors, *Multigrid Methods*, Lectures Notes in Mathematics No. 960, pages 368–386, Berlin Heidelberg New York, 1982. Springer-Verlag.

[11] D. Braess. Towards algebraic multigrid for elliptic problems of second order. *Computing*, 55:379–393, 1995.

[12] D. Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics.* Cambridge University Press, Cambridge, 1997.

[13] D. Braess and W. Hackbusch. A new convergence proof for the multigrid method including the V-cycle. *SIAM J. Numer. Anal.*, 20:967–975, 1983.

[14] James H. Bramble, Joseph E. Pasciak, Junping Wang, and Jinchao Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Math. Comp.*, 57:23–45, 1991.

[15] A. Brandt, S. F. McCormick, and J. W. Ruge. Algebraic multigrid (amg) for sparse matrix equations. In D. J. Evans, editor, *Sparsity and its Application*, pages 257–284. Cambridge University Press, Cambridge, 1984.

[16] Achi Brandt. Algebraic multigrid theory: the symmetric case. *Appl. Math. Comput.*, 19:23–56, 1986.

[17] Achi Brandt. Rigorous quantitative analysis of multigrid, i: constant coefficients two-level cycle with l2-norm. *SIAM J. Numer. Anal.*, 31:1695–1730, 1994.

[18] M. Brezina, A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge. Algebraic multigrid based on element interpolation (AMGe). *SIAM J. Sci. Comput.*, 22:1570–1592, 2000.

[19] William L. Briggs, Van Emden Henson, and Stephen F. McCormick. *A multigrid tutorial.* SIAM, Philadelphia, PA, 2000.

[20] V. E. Bulgakov. Multi-level iterative technique and aggregation concept with semi-analytical preconditioning for solving boundary-value problems. *Comm. Numer. Methods Engrng.*, 9:649–657, 1993.

[21] T. Chartier, R. D. Falgout, V. E. Henson, J. Jones, T. Manteuffel, S. McCormick, J. Ruge, and P. S. Vassilevski. Spectral AMGe (ρAMGe). *SIAM J. Sci. Comput.*, 25:1–26, 2004.

[22] Robert D. Falgout and Panayot S. Vassilevski. On generalizing the algebraic multigrid framework. *SIAM J. Numer. Anal.*, 42:1669–1693, 2005.

[23] Robert D. Falgout, Panayot S. Vassilevski, and Ludmil T. Zikatanov. On two-grid convergence estimates. *Numer. Lin. Alg. Appl.*, 12:471–494, 2005.

[24] G. H. Golub and C. F. van Loan. *Matrix Computations.* The John Hopkins University Press, Baltimore, Maryland, 1996. Third ed.

[25] M. Griebel and P. Oswald. On the abstract theory of additive and multiplicative Schwarz algorithms. *Numer. Math.*, 70:163–180, 1995.

[26] W. Hackbusch. Multi-grid convergence theory. In W. Hackbusch and U Trottenberg, editors, *Multigrid Methods*, Lectures Notes in Mathematics No. 960, pages 177–219, Berlin Heidelberg New York, 1982. Springer-Verlag.

[27] W. Hackbusch. *Multi-grid Methods and Applications.* Springer, Berlin, 1985.

[28] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952.

[29] R. Hiptmair. Multigrid method for Maxwell's equations. *SIAM Journal on Numerical Analysis*, 36(1):204–225 (electronic), 1999.

[30] J. Jones and B. Lee. A multigrid method for variable coefficient Maxwell's equations. *SIAM Journal on Scientific Computing*, 27(5):1689–1708, 2006.

[31] J. E. Jones and P. S. Vassilevski. AMGe based on element agglomeration. *SIAM J. Sci. Comput.*, 23:109–133, 2001.

[32] H. Kim, J. Xu, and L. Zikatanov. A multigrid method based on graph matching for convection-diffusion equations. *Numer. Lin. Alg. Appl.*, 10:181–195, 2003.

[33] Tzanio V. Kolev and Panayot S. Vassilevski. Amg by element agglomeration and constrained energy minimisation interpolation. *Numer. Lin. Alg. Appl.*, 13:771–788, 2006.

[34] Michael Lauzon and Sergei Treil. Common complements of two subspaces of a Hilbert space. *Journal of Functional Analysis*, 212:500–512, 2003.

[35] B. Lee and C. Tong. A novel algebraic multigrid-based approach for Maxwell's equations. *Preprint UCRL-JC-218750*, 2006.

[36] R. B. Lehoucq, D. C. Sorensen, and C. Yang, editors. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Philadelphia, PA, 1998.

[37] Jan Mandel, Steve F. McCormick, and John W. Ruge. An algebraic theory for multigrid methods for variational problems. *SIAM J. Numer. Anal.*, 25:91–110, 1988.

[38] S. F. McCormick. Multigrid methods for variational problems: general theory for the V-cycle. *SIAM J. Numer. Anal.*, 22:634–643, 1985.

[39] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM, Philadelphia, 2000.

[40] K. W. Morton and D. F. Mayers. *Numerical solution of partial differential equations: an introduction.* Cambridge University Press, Cambridge, 2005.

[41] A. C. Muresan and Y. Notay. Analysis of aggregation-based multigrid. *SIAM J. Sci. Comput.*, 30:1082–1103, 2008.

[42] S. Nakamura. *Computational Methods in Engineering and Science.* John Wiley & Sons, New York, 1977.

[43] Michael K. Ng. *Iterative methods for Toeplitz systems.* Oxford University Press, Oxford, 2004.

[44] Y. Notay. Algebraic analysis of two-grid methods: the nonsymmetric case. *Numer. Lin. Alg. Appl.* published online in Wiley InterScience. DOI: 10.1002/nla.649, 2009.

[45] Y. Notay. Flexible conjugate gradients. *SIAM J. Sci. Comput.*, 22:1444–1460, 2000.

[46] Y. Notay. Algebraic multigrid and algebraic multilevel methods: a theoretical comparison. *Numer. Lin. Alg. Appl.*, 12:419–451, 2005.

[47] Y. Notay. Convergence analysis of perturbed two-grid and multigrid methods. *SIAM J. Numer. Anal.*, 45:1035–1044, 2007.

[48] Y. Notay. An aggregation-based algebraic multigrid method. *Electronic Trans. Numer. Anal.*, 2009. To appear.

[49] Y. Notay and P. S. Vassilevski. Recursive Krylov-based multigrid cycles. *Numer. Lin. Alg. Appl.*, 15:473–487, 2008.

[50] P. Oswald. *Multilevel Finite Element Approximation: Theory and Applications.* Teubner Skripte zur Numerik. Teubner, Stuttgart, 1994.

[51] P. Oswald. *Subspace Correction Methods and Multigrid Theory*, pages 533–572. In Trottenberg et al. [61], 2001. Appendix A.

[52] S. Reitzinger and J. Schöberl. An algebraic multigrid method for finite element discretizations with edge elements. *Numerical Linear Algebra with Applications*, 9(3):223–238, 2002.

[53] J. W. Ruge and K. Stüben. Algebraic multigrid (AMG). In S. F. McCormick, editor, *Multigrid Methods*, volume 3 of *Frontiers in Applied Mathematics*, pages 73–130. SIAM, Philadelphia, PA, 1987.

[54] Y. Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, Philadelphia, PA, 2003. Second ed.

[55] Y. Saad and H. van der Vorst. Iterative solution of linear systems in the 20-th century. *J. Comput. Appl. Math.*, 123:1–33, 2000.

[56] Gordon D. Smith. *Numerical solution of partial differential equations: finite difference methods.* Oxford University Press, Oxford, 1985.

[57] R. Stevenson. *On the validity of local mode analysis of multi-grid methods.* PhD thesis, Utrecht University, 1990.

[58] K. Stüben. Algebraic multigrid (AMG): experiences and comparisons. *Appl. Math. Comput.*, 13:419–452, 1983.

[59] K. Stüben. *An Introduction to Algebraic Multigrid*, pages 413–532. In Trottenberg et al. [61], 2001. Appendix A.

[60] K. Stüben and K. U. Trottenberg. Multigrid methods: Fundamental algorithms, model problem analysis and applications. In W. Hackbusch and U. Trottenberg, editors, *Multigrid Methods*, Lectures Notes in Mathematics No. 960, pages 1–176, Berlin Heidelberg New York, 1982. Springer-Verlag.

[61] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid.* Academic Press, London, 2001.

[62] Henk A. van der Vorst. *Iterative Krylov methods for large linear systems.* Cambridge University Press, Cambridge, 2003.

[63] P. Vaněk, J. Mandel, and M. Brezina. Algebraic multigrid based on smoothed aggregation for second and fourth order problems. *Computing*, 56:179–196, 1996.

[64] Petr Vaněk, Marian Brezina, and Radek Tezaur. Two-grid method for linear elasticity on unstructured meshes. *SIAM J. Sci. Comput.*, 21:900–923, 1999.

[65] P. S. Vassilevski. *Multilevel Block Factorization Preconditioners.* Springer, New York, 2008.

[66] John L. Volakis, Arindam Chatterjee, and Leo C. Kempel. *Finite Element Method for Electromagnetics.* IEEE Press, New York, 1998.

[67] P. Wesseling. *An Introduction to Multigrid Methods.* J. Wiley and Sons, Chichester, 1992.

[68] R. Wienands and W. Joppich. *Practical Fourier Analysis for multigrid methods.* Chapman & Hall/CRC Press, Boca Raton, Florida, 2005.

[69] Roman Wienands and Cornelis W. Oosterlee. On three-grid Fourier analysis for multigrid. *SIAM J. Sci. Comput.*, 23:651–671, 2001.

[70] G. Wittum. Linear iterations as smoothers in multigrid methods: Theory with applications to incomplete decompositions. *Impact of Computing in Science and Engineering*, 1:180–215, 1989.

[71] G. Wittum. On the robustness of ILU-smoothing. *SIAM J. Sci. Statist. Comput.*, 10:699–717, 1989.

[72] J. Xu. *Theory of Multilevel Methods.* PhD thesis, Cornell University, 1989.

[73] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581–613, 1992.

[74] J. Xu and L. T. Zikatanov. The method of alternating projections and the method of subspace corrections in hilbert space. *J. Amer. Math. Soc.*, 15:573–597, 2002.

[75] H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numerica*, 2:285–326, 1993.

[76] O.C. Zienkiewicz. *Finite Elements: Introductory lectures on the finite element method.* McGraw-Hill, London, 1977.

[77] Ludmil T. Zikatanov. Two-sided bounds on the convergence rate of two-level methods. *Numer. Lin. Alg. Appl.*, 15(5):439–454, 2007.